Neurocomputing 363 (2019) 375–397

Contents lists available at ScienceDirect

Neurocomputing

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

A fast and robust 3D face recognition approach based on deeply learned face representation

Ying Cai[b,c], Yinjie Lei[d], Menglong Yang[a,c,*], Zhisheng You[a,c], Shiguang Shan[e]

[a] School of Aeronautics and Astronautics, Sichuan University, Chengdu, Sichuan, China
[b] School of Electrical and Information Engineering, Southwest Minzu University, Chengdu, Sichuan, China
[c] Wisesoft Software Co., Ltd., Chengdu, Sichuan, China
[d] School of Electronics and Information Engineering, Sichuan University, Chengdu, China
[e] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, China

ARTICLE INFO

Article history:
Received 9 April 2018
Revised 28 March 2019
Accepted 16 July 2019
Available online 23 July 2019

Communicated by Dr Xiaoming Liu

Keywords:
3D face recognition
Deep learning
Face preprocessing
Multiple data augmentation

ABSTRACT

With the superiority of three-dimensional (3D) scanning data, e.g., illumination invariance and pose robustness, 3D face recognition theoretically has the potential to achieve better results than two-dimensional (2D) face recognition. However, traditional 3D face recognition techniques suffer from high computational costs. This paper proposes a fast and robust 3D face recognition approach with three component technologies: a fast 3D scan preprocessing, multiple data augmentation, and a deep learning technique based on facial component patches. First, unlike the majority of the existing approaches, which require accurate facial registration, the proposed approach uses only three facial landmarks. Second, the specifical deep network with an improved supervision is designed to extract complementary features from four overlapping facial component patches. Finally, a data augmentation technique and three self-collected 3D face datasets are used to enlarge the scale of the training data. The proposed approach outperforms the state-of-the-art algorithms on four public 3D face benchmarks, i.e., 100%, 99.75%, 99.88%, and 99.07% rank-1 IRs with the standard test protocol on the FRGC v2.0, Bosphorus, BU-3DFE, and 3D-TEC datasets, respectively. Further, it requires only 0.84 seconds to identify a probe from a gallery with 466 faces.

1. Introduction

Face recognition (FR) has been one of the most significant research areas in computer vision. The problem of that two-dimensional (2D) facial imaging is sensitive to pose, illumination, and expression has been plaguing researchers [1]. However, three-dimensional (3D) facial imaging has shown several advantages in terms of FR. For example, (1) 3D scans provide more geometrical (depth) information than 2D images; (2) 3D scans are relatively invariant to scaling, rotation, and illumination [2]; (3) feature extracted from 3D scans are more robust for facial expression and pose variations [3–5]. Moreover, the 3D face model is widely used in auxiliary 2D FR [6–8]. Further, the attacking costs of 3D FR are higher than 2D FR in many scenes [9,10]. Consequently, the 3D FR has always been considered to be significant in research and application. The recent emergence of low cost 3D sensors (e.g., Microsoft Kinect, Intel RealSense) and easy 3D acquisition methods further accelerate the research of 3D FR. Many existing researches of 3D FR have achieved high accuracy in controlled scenarios [11–13].

Nevertheless, the majority of the existing systems are computationally expensive owing to the complicated feature extraction process or precise facial point registration required for alignment. For example, Li et al. [14] required 75 s for one recognition based on three curvature-based 3D keypoint descriptors. Al-Osaimi et al. [15] consumed 36.2 s and Lei et al. [12] required 8.49 s. Further, these approaches typically use handcrafted features to address specific 3D FR tasks. such as Lei et al. [12] addressed the partial situation using local Keypoint-based Multiple Triangle Statistics. Creusot et al. [2] used features extracted from target landmark shapes. Wang et al. [16] used Gabor features.

Because learning features are thought to have greater generalizing ability than handcrafted features, recent researches have learned features for 3D FR. Wang et al. [17] used boosting technology to enhance the low-level features resulting in a significantly improved recognition accuracy. However, they required 3.6 s per recognition based on a gallery of 1000 faces. Lei et al. [18] trained Kernel Principal Component Analysis (KPCA) for feature

* Corresponding author at: School of Aeronautics and Astronautics, Sichuan University, Chengdu, Sichuan, China.
E-mail address: mlyang@scu.edu.cn (M. Yang).

https://doi.org/10.1016/j.neucom.2019.07.047
0925-2312/

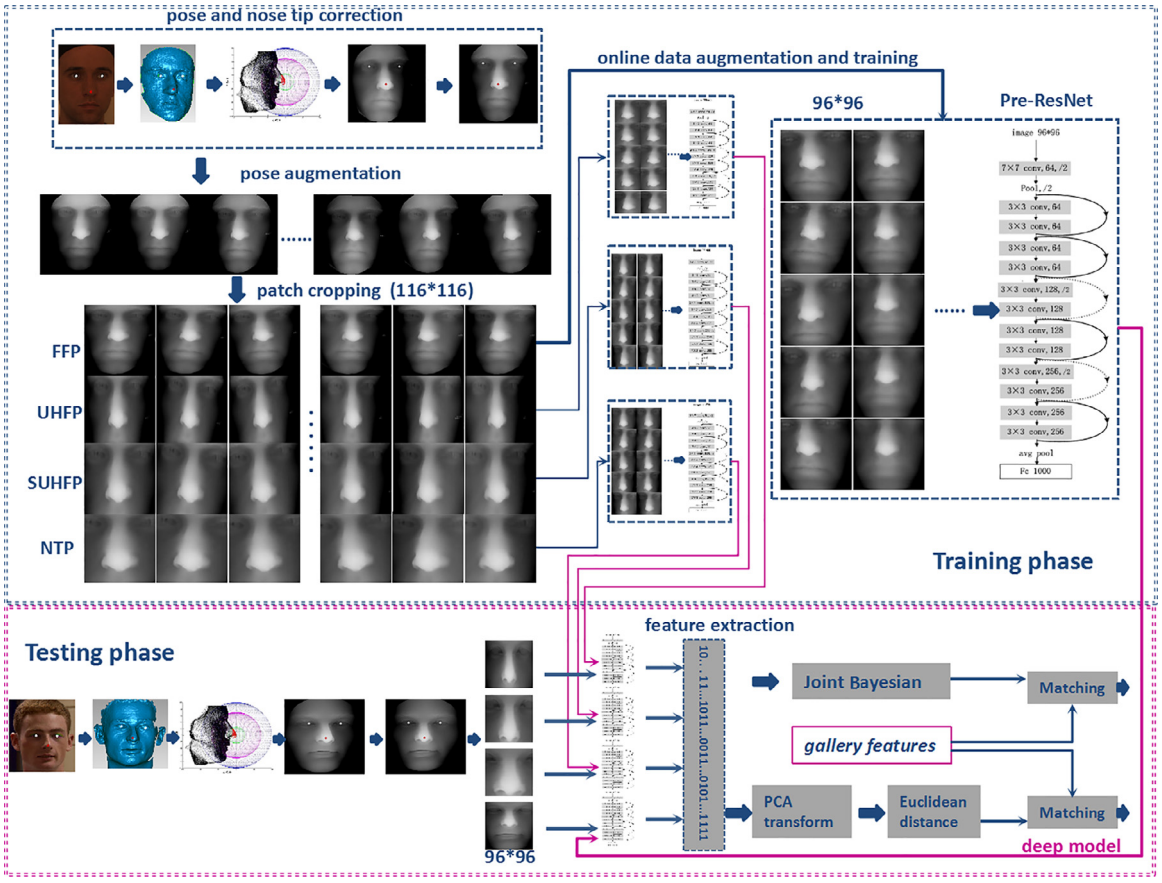


Fig. 1. Overview of proposed 3D face recognition system.

representations. Ballihi et al. [19] adopted the Adaboost algorithm to select robust features.

However, the recent breakthrough in feature learning obtained by deep learning technologies has demonstrated their superiority compared with shallow learning in terms of generalizing ability. Unfortunately, although the successful application of deep learning technologies in 2D FR, its corresponding applications in 3D FR remain in a state of infancy. There are two important reasons. Firstly, the size of existing 3D face datasets cannot compare with their 2D counterpart, which leads to a set of training difficulties (i.e., overfitting and low discrimination of features). Secondly, the majority of the existing deep networks are designed specifically for 2D face images, rather than 3D faces, owing to the order-less nature of the 3D data.

The first application of deep learning in 3D FR was proposed by Kim et al. [20]. Because of a lack of training data, they leveraged the public 2D face deep model (VGG-Face [21]) to obtain 3D deep representations. That is, they applied a 2D deep model, which was intrinsically designed for the 2D FR task, without performing specific optimization for the 3D FR task. The direct transfer of deep models from 2D to 3D is not the best approach when applying deep learning to 3D FR. Further, they required 3.25 s for one identification against a gallery with 466 faces. Recently, Gilani and Mian [22] proposed a large-scale 3D face recognition approach, but they did not provide their time cost analysis.

This paper proposes a fast and robust 3D FR approach for real-world applications. Firstly, a simple and fast preprocessing algorithm is proposed to obtain a 2.5D range image from raw 3D data. It relies only on the center of the two pupils and nose tip for alignment. Secondly, rather than using only the entire face, we cropped four overlapping facial patches from range images, including global, upper-semi-global, and different rigid-local information. These patches can provide different contributions to the recognition accuracy. Then, a facial data augmentation method (including pose augmentation, online resolution augmentation, and transformational augmentation) and three self-collected 3D face datasets (containing 13,000 textured 3D face scans from 3000 individuals) are adopted to enlarge the scale of the training dataset.

Next, the four overlapping facial patches are respectively fed into the selected deep pre-activation residual convolutional neural networks (Pre-ResNet) with an improved multi-scale triplet loss. The four 128-dimensional features extracted from the different patches then are concatenated to a 512-dimensional feature vector. Joint Bayesian [23] and Euclidean distance are respectively used to feature match. Test experiments are conducted on four mainstream public 3D face benchmarks and the results demonstrate that the proposed approach achieves the best recognition performance and highest computational efficiency compared with the state-of-the-art. An overview of the proposed approach is illustrated in Fig. 1.

The remainder of this paper is organized as follows: Section 2 provides a brief literature review of closely related work on traditional 3D FR algorithms, current deep learning in 3D representation, and current deep convolutional neural network (CNN) technologies in 2D image tasks. Section 3 describes the proposed approach in detail. Section 4 presents the experimental results and result analysis. Section 5 concludes this paper.

2. Related work and contributions

In this section, we review the literature closely related to the proposed approach. It is divided into traditional 3D FR algorithms,

current deep learning in 3D representation, and current deep CNN technologies in 2D image tasks.

2.1. Traditional 3D face recognition algorithms

According to the different types of features, traditional 3D FR approaches can be categorized into holistic features, local features, and holistic/local hybrid features.

Holistic features. The algorithms of [24–26] used an approach with entire surface registration to compute the distance between two faces. Mohammadzade and Hatzinakos [27] is based on closest normal points (CNPs), Liu *et al.* [28] extracted spherical harmonic features (SHF), and Gilani *et al.* [29] generated a 3D deformable model (R3DM). These approaches described the entire 3D face by defining a set of global features. The main disadvantage is their sensitivity to facial expressions.

Local features. A survey of 3D face recognition based on local features is presented in [30]. Creusot et al. [2] combined different local surface descriptors to vectors based on a learned distribution according to a given target landmark shape. Lei et al. [31] leveraged low-level geometric features collected from the eyes-forehead and the nose regions to achieve robustness to the facial expressions. Elaiwat et al. [32] proposed a local feature descriptor by integrating different Curvelet elements of different orientations. In general, local matchine-based approaches are known to be more robust to facial expressions, scale, and occlusion variations [33]. However, a single local descriptor has less discrimination power and can be affected by noise.

Holistic/Local hybrid features. Al-Osaimi et al. [34] structured a compact representation of a facial scan based on both local and global geometrical cues. It is an expression-robust 3D FR algorithm; however, it is difficult to determine the weights of the global and local contributions when combining the two representations. Huang et al. [35] used the Scale Invariant Feature Transform (SIFT) algorithm to extract Multi-Scale Local Binary Pattern (MS-LBP) features from a local depth map and Shape Index (SI) map from a global shape. Holistic/Local hybrid descriptors combine global and local information, which can compensate each other. Consequently, these approaches can obtain good performance.

3D FR approaches can also be categorized into learning-based and matching-based according on whether they involve a training phase.

Learning-based approaches. Ballihi et al. [19] adopted the Adaboost algorithm to select features from a large set of geometric curve features extracted from circular curves and radial curves of the Euclidean distance based on facial surface. They proposed a composite classifier that achieves high performance with a minimal set of features. Lei et al. [18] trained kernel principal component analysis (KPCA) to transform the Angular Radial Signature (ARS), which is extracted from the semi-rigid region of the face, to mid-level feature representations aiming to improve the discrimination ability. Wang et al. [17] used boosting to optimally select local features and train them as weak classifiers for assembling three collective strong classifiers to improve the accuracy. To avoid the registration between an input face and every face in the gallery, they used a fast posture alignment method. These learning-based approaches can improve, to some extent, the generalization ability of the entire 3D FR system.

Matching-based approaches. Lu et al. [36] and Mohammadzade and Hatzinakos [27] applied Iterative Closest Point (ICP) to match face surfaces. ICP is an iterative process that requires expensive computation. Elaiwat et al. [32] calculated the Curvelet transform to extract features from semi-rigid regions (eyes-forehead and nose) and proposed a multi-modal face identification approach. These matching-based approaches commonly require high computational expense.

2.2. Current deep learning in 3D representation

The deep learning technology applied to 3D vision tasks was not well developed until approximately three years ago. There are three main difficulties: (1) How to design an appropriate 3D object representation as input to a deep learning model, (2) Lacking of large-scale 3D datasets, and (3) How to control the computational cost in 3D tasks. The good news is that with the rapid technological development of low-cost 3D devices [37–39], the acquisition of 3D data has become easier, and the advancement of computing devices has further facilitated the application of deep learning in 3D tasks.

Existing approaches addressing 3D tasks based on deep learning can be classified into four categories based on the different inputs for training.

1) Inputting handcrafted features. These approaches used the handcrafted features extracted from 3D data as input to the deep network for learning high-level features. They leveraged the deep learning technology to improve the representation ability of the existing handcrafted features. Bu et al. [40] used bag-of-words to change the low-level 3D shape descriptors consisting of heat kernel signature and average geodesic distance to mid-level features; they then learned the high-level features using deep belief networks (DBNs) from the mid-level features. Jin et al. [41] computed multi-scale shape distributions and trained a deep auto-encoder for each scale. They connected all of these deep auto-encoders into a feature descriptor that is used for 3D object retrieval. However, using the handcrafted features in these approaches results in the inefficient utilization of the deep learning and thus their performance is determined by the discriminative ability of the handcrafted features.

2) Inputting projection image. The 3D objects are projected into a 2D space in these approaches and the resulting images are used to train the deep learning model as in typical 2D imaging approaches. There are two advantages to these approaches. Completed 2D deep learning models can be used to fine-tune the 3D deep representation with a time cost that is relatively small based on the projective image compared with the time cost based on 3D point clouds. Shi et al. [42] performed cylinder projection around the principal axis for each 3D object to convert a 3D shape into a panoramic view image to train the deep CNNs. Because a row-wise max-pooling layer was inserted between the convolution and fully-connected layers, the learned representations are invariant to the rotation around the principal axis. Sinha et al. [43] converted the original 3D model into a flat and regular geometry image using authalic parametrization on a spherical domain and a standard CNN is used directly to learn the features. Kalogerakis et al. [44] obtained a series of shaded images and range images with multi-views and multiscales from 3D scans and fed these images into a fully convolutional network. Their results are competitive in the currently largest segmentation benchmark. Huibin et al. [45] represented each textured 3D face scan as six types of 2D facial attribute maps (i.e., geometry map, three normal maps, curvature map, and texture map) and used each map as input to a branch of deep CNNs. Then, fusion layers were used to fuse the outputs of each branch and generate the final output of the entire model. They obtained the best results in expression recognition. Kim et al. [20] fine-tuned the public face deep model (VGG-Face [21]) to obtain their 3D deep representations using depth map images of 3D face scans. They achieved acceptable performance in 3D FR; however, owing to the rigid-ICP iterations required for face registration, this approach suffers from an expensive computational cost.

3) Inputting 3D voxel. These approaches represent a geometric 3D shape as a probability distribution of binary variables on a 3D voxel grid. Each 3D mesh is represented as a binary tensor. The main superiority of these approaches is that the representation

using 3D voxel, which includes the geometric shape, is beneficial to the discrimination ability. Wu et al. [46] adopted "1" to indicate that the voxel is inside the mesh surface and "0" to indicate that the voxel is outside the mesh. Then, they designed a Convolutional Deep Belief Network (CDBN) for learning the complex probabilistic distribution of the binary tensor. Xu and Todorovic [47] formulated CNN learning as a beam search aimed to form an optimal 3D ShapeNets architecture, namely the number of layers, nodes, and their connectivity in the network. Li et al. [48] represented 3D spaces as volumetric fields to address the sparsity of voxels and employed field probing filters to replace the convolutional layers in deep CNNs to extract features. Li et al. [48] generated 3D objects from a probabilistic space based on the volumetric convolutional networks and generative adversarial nets. Their unsupervised learned features have competitive performance on 3D object recognition. However, the main weakness of these approaches is that the complexity of training increases cubically with respect to voxel resolution. The common voxel resolution is $30 \times 30 \times 30$ for a 3D object; unfortunately, it remains seriously lacking for a 3D face shape.

4) Inputting raw data. These approaches can address the irregularity of 3D raw data when designing the input of a deep model. Commonly, mesh and point clouds are used in these approaches. Han et al. [49] proposed a mesh convolutional restricted Boltzmann machine (MCRBM) to unsupervisedly learn the features of a 3D object. Han et al. [50] used a circle convolutional restricted Boltzmann machine (CCRBM) to learn the unsupervised 3D local features. They sampled the extra points in each 3D local region and projected onto the tangent plane of the center of the region. The Fourier transform modulus was used to transform the projection distance distribution (PDD) into the Fourier domain, which was then conveyed to CCRBM for training. Qi et al. [51] designed a neural network that is directly based on point clouds, called PointNet. It can maintain the invariance of the input point cloud. They provided a unified approach to 3D object classification, part segmentation, and semantic segmentation. These approaches continue to suffer from expensive computational cost.

2.3. Current deep CNN technologies in 2D image tasks

In this section, we review the current deep CNN technologies in 2D image tasks from three perspectives: the scale of the training data, architecture of the network, and type of supervision.

Scale of training data. DeepFace [52] used a dataset of four million examples spanning 4000 unique identities to train a Siamese deep CNN and achieved acceptable performance on the LFW [1] and YFW [53] benchmarks. Then, the DeepId series of papers by Sun et al. [54–57] continuously increased the performance on LFW and YFW using a combination of CelebFaces [54] and WDRef [23] containing 0.3 M image face pairs of 1.3 K face identities. VGG Face [21] demonstrated how a large-scale dataset (2.6 M images, more than 2.6 K people) could be assembled by a combination of automation and "human in the loop" and discussed the tradeoff between data purity and time. FaceNet [58] used a large training dataset containing 200 M face images of 8 M face identities and listed the comparison of performance with different magnitudes of training data in detail. The increase of training samples for each category results in an improved recognition performance; however, it also leads to an increased difficulty of training.

Network architecture. CNN as a classic network architecture has significantly improved the state-of-the-art in numerous 2D image-processing tasks. The literature [21,52,55–58] indicates that the performance of CNN feature extractors is considerably superior to traditional handcrafted feature descriptors such as SIFT, LBP, and HOG [53,59,60] in 2D FR. Consequently, the network architecture of CNNs have realized significant progress. Sun et al. [54,55] connected the last hidden layer to both the third and fourth convolutional layers for learning multiscale features. Sun et al. [56,57] manifolded convolutional layers to improve performance and added fully connected layers after convolutional layers or pooling layers to enhance the supervision. DeepFace [52] used a Siamese architecture deep CNN to extend the network to double size. FaceNet [58] and GoogLeNet [61] trained a CNN with inception modules that could increase the depth and width of the network while maintaining the computational budget constant. ResNet [62] plugged several stacked residual units in the CNN architecture to ease network optimization and improve accuracy from the considerably increased depth. They won "1st place" in the ILSVRC 2015 classification competition. Batch Normalization [63] addressed the problem of internal covariate shift in training deep neural networks by normalizing layer inputs that eased the training processing and improved the performance of the image classification. He et al. [64] proposed a Pre-ResNet that suggests adding a pre-activation before every weight layer and using identity mappings as the skip connections in the ResNet architecture; they archived superior results compared to ResNet [62].

Types of supervision. Supervision in terms of loss function has a key role in deep learning. Literatures [52,54] used a softmax loss function with richer identity-related information to supervise a deep CNN. Literatures [55,65] joined the identity and verification signal to learn more discriminative features. Sun *et al.* [56] extended the supervision to each convolutional layer to obtain improved performance. Recently, FaceNet [58] proposed a triplet loss as supervision. They ensemble-minimized the distance between an anchor and a positive sample and maximized the distance between an anchor and a negative sample with deep embedding and obtained the best accuracy 99.63% in the LFW benchmark. Then, center loss [66] simultaneously learned a center for the deep features of each class and penalized the distances between the deep features and their corresponding class centers. They achieved the best results on MegaFace [67] under the protocol of small training set (containing less than 500,000 images and less than 20,000 individuals).

2.4. Our contributions

Aimed at two problems in the 3D FR task, i.e., the computational efficiency and recognition performance improvement, the main contributions of this paper are summarized as follows:

- First, to reduce the computational expense, a fast 3D scan preprocessing method is proposed. Specifically, a fast Principal Component Analysis (PCA) pose correction and a nose-tip refining on the raw 3D points cloud have been conducted firstly. Then, the 3D scan is projected to a range image and normalized in scale with only three facial landmarks. Because accurate yet expensive face alignment is avoided, the proposed preprocessing procedure requires only 0.76 s, which is approximately a third of that of the most current approaches. It makes the 3D FR easier to apply in real-world scenarios.
- Second, we combined multiple data augmentation methods and additional facial data collection to train the deep networks with a relatively small-scale training set. The multiple data augmentation based on the characteristics of 3D face data is used to construct the training set by modifying the pose, resolution, and geometry of the training faces. With the variations augmented, the over fitting problem of the learning is alleviated, resulting in improved recognition accuracy.
- As a third contribution, rather than applying a single deep network in a straightforward manner to learn the entire preprocessed range image, we adopt four deep networks to exploit the information in three overlapping face components

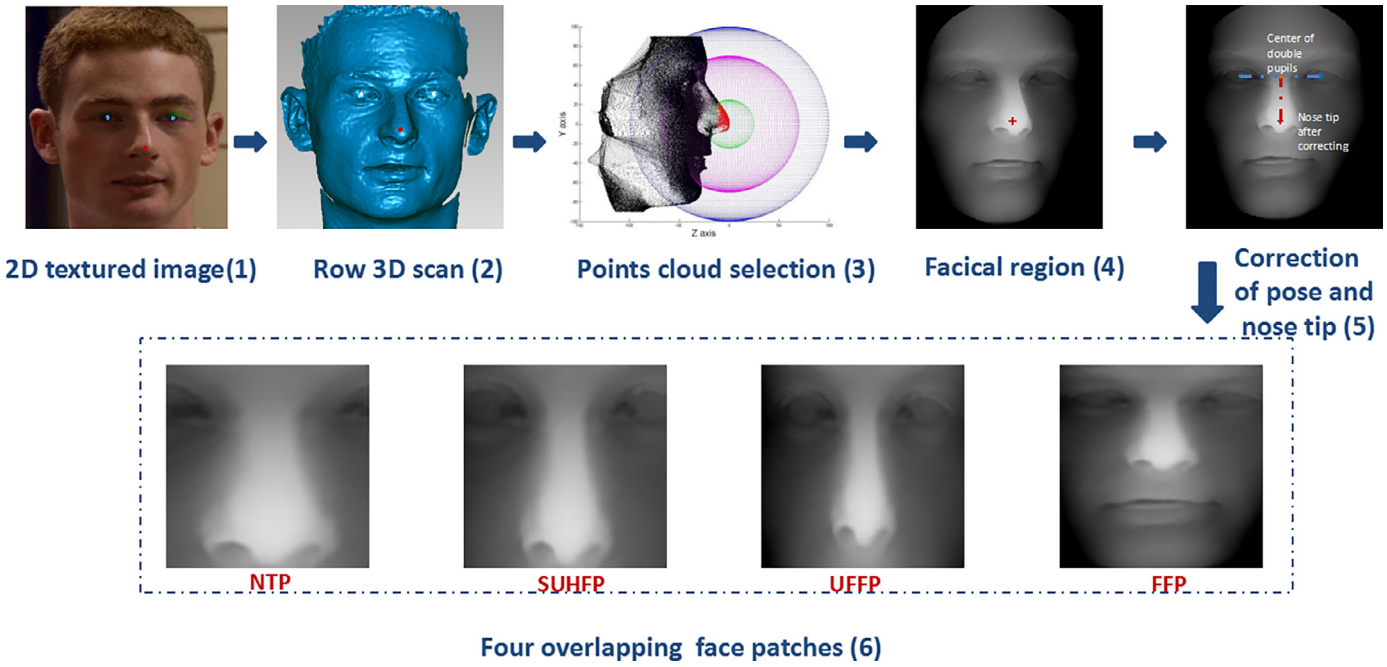


Fig. 2. Diagram of 3D face preprocessing.

and one entire facial region respectively. Their output features are then fused as the final face representation. Further, the deep networks are improved using a multiscale triplet loss function for faster convergence and superior discrimination ability.

With the above contributions, this paper presents a systematic work for 3D FR using specifically designed deep learning method rather than the existing public 2D options and achieves improved performance in both computational efficiency and recognition accuracy.

3. 3D face recognition based on deep learning

3.1. Facial scan preprocessing

The proposed preprocessing algorithm for 3D facial scan is composed of the following five steps. *Step 1:* Three facial landmarks (nose tip and center of pupils) are detected by DLib [68] from the corresponding 2D textured image, which is the texture data of the corresponding 3D face data. DLib is a continuous update toolkit; the details of the landmark algorithm can be found in the web page after [68]. Then, three landmarks on the 3D model are computed according to the correspondence between the 2D image and 3D scan. *Step 2:* Based on the observations, i.e., average size of the real face is slightly smaller than (approximately) 100 mm, we use a sphere centered on the nose tip with a radius of 100 mm (i.e., the blue sphere in sub-image-3 of Fig. 2.) to crop the face region. The points outside the blue sphere are removed, including the spikes. Because the proposed approach has tolerance to slight pose deviation, a simplified (omitting the aspect ratio adjustment and maintaining the number of iterations to less than five) PCA pose correction method based on the work of [26] is used to normalize the pose as sub-image-5 displayed in Fig. 2. Then, we remove the roll pose deviation by maintaining the two pupils in a level orientation. *Step 3:* Nose-tip detection based on a 2D textured image has a consideration of less computational cost compared with a similar detection in a 3D domain. However, the 2D domain detection has reduced detection accuracy, especially when pose variations are present. An example of such a case is illustrated in sub-image-4 in Fig. 2. Therefore, a nose-tip refine process utilizing depth information is performed to limit the ambiguity of the proposed nose-tip detection. We use a sphere centered on the nose tip with a radius of 25 mm (i.e., the green sphere in sub-image-3 of Fig. 2.) to crop the nose region and select the points whose deep distance to the nose tip is less than a certain threshold (15 mm in this paper, i.e., the red part in sub-image-3 of Fig. 2.) from the green sphere. Then, we smooth the red area using an average filter and perform interpolation. Because we have corrected pose, the highest point near the vertical centerline of the red area is the new nose tip. We extract seven slices in the middle part of the red area (where the point distribution is relatively uniform and dense) following [26] and use the mean of the horizontal midpoint of all slices to locate the vertical centerline of the face. This process is illustrated in Fig. 3, where we have increased the interpolation density for improved visibility. *Step 4:* We use the nose tip as the center of the image and normalize the facial mesh onto a square grid of 301×301 mm with a uniform resolution of 1 mm (i.e., the coordinate of the nose tip is 150,150). Then, we unify the vertical distance between the center of the two pupils and nose tip of all range images to a fixed value (40 pixels). *Step 5:* Many 3D face recognition methods [18,26] segment a face into rigid (nose), semi-rigid (eyes-forehead), and nonrigid (mouth) regions or semi-rigid (upper region) and nonrigid (lower region). The rigid facial regions are less affected by facial expressions resulting in a reliable descriptor [26] and 3D face recognition accuracy is highly influenced by the nonrigid facial regions [31]. The 2D face recognition method based on deep learning [55] selected 25 effective and complementary face patches from 400, including 12 upper regions, 10 global, and only three lower regions. Inspired by this, we combine global, upper-semi-global, and rigid-local descriptors to represent a 3D face. The global descriptors contribute more in identity discrimination. Upper-semi-global and rigid-local descriptors approve essential compensations. The four facial component patches with the same size of 116×116 pixel are generated as indicated in sub-image-6 of Fig. 2. In these facial patches, one including the entire face, FFP, one including the upper half face, UHFP, one including the small upper half face, SUHFP, and one including only the nose tip, NTP.

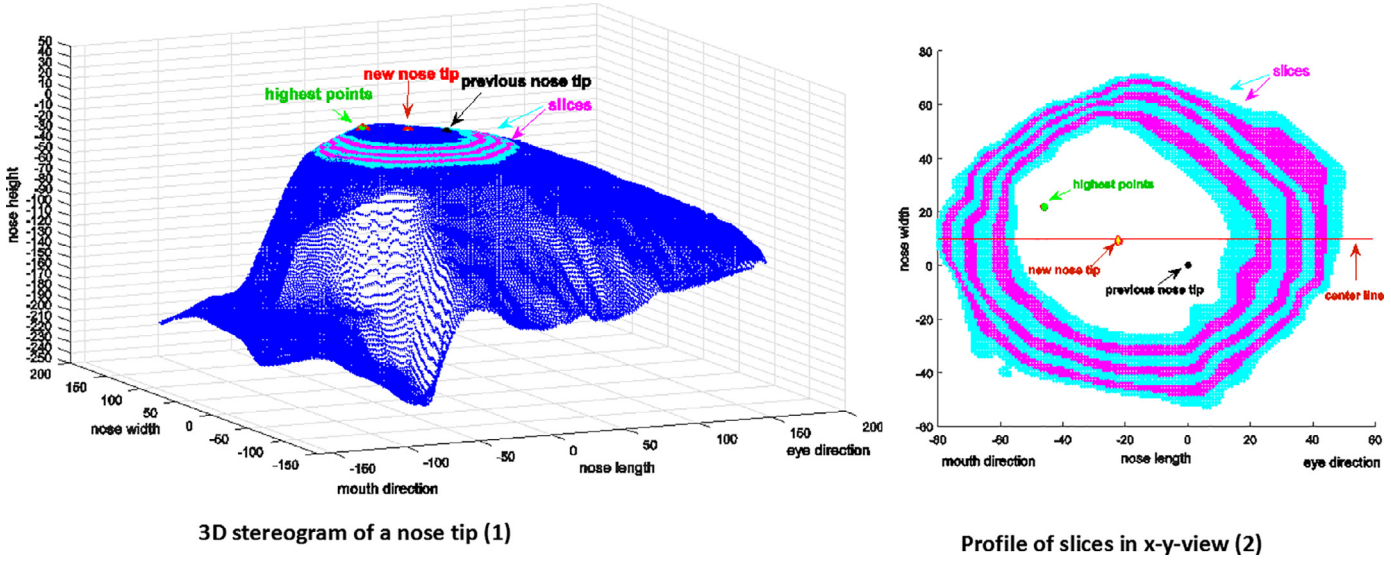


Fig. 3. Diagram of nose-tip refining method. Sub-fig-1 uses a 3D stereogram to indicate the result of nose-tip refining. Sub-fig-2 displays slices from the *x*-*y*-view, where the red line is the vertical centerline of face.

Table 1
Details of three self-building datasets.

Dataset name	Individual size	Gender	Age	Camera name	Number of 3D scans	Pose	Expression	Resolution
WS3D_set1	1000	Equal distribution	22–58	WS-FC120-II	1000	Front	Arbitrary	High
WS3D_set2	1000	Male	18–26	WS-FC120-II	1000	Front	Neutral	High
				WS-FX	1000	Front	Neutral	Medium
				RealSense	1000	Front	Neutral	Low
WS3D_set3	1000	Female	18–26	WS-FC120-II	5000	Front Pitch$^{\pm 10}$ Yaw$^{\pm 15}$	Neutral	High
				WS-FC120-II	1000	Front	Smile	High
				RealSense	1000	Front	Arbitrary	Low
				RealSense	1000	Front	Smile	Low
				RealSense	1000	Front	Neutral	Low

3.2. Construct training data

Extremely large 2D facial datasets have been collected (hundreds of millions), which significantly boosts 2D FR. However, there are significantly fewer 3D facial dataset because of the inherent drawbacks caused by 3D sensors (high cost, low portability, and unfriendly acquisition). For example, the largest public 3D dataset is ND 2006 [69], which contains 13,450 scans of 888 individuals. This is significantly less than the amount required to train a 3D deep learning model. To enlarge the 3D facial dataset to fulfil the requirements of training a 3D deep learning model, we first generate three self-built 3D facial datasets. Then, we conduct a multiple facial data augmentation to further enlarge the dataset.

3.2.1. Introduction of dataset

The proposed approach uses seven datasets including four public 3D datasets (FRGC v2.0, Bosphorus, BU-3DFE, and 3D-TEC) and three self-built datasets. Wisesoft Software Co., Ltd. and our laboratory are committed to the development of 3D face acquisition equipment and 3D FR technology. We created three 3D face datasets named WS3D_set1, WS3D_set2, and WS3D_set3, where each dataset includes 1000 individuals. Three 3D cameras were used for data acquisition in this paper. They included a self-developed high-accuracy camera (WS-FC120-II), a self-developed middle-accuracy device (WS-FX), and a low-accuracy device (Intel®RealSense^{TM}). The accuracy of the measurement of the corresponding cameras were 0.1 mm, 0.3–0.5 mm, and 1–1.5mm, respectively. Details of the three datasets are presented in Table 1 and examples of the three different datasets are displayed in Figs. 4–6, respectively. In total, the three datasets contain 13,000 textured 3D face scans of 3 K individuals. These three self-built 3D face datasets are used only to train the deep network.

There are two reasons for using mixed resolution data to form the training datasets. One is to expand the number of training sets to prevent overfitting on high resolution. The other is to increase the robustness of the features to the resolution. Although low-resolution training data is not sufficiently precise for face details, the overall trend of depth information remains consistent with the real face distribution, which has a positive influence on performance. In Section 4.2.2 regarding the data contribution analysis, the experimental analysis of the effect of low-precision training data on the recognition accuracy is provided.

3.2.2. Multiple data augmentation

In this paper, a multiple data augmentation process is conducted to enlarge the size of the training data, which includes pose augmentation based on 3D scan, resolution, and transformational augmentation based on range images.

(1) Pose augmentation for 3D model. One advantage of the 3D model is that different views can be obtained from one 3D model via rotation. Consequently, we rotate one 3D face and obtain the corresponding range images with different poses before normalization in Step4 discussed in Section 3.1. Specifically, five random poses in each rotation angle of pitch, yaw, roll (within the range of $\pm 10^{o}$ with an arbitrary step, 15 in total) and another five arbitrary poses in combining style are generated. Including the row frontal-view image, 21 range images are obtained for each 3D face by pose augmentation.

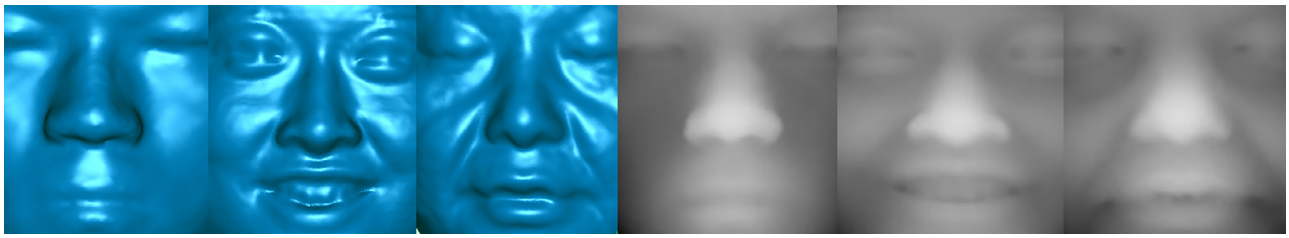


Fig. 4. Examples belonging to three different individuals in dataset WS3D_set1; first three images are raw 3D scans and latter three images are corresponding range images.

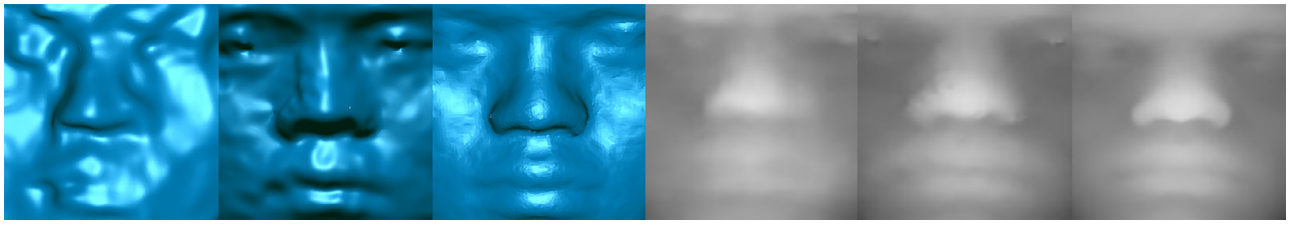


Fig. 5. Examples belonging to one individual in dataset WS3D_set2; first three images are raw 3D scans and latter three images are corresponding range images.

Fig. 6. Examples belonging to one individual in dataset WS3D_set3; first six images captured by high-resolution camera and letter three images captured by low-resolution camera.

(2) Transformational augmentation for range images. We apply minor random affine transformation (including rotation, shearing, and zooming), projection transformation, twisting, and horizontal flipping on training images in each iteration in the training process. The online transformation is aimed at obtaining a different transformation in different iterations; a random two-thirds of the training samples are applied in the online transformation in each iteration.

(3) Resolution augmentation for range images. Many 3D face datasets are acquired with different z-axis resolution because of the varied specifications of the corresponding 3D sensors. For example, in the BU-3DFE dataset, the z-axis resolution was clearly less than the other three public test datasets in our experiment. Such differences can cause different presentations of corresponding range images (i.e., a high-resolution results in a sharp range image and vice versa).

To accommodate more 3D facial scans with different z-axis resolution, we manually smoothen the range images simulating different z-axis resolution. Specifically, we randomly select four smooth templates from four median filters and four averaging filters with sizes 3×3, 5×5, 7×7, and 9×9 to perform an online smoothing operation on the training images, i.e., there are four range images with random resolution derived from the original range image if it is selected for resolution augmentation. In the training process, we randomly select one tenth of the training samples to conduct resolution augmentation in each iteration.

Leveraging this augmentation method, a multiresolution, multipose, minor multi-transformation training data set are constructed to guarantee the learning of the deep face representation. Further, after augmentation, all the patches are cropped to 96×96 pixels.

Sample images of our multiple data augmentation are presented in Fig. 7. The first column is the original range images. In the first row, the first image belongs to Bosphorus and the other images are derived from it using random transformational augmentation. In the second row, the first image belongs to BU-3DFE and the other images are derived from it using random pose augmentation. In the third row, the first and sixth images without augmentation belong to Bosphorus and FRGC V2.0, respectively, and the subsequent images are derived from the two images using random resolution augmentation.

3.3. Improvement of supervision and architecture of Pre-ResNets

The residual networks proposed by [62] have been widely used in the vision field because they are easier to optimize and can gain accuracy from the increased depth of the network. In this paper, three-deep residual networks with different layers are designed for our 3D FR task. We use the same Residual Unit as [64] and adopt the suggestion from [64], i.e., the identity mappings are used as the skip connections for improved recognition accuracy. Furthermore, activation functions ReLU and BN [63] are applied as pre-activation before every weight layer in our network. They are denoted as Pre-ResNet-14, Pre-ResNet-24, and Pre-ResNet-34. The architecture of Pre-ResNet-34 is the same as the 34-layer residual network in [62]; however, the input image size

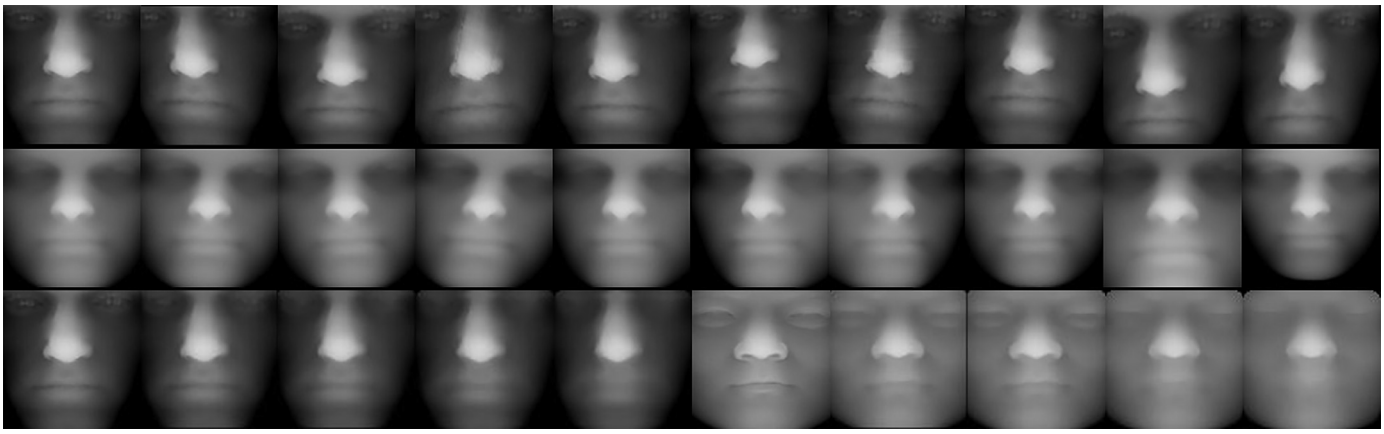


Fig. 7. Sample range images of data augmentation.

Table 2
Architectures of Pre-ResNet-34, Pre-ResNet-24, and Pre-ResNet-14. Every bracket is a building block with a shortcut. Down-sampling is performed by conv3_1, conv4_1, and conv5_1 with a stride of "2".

Layer name	Output size	34-layer	24-layer	14-layer
conv1	48 × 48	7 × 7, 64, stride 2		
		3 × 3 max pool, stride 2		
conv2_x	24 × 24	$\begin{bmatrix} 3\times3,\ 64 \\ 3\times3,\ 64 \end{bmatrix}\times3$	$\begin{bmatrix} 3\times3,\ 64 \\ 3\times3,\ 64 \end{bmatrix}\times3$	$\begin{bmatrix} 3\times3,\ 64 \\ 3\times3,\ 64 \end{bmatrix}\times2$
conv3_x	12 × 12	$\begin{bmatrix} 3\times3,\ 128 \\ 3\times3,\ 128 \end{bmatrix}\times4$	$\begin{bmatrix} 3\times3,\ 128 \\ 3\times3,\ 128 \end{bmatrix}\times4$	$\begin{bmatrix} 3\times3,\ 128 \\ 3\times3,\ 128 \end{bmatrix}\times2$
conv4_x	6 × 6	$\begin{bmatrix} 3\times3,\ 256 \\ 3\times3,\ 256 \end{bmatrix}\times6$	$\begin{bmatrix} 3\times3,\ 256 \\ 3\times3,\ 256 \end{bmatrix}\times4$	$\begin{bmatrix} 3\times3,\ 256 \\ 3\times3,\ 256 \end{bmatrix}\times2$
conv5_x	3 × 3	$\begin{bmatrix} 3\times3,\ 512 \\ 3\times3,\ 512 \end{bmatrix}\times2$		
	1 × 1	average pool, 1024-d fc1, triplet loss 128-d fc2, softmax+triplet loss		

of [62] is 224 × 224 and our size is 96 × 96. The architecture of the three networks are presented in Table 2. The down sampling is performed in every row and a shortcut connection is added to each pair of 3 × 3 filters as in [62].

The supervisory method of CNNs has become an active research topic with the widespread use of CNNs. Effective supervision can reduce the complexity of training and improve the feature discrimination ability. Our training set has three distinct features: (1) The total number of individuals is less than 4000, which is considerably less than the 2D face dataset. (2) Our training samples are acquired using different 3D sensors, which make the samples of different quality. (3) Many individuals have only a single 3D scan. The above observations add difficulty to training in terms of the selection of a supervision method. Schrok *et al.* [58] using triplet loss trained two networks for 2D FR and achieved state-of-the-art accuracy on Labeled Faces in the Wild (LFW) dataset. A similar work was proposed by Liu *et al.* [70]. The triplet loss ensemble minimizes the distance between an anchor and a positive sample. It also maximizes the distance between an anchor and a negative sample. Consequently, it enables a force separation between the positive pair and the negative by a distance margin.

Traditional triplet loss can, on occasion, increase the distance between all samples, resulting in the negative samples contributing significantly less. To avoid this drawback of triplet loss, the L2 distance between the positive sample into the triplet loss has been added as a constraint to avoid this situation and guarantee the sufficient contribution of the negative sample, as indicated in formula (1). Furthermore, the supervision are used on both the last fully connected layers and last feature layer. Finally, joining the two layers' triplet loss and the softmax loss, we construct the complete multi-scale triplet loss supervision for our training.

$$L = \sum_{i=1}^{N} [d_{i+} - d_{i-} + m] + \lambda d_{i+} \tag{1}$$

The parameter m is a margin that is enforced between positive and negative pairs. According to the dimension of our features, it is set to 6000. The parameter λ is a constant parameter, and we set it to 1 aiming to coequally constrain the two parts of the equation. The d_{i+} is the $L2$ distance of the positive pair and the d_{i-} is the $L2$ distance of the negative pair. The cardinality of all possible triplets is N.

We selected all possible positive sample pairs and randomly selected two negative samples for each positive sample, which constituted two triplets. Finally, we randomly sorted of all triplets. The proposed approach has achieved good performance without any other especial policy of the triplets. A careful designed selectional strategy may further improve the accuracy.

3.4. Recognition

After the deep networks are trained, the four groups' embedded 128-dimensional feature vectors from different face patches can be used for 3D face recognition singly or in combination. Two methods are used to match the features between the gallery and probe. (1) Following the existing approaches with high recognition accuracy [54], the Joint Bayesian are trained as a classifier to score the matching. To reduce the time cost, we use only the public 3D datasets for training. Except where otherwise stated, one dataset is used for testing, the remaining three public datasets are used to train the Joint Bayesian. (2) The features transformed by PCA can

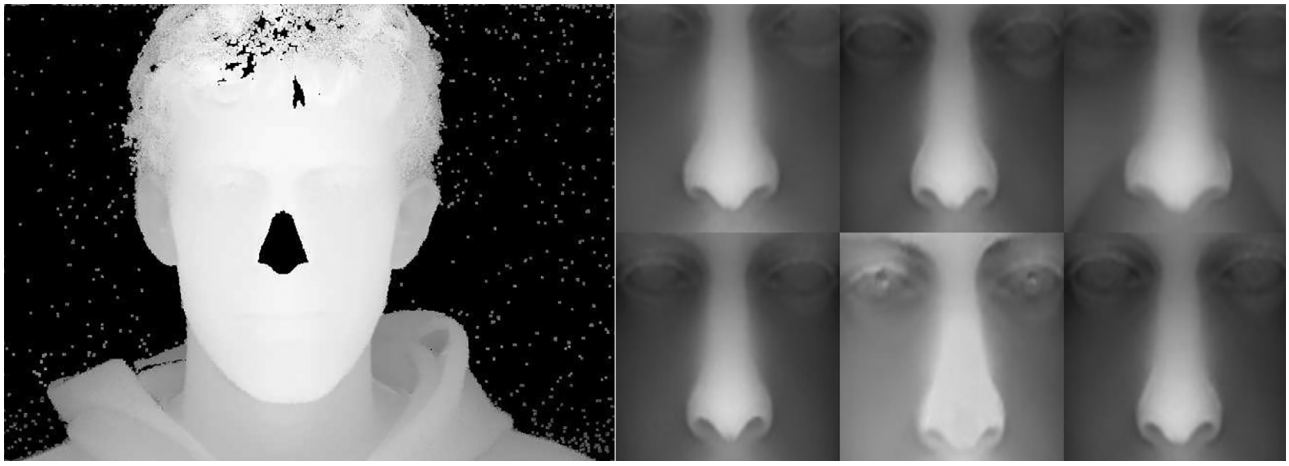


Fig. 8. Left image is raw data of scan 04505d222 with missing nose-tip region; right six range images belong to individual 04505. Middle image in second row is range image of scan 04505d222.

be used for face recognition by minimizing the Euclidean distance between the probe and gallery faces. The PCA is performed on the features of the probe set with the features of the gallery set.

4. Experimental results

We performed experiments on four challenging 3D face datasets, namely FRGC v2.0, Bosphorus, BU-3DFE, and 3D-TEC. In this section, we first perform a comparison between the results of the proposed approach and the state-of-the-art methods on the four 3D face datasets in Section 4.1. Then, we analyze the key modules in the proposed algorithm in Section 4.2. Finally, a time cost analysis is presented in Section 4.3.

Except where otherwise stated, all experiments were conducted with the following four conditions: (1) One dataset was used for testing and the remaining six were used to train the deep networks. (2) Because Pre-ResNet-24 obtained the best performance compared with the other ResNet in the next Section, 4.2.5, we used it for all experiments. (3) Our modified multi-scale triplet loss was used for supervision in all experiments. (4) All experiments regarding the FRGC dataset used features from only one face patch SUHFP because it was sufficient to obtain full marks; we use SUHFP in the subscript of the name of the test method to indicate this situation. The experiments regarding other datasets using feature fusion from all four face patches use ALLP in the subscript to indicate such.

4.1. Performances on the 3D datasets

4.1.1. Results on the FRGC v2.0

The FRGC v2.0 dataset is divided into three partitions based on the season of acquisition: Spring2003 set (943 scans of 277 individuals), Fall2003 set, and Spring2004 set (4007 scans of 466 subjects). We follow the same protocol [71] to use the Spring2003 for training and the remaining two partitions for validation. FRGC v2.0 has the presence of large variations in facial expressions, illuminations, and limited pose variations. Owing to the superior quality and less expression change compared with other 3D face datasets (e.g., the Bosphorus and BU-3DFE), the majority of the existing literature has achieved acceptable performance with this dataset. Five experiments were conducted to evaluate the proposed approach on FRGC v2.0; two of these are more complicated than in the majority of the literature.

We used the neutral scan of each individual (466 in total) to form the gallery set and three probe subsets were selected; the details are as follow: (1) “neutral vs. neutral” (1944 probes), (2) “neutral vs. non-neutral” (1597 probes), and (3) “neutral vs. all” (3541 probes). Note that only using the SUHFP single patch and Euclidean distance, the proposed approach obtained 100% VR at 0.1% FAR in all experiments and 100%, 99.88%, and 99.94% rank-1 IR in “neutral vs. neutral”, “neutral vs. non-neutral”, and “neutral vs. all” experiments, respectively (we denote the testing mode as ours$^{eu}_{suhfp}$ for the later experiments).

We had one scan of “04505d222” in the FRGC v2.0 without a nose-tip region, as displayed in Fig. 8 (we manually located the nose tip of the scan for testing; it is the sole manual operation in this work). When we adopted the Joint Bayesian trained using the Spring2003 subset of FRGC v2.0 as a classifier, the rank-1 IR in all experiments increased to 100% (we denote the testing mode as ours$^{jb}_{suhfp}$). Performance comparisons between the proposed approach and the existing approaches with the same test protocols are presented in Table 3. Further, Fig. 8 displays the comparisons between six range images belonging to individual “04505”; the middle range image in the second row is the range image of “04505d222”.

Then, to evaluate the robustness of the proposed approach between different sessions (Spring2003, Fall2003, and Spring2004), we performed the standard ROCIII experiment using public FRGC masks [71]; specifically, the gallery images collected in the Fall semester (1,893) and probe images collected in the Spring semester (2,114). The gallery and probe samples derive from different semesters in the ROCIII experiment. The comparisons between the proposed approach and the state-of-the-art approaches are presented in Table 3(a). We again achieved the highest 0.1% FAR VR of 100% using ours$^{eu}_{suhfp}$.

Finally, aiming to increase the difficulty of the experiment to evaluate the performance of the proposed approach under serious facial deformation, we performed an open-mouth experiment, which was first proposed by Berretti *et al.* [72]. Specifically, the neutral scans of each individual (466 in total) were again used as the gallery set and all the facial scans with an open mouth were selected as the probes[1].

Although an open mouth significantly deforms a facial surface, we again obtained 100% both on the rank-1 IR and the 0.1% FAR VR using ours$^{eu}_{suhfp}$. A comparisons of the result is presented in Table 3(c). The above observations clearly demonstrate the effectiveness of the proposed approach.

[1] All 816 probes were manually selected by Berretti et al. [72]; a list is available at: http://www.dsi.unifi.it/~berretti/frgcv2.0/.

Table 3
Comparisons between proposed approach and state-of-the-art approaches on FRGC v2.0 dataset.

(a) 0.1% FAR VRs for "neutral vs. neutral", "neutral vs. non-neutral", "neutral vs. all", and "ROCIII" experiments.

Approaches	*n* vs. *n*	*n* vs. *nn*	*n* vs. all	ROCIII
Maurer et al. [73] (2005)*	97.8%	–	86.5%	92.0%
Kakadiaris et al. [74] (2007)	–	–	–	97.0%
Mian et al. [26] (2007)*	99.4%	97%	98.5%	–
Mian et al. [75] (2008)*	99.9%	92.7%	97.4%	–
Al-Osaimi et al. [15] (2009)	98.4%	97.8%	98.1%	94.1%
Wang et al. [17] (2010)	–	–	98.6%	98.0%
Berretti et al. [72] (2010)	97.7%	91.4%	95.5%	–
Huang et al. [76] (2012)	99.6%	97.2%	98.4%	95.0%
Lei et al. [18] (2014)	–	97.8%	–	96.7%
Elaiwat et al. [32] (2015)*	99.9%	93.1%	–	97.8%
Lei et al. [12] (2016)	99.9%	96%	98.3%	–
ours$^{eu}_{suhfp}$	**100%**	**100%**	**100%**	**100%**

(b) rank-1 IRs for the "neutral vs. all" experiment.

Approaches	*n* vs. all
Cook et al. [77] (2006)	94.6%
Kakadiaris et al. [74] (2007)	97%
Mian et al. [26] (2007)*	96.2%
Mian et al. [75] (2008)*	93.5%
Al-Osaimi et al. [15] (2009)	96.5%
Aly et al. [15] (2009)	97.5%
Wang et al. [17] (2010)	98.4%
Queirolo et al. [78] (2010)	98.4%
Berretti et al. [72] (2010)	≈94%
Huang et al. [76] (2012)	97.6%
Elaiwat et al.[32] (2015)*	97.1%
Li et al.[14] (2015)	96.3%
Lei et al.[12] (2016)	96.3%
ours$^{eu}_{suhfp}$	99.94%
ours$^{jb}_{suhfp}$	**100%**

(c) rank-1 IRs and VRs for "neutral vs. open-mouth" experiment.

Approaches	IRs	0.1% FAR VRs
Berretti et al. [72]. (2010)	91%	-
Lei et al. [12]. (2016)	89.2%	93.8%
ours$^{eu}_{suhfp}$	**100%**	**100%**

4.1.2. Results on Bosphorus

The Bosphorus dataset [79] includes 4666 scans collected from 105 individuals (60 men and 45 woman aged between 25 and 35). These scans have been acquired under pose changes, expression variations (both emotion and action units), and typical occlusions. In our experiments, we first evaluated the performance of the proposed approach on the expression variations.

Consequently, a subset containing only expression face scans was collected, which contained six facial expressions (i.e., anger, disgust, fear, happiness, sadness, and surprise) and 28 facial action units (i.e., 20 Lower AUs, five Upper AUs, and three combined AUs). Thus, there were 2902 scans with approximately 34 different expressions in the expression subset. We performed the three experiments following the same protocol as [20]; the gallery contained 105 neutral scans in all three experiments, for details: (1) neutral vs. neutral (194 probes), (2) neutral vs. non-neutral (2603 probes), (3) neutral vs. all (2797 probes).

The proposed approach achieved 100%, 99.73%, and 99.75% rank-1 IRs and 100%, 98.30%, and 98.39% VRs at 0.1% FAR, respectively. All of the above results were obtained using the fusion of all features from the four patches (denoted as: ALLP) and trained the Joint Bayesian as classifier. The results based on ALLP and Euclidean distance are marginally worse, i.e., 100%, 99.42%, and 99.46% rank-1 accuracies and 100%, 97.46%, and 97.64% VRs at 0.1% FAR, respectively. The ROC and CMC curves are presented in Fig. 10. The rank-1 IR comparisons with the state-of-the-art methods on the neutral vs. all experiment are provided in Table 4, which indicates that the 99.75% and 99.46% obtained by the proposed approach are the first and second accuracy, respectively.

Table 4
Comparisons of rank-1 IRs in "neutral vs. all" experiment on facial expression subsets of Bosphorus dataset.

Approaches	*n* vs. all IRs
Ocegueda et al. [80] (2011)	98.6%
Berretti et al. [72] (2013)	95.7%
Smeets et al. [81] (2013)	97.7%
Li et al. [82] (2014)	95.4%
Li et al. [14] (2015)	98.8%
Lei et al. [12] (2016)	98.9 %
Kim et al. [20] (2017)	99.2%
Ours$^{eu}_{allp}$	99.46%
Ours$^{jb}_{allp}$	**99.75**%

The proposed approach required only a simple alignment method based on three landmarks, which is computationally efficient compared with the majority of existing alignment-based approaches. Although it is without any complicated alignment, the robust feature descriptors extracted by the proposed deep learning framework can effectively address the facial expression variations. This demonstrates that learning expression-invariant feature is a solution that is more efficient than alignment.

During the training phase, a pose augmentation process was performed to enlarge the training set. Consequently, the proposed approach can accommodate minor pose variations. To evaluate the

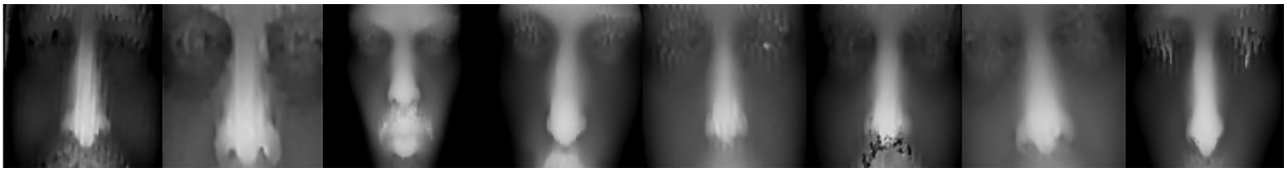


Fig. 9. UHFP patch of range images with defective preprocessing in Bosphorus dataset; third image presents serious cropping error due to incorrect nose-tip location.

Table 5
Comparison with state-of-the-art methods on BU-3DFE dataset in terms of 0.1% FAR VRs and rank-1 IRs for "neutral vs. al" experiment.

Approaches	*n* vs. all VRs	*n* vs. all IRs
Berretti et al. [83] (2013)	–	87.50%
Ocegueda et al. [84] (2013)	96.30%	99.30%
Li et al. [82] (2014)	–	92.20%
Elaiwat et al. [32] (2015)*	81.50%	–
Lei et al. [12] (2016)	94.00%	93.20%
Kim et al. [20] (2017)	–	95.00%
Ours^{eu}_{allp}	**98.92%**	**99.88%**
Ours^{jb}_{allp}	98.25%	99.67%

ability against pose variation of the proposed approach, three experiments were conducted as described following. The 105 Neutral scans were selected as the gallery and the scans with $\pm 10^0$ pose (105 probes), $\pm 20^0$ pose (105 probes), and $\pm 30^0$ pose (105 probes) in yaw rotational angle were selected as probes. The proposed approach achieved 100% rank-1 IR and 100% VR at 0.1% FAR in the "neutral vs. $\pm 10^0$ pose" experiment, 100% rank-1 IR and 98.1% VR at 0.1% FAR in the "neutral vs. $\pm 20^0$ pose" experiment, and 96.50% rank-1 IR and 92.3% VR at 0.1% FAR in the "neutral vs. $\pm 30^0$ pose" experiment. The performance on "neutral vs. $\pm 30^0$ pose" decreased 3.5% and 5.8% on rank-1 IR and 0.1% FAR VR, respectively, compared with "neutral vs. $\pm 20^0$ pose" because we did not restore the missing data caused by pose.

It can be observed from the CMC curves in Fig. 10(a) and (b) that the IR remains less than 100% until rank $n = 20$ in the "natural vs. all" experiment. This is mainly a consequence of the errors in nose-tip detection, pose correction, and serious noise (indicated in Fig. 9). The reason is that the proposed method does not perform complicated preprocessing in consideration of efficiency.

4.1.3. Results on the BU-3DFE

The BU-3DFE dataset [85] contains 2500 3D textured facial scans of 100 individuals (44 males and 56 females). Each individual has with six expressions (e.g., happiness, anger, fear, disgust, sadness, and surprise) and each expression has four levels of intensity. Levels 1 and 2 are considered as low intensity and Levels 3 and 4 are high intensity. The depth resolution of the BU-3DFE dataset is low compared with the other datasets and includes a considerable range of diversity in age and racial ancestries. Therefore, the BU-3DFE dataset is considered to be more challenging than other public datasets.

The neutral scan of each individual was selected to form the gallery (100 gallery) and three experimental scenarios depending on the intensity of expression were conducted as follow. (1) Neutral vs low-intensity (1200 probes), (2) neutral vs high-intensity (1200 probes), and (3) neutral vs. all (2400 probes). The proposed approach achieved rank-1 IRs of 99.92%, 99.83%, and 99.88% and 0.1% FAR VRs of 99.67%, 97.85%, and 98.92% in the three experiments, respectively. The comparisons between the proposed approach and the state-of-the-art approaches in the "neutral vs. all" experiment are listed in Table 5. We also achieved first place in both IRs and 0.1% FAR VRs using Euclidean distance. The performance degradation from "neutral vs. low-intensity" to "neutral vs. high-intensity" was only 0.09% in terms of the IR. These results clearly demonstrate the stability of the proposed approach. The CMC and ROC curves of the three experiments based on Joint Bayesian and Euclidean distance, respectively, are indicated in Fig. 11 and the results of rank-1 IRs and 0.1% FAR VRs are presented in Table 6.

Table 6
Rank-1 IRs and 0.1% FAR VRs results based on Joint Bayesian and Euclidean distance of "neutral vs. low-intensity ", "neutral vs. high-intensity", and "neutral vs. all" experiments on BU-3DFE.

Test protocol	IRs-JB	IRs-EU	VRs-JB	VRs-EU
Neutral vs. low-intensity	99.92%	99.92%	99.42%	99.67%
Neutral vs. high-intensity	99.42%	99.83%	97.17%	97.85%
Neutral vs. all	99.67%	99.88%	98.25%	98.92%

Table 7
Comparison of rank-1 IRs with state-of-the-art methods for "Case I", "Case II", "Case III", and "Case IV" experiment on 3D-TEC dataset.

Approaches	Case I	Case II	Case III	Case IV
Kakadiaris et al. [74] (2007)	98.1%	98.1%	91.6%	93.5%
Faltemier et al. [25] (2008)	94.4%	93.5%	72.4%	72.9%
Huang et al. [87] (2011)	91.1%	93.5%	77.1%	78.5%
Huang et al. [88] (2011)	91.6%	93.9%	68.7%	71.0%
Li et al. [82] (2014)	94.4%	96.7%	90.7%	92.5%
Kim et al. [20] (2017)	94.8%	94.8%	81.3%	79.9%
Ours^{eu}_{allp}	98.13%	98.13%	93.46%	**94.86%**
Ours^{jb}_{allp}	**99.07%**	**99.07%**	**94.39%**	94.39%

4.1.4. Results on the 3D-TEC

The 3D-TEC dataset consisted of 107 pairs of twins (two of the triplets were included as the 107th set of twins, total 214 subjects). Each subject contained two scans: one neutral scan and one smile scan. More details can be found in [86]. This dataset thus enables a more challenging evaluation of 3D FR algorithms because the pairs of facial scans of twins include only minor difference in presence. The standard testing protocols for 3D-TEC included four scenarios (Case I, II, III, and IV); the details are described in [82]. Case III and IV stages were the most challenging scenarios for 3D FR owing to the fact that the expression of a probe was different from the gallery face yet similar to the twin's expression in the gallery.

In these experiments, we are in accordance with standard protocols to evaluate the performance of the proposed approach. We obtained the best results in terms of rank-1 IRs (i.e., 99.07%, 99.07%, 94.39%, and 94.86%) on the four test protocols compared with the state-of-the-art methods as indicated in Table 7. Note that the proposed approach obtained a rank-1 IR 5% higher than the best reported result. We also obtained the highest 0.1% FAR VRs (i.e., 96.26%, 96.26%, 93.46%, and 93.46%). Fig. 12 displays the ROC and CMC curves on the 3D-TEC dataset for all four experiments. It can be observed from Table 7 that the robustness of the proposed deep facial features is effective even in distinguishing twins.

4.2. Analysis of modules

In this section, we analyze the effect of each module on the proposed framework, including facial component patches scheme,

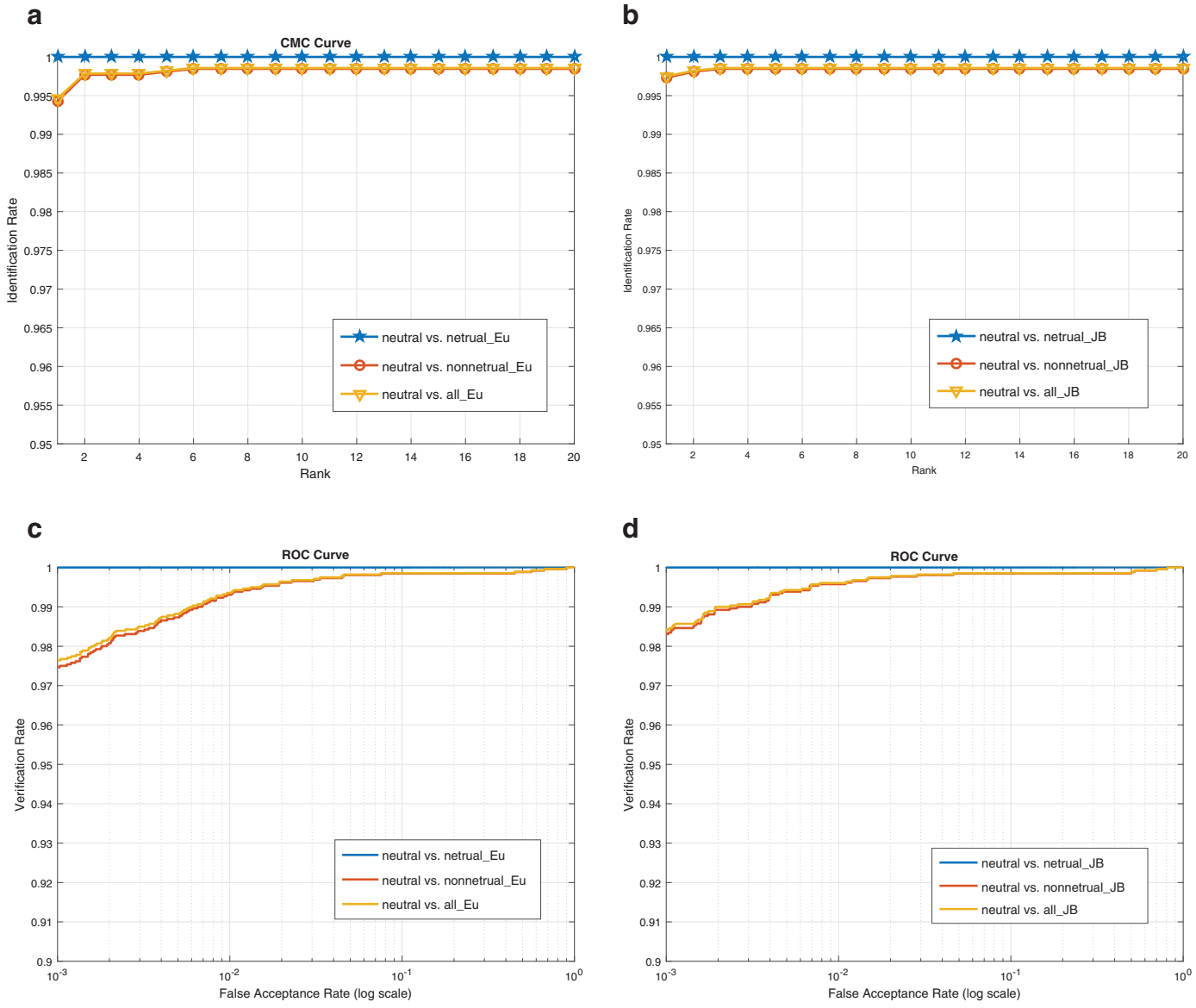


Fig. 10. CMC and ROC curves of "neutral vs. neutral", "neutral vs. Non-neutral", and "neutral vs. all" experiments on Bosphorus dataset. (a) CMC curves based on Euclidean distance. (b) CMC curves based on Joint Bayesian. (c) ROC curves based on Euclidean distance. (d) ROC curves based on Joint Bayesian.

multiple augmentation method, facial preprocessing, the proposed modified supervision, and different Pre-ResNet architectures. Excepting the first sub-chapter 4.2.1, all experiments conducted in this subsection were based on the "neutral vs. all" experiment using ours$^{eu}_{suhfp}$ method on the FRGC V2.0 dataset (mentioned in Section 4.1.1), namely, the single SUHFP patch was used for feature extraction and the Euclidean distance was used for classification.

4.2.1. Key role of facial component patches

We compared the performance of the four facial component patches in the "neutral vs. all" experiment on Bosphorus, BU-3DFE, and "Case I" experiment on 3D-TEC. The accuracies are displayed in Table 8 and the ROC and CMC curves of the three experiments are presented in Fig. 8. From these tables and figures we can observe that the performance of UHFP and SUHFP always outperformed FFP and NTP.

Based on the above results, two observations can be drawn. Firstly, the lower half of a face can be significantly affected by expressions and the up half can be considered as a rigid part against expression. Consequently, the FFP obtained the worst performance. Secondly, the nose area is less affected by expression. However, the eye and brow area of a face can provide more discrimination information. Therefore, the combination of these semi-rigid areas (nose, eyes, and brow)is superior to only the nose area and entire face area (containing lower half of a face) in terms of FR. In particular, the SUHFP indicated excellent and stable discrimination ability, achieving 100% rank-1 IRs and 100% VRs at 0.1% FAR in FRGC V2.0 with only one patch.

Although we have chosen to use the combination of all the four patches to perform face recognition, it is interesting to further investigate the contribution of each of the patches to this combination. Four sessions of "neutral vs. all" experiments on the BU-3DFE dataset were conducted; each session had one patch absent. The results are presented in Table 9 and Fig. 13 presents the ROC and CMC curves. From Table 9 and Fig. 14, although the FFP patch contributes less when a single patch is used, the FFP patch significantly contributed to the combination scheme, namely the recognition accuracy decreased the most when FFP was absent.

It is worth mentioning that although the probe set (214 images) in the twins experiment was considerably smaller than the other

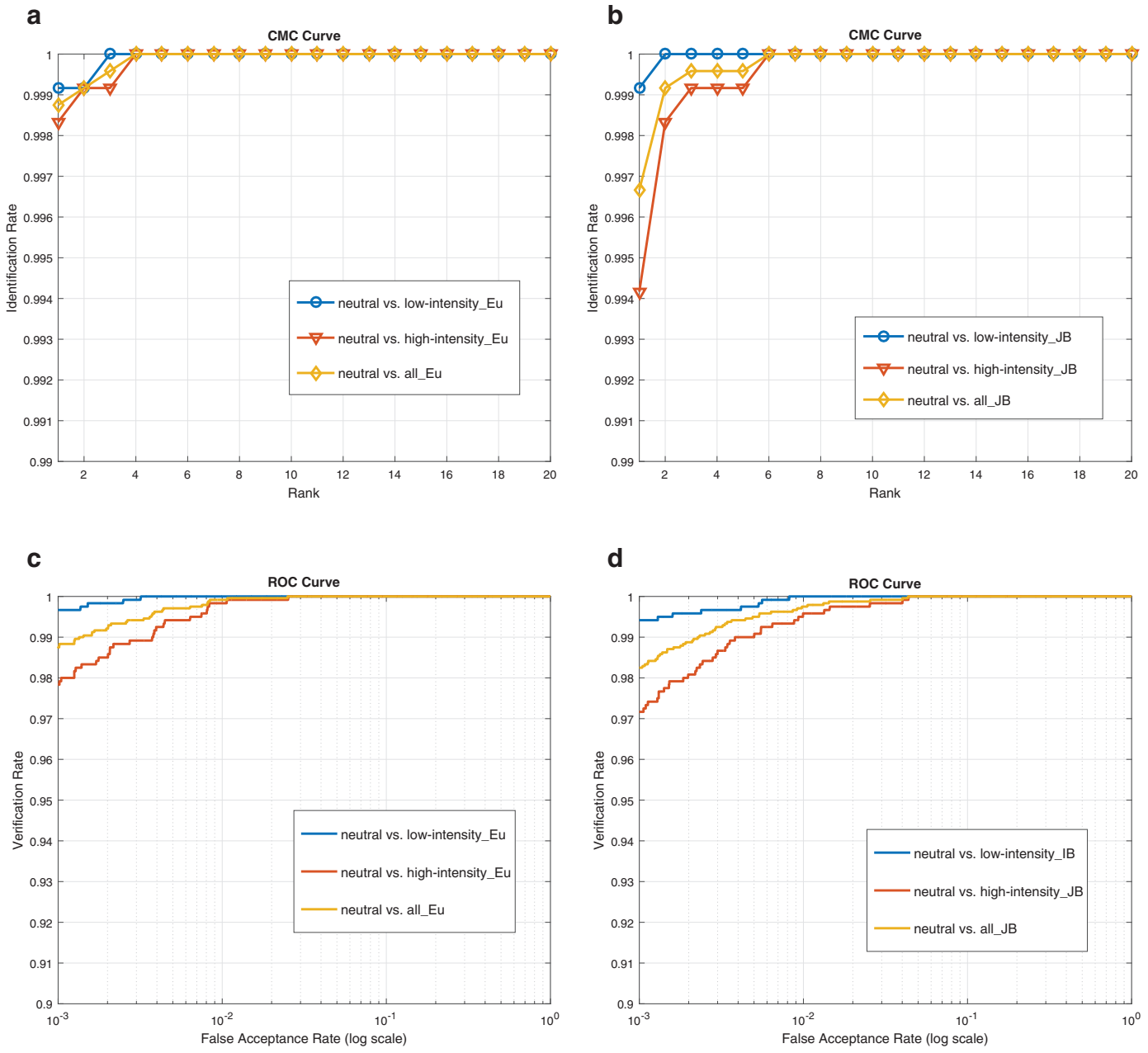


Fig. 11. CMC and ROC curves of "neutral vs. low-intensit", "neutral vs. high-intensity", and "neutral vs. all" experiments on the BU-3DFE dataset. (a) CMC curves of three experiments using Euclidean distance. (b) CMC curves of three experiments using Joint Bayesian. (c) ROC curves of three experiments using Euclidean distance. (d) ROC curves of three experiments using Joint Bayesian.

dataset, from Table 8(c) we can observe that the rank-1 IRs of the Case I experiment on 3D-TEC based on Euclidean distance using UHFP, as using ALL patches, was 98.13%. Namely, four probes were incorrectly recognized based on Euclidean distance. Fig. 15 displays the four incorrect probe images. Each row contains four images of one pair of twins. The first and second belong to one individual and the third and fourth belong to another. The first and third with neutral expression are from the gallery and the second and fourth with the smile expression are probes in the Case I testing protocol. In every row, the image marked with the red circle was incorrectly recognized as the image with blue circle mark; the image with the orange circle mark was the correct gallery image. Finally, the two rows beginning with the red bars are the mistaken probes based on Joint Bayesian. Fig. 15 indicates that such twin probes are extremely difficult to distinguish, even for the human eye.

4.2.2. Analysis of contribution of data augmentation and low-resolution training data

In this section, we perform two sets of comparative experiments to evaluate the advantage of the proposed multiple data augmentation and low-resolution training data.

In the first series of experiments, we compared the change of performance on "FRGC_NVALL_ours$^{eu}_{suhfp}$" experiments using different training data. (1) Removing the 3D pose augmentation from training data. (2) removing the range image transformational augmentation from training data. (3) removing the resolution augmentation from the training data. (4) removing the 4k low-resolution data from the training data. A comparison of the results with "FRGC_NVALL_ours$^{eu}_{suhfp}$" in Section 4.1.1 using training data with all augmentation is presented in Table 10(a). The CMC and ROC curve are presented in Fig. 16.

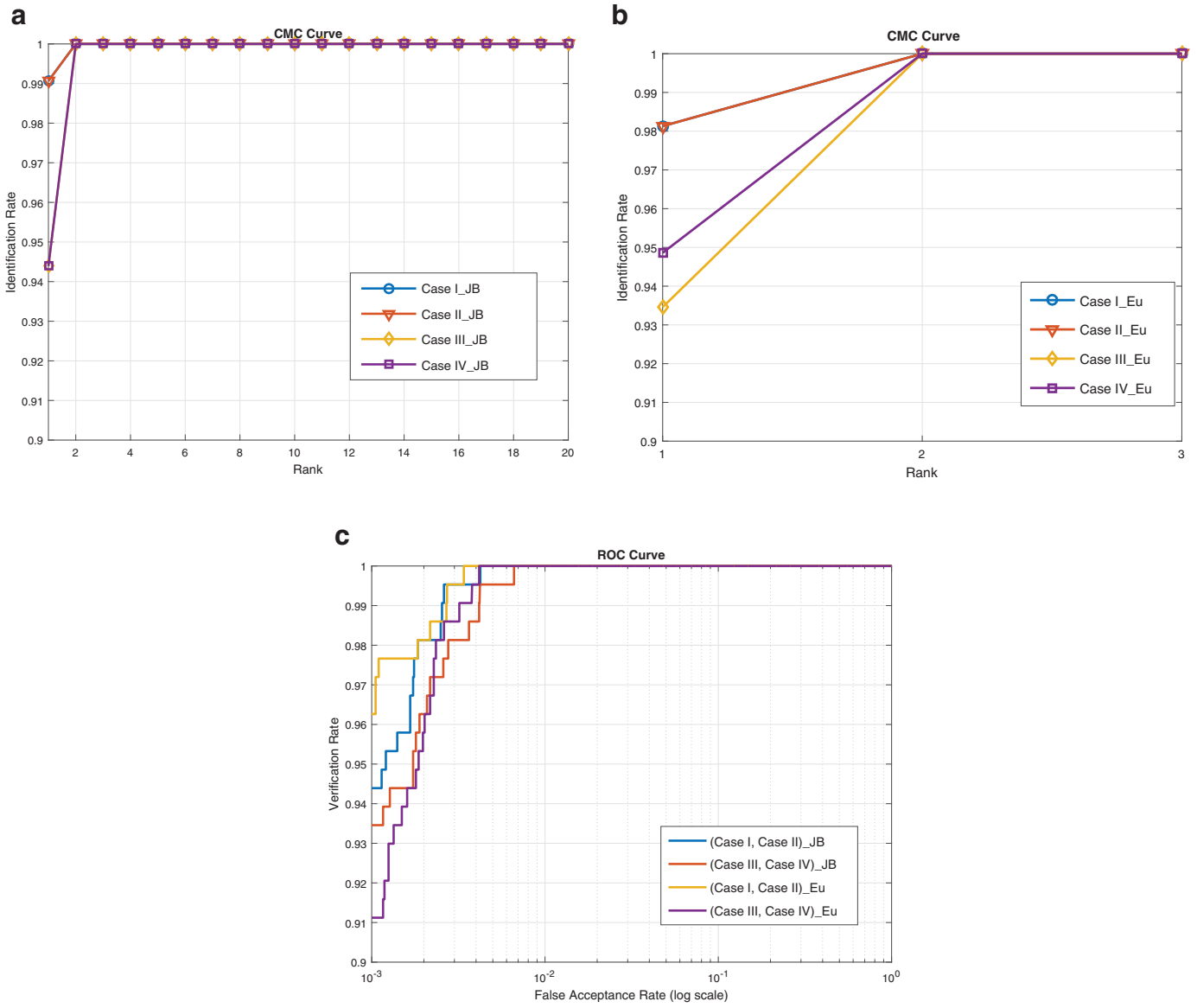


Fig. 12. CMC and ROC curves on 3D-TEC dataset. (a) CMC curves on Case I, II, III, and IV using Joint Bayesian. (b) CMC curves on Case I, II, III, and IV using Euclidean distance. (c) ROC curves on (Case I and II) and (Case III and IV) using Joint Bayesian and Euclidean distance, respectively.

Because the resolution of the BU-3DFE dataset is less than the other test datasets, in the second series of experiments, we aimed to spotlight the contribution of resolution augmentation and low-resolution training data. We compared the BU-3DFE "neutral vs. all" Ours$^{eu}_{allp}$ (mentioned in Section 4.1.3) results with it of experiments removing the resolution augmentation or low-resolution data. Table 10(b) displays the comparison results.

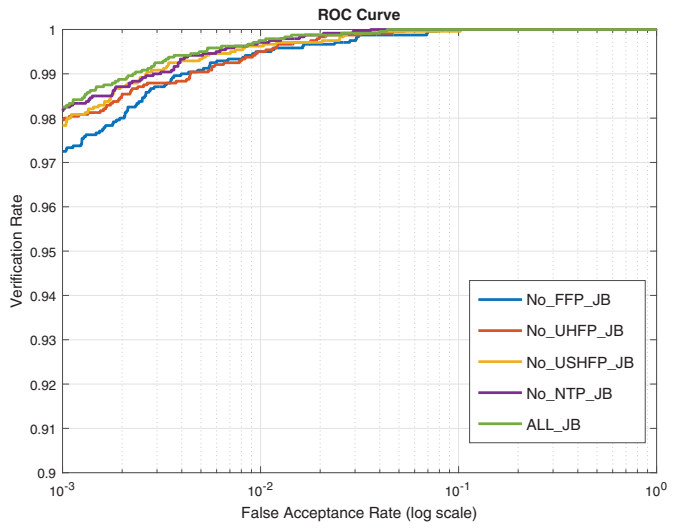
From the second row of Table 10(a) and the third row of Table 10(b), we can see that the accuracy of both 0.1% FAR VR and rank-1 IR decreased by approximately 4%. The reason for this decline is that removing the range image transformational augmentation or the 4k low-resolution data from the limited training data can easily result in rapid overfitting. Clearly, the advantage of resolution augmentation is remarkable in low-resolution dataset. The two compared tables both indicate that the "all augmentation" tactics are contributing and the low-resolution training data cannot be absent.

4.2.3. Performance analysis of preprocessing technology

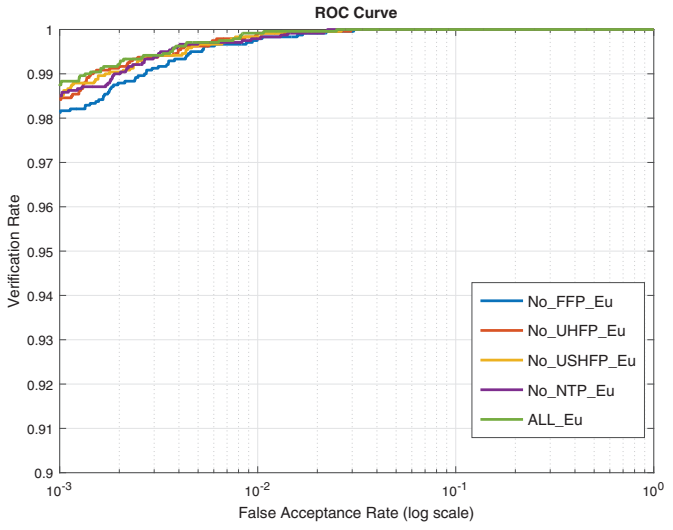
In this subsection, we performed two "FRGC_NVALL_ours$^{eu}_{suhfp}$" experiments using different 3D scan preprocessing for the test range images, aiming to evaluate the fast PCA pose correction and proposed nose-tip relocation method. There are three 3D scan preprocessing schemes: (1) both pose and nose-tip relocation, performed in Section 4.1.1, (2) without nose-tip relocating, and (3) without both pose and nose-tip relocation, represented as "Both_NT_P", "No_NT", and "Both_No_NT_P", respectively.

The rank-1 IRs decreased 0.87% and 3.05%; 0.1% FAR VRs decreased 0.9% and 3.11% using "No_NT" and "Both_No_NT_P" preprocessing schemes, respectively, compared with "Both_NT_P". Table 11 displays the results of the three preprocessing schemes; the CMC and ROC curve are presented in Fig. 17.

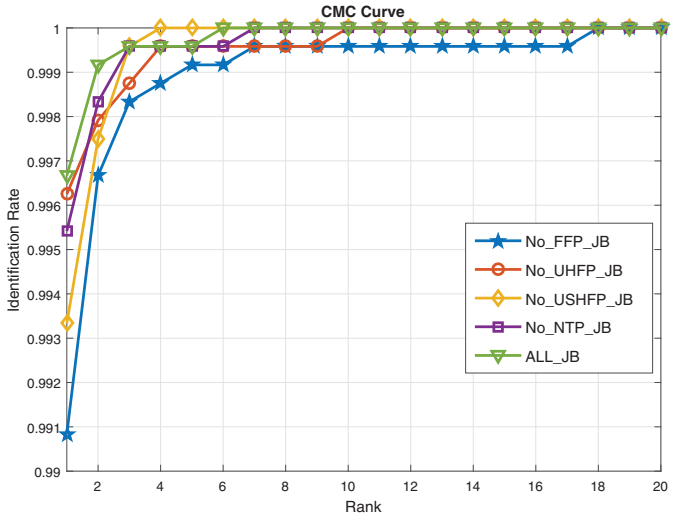
One observation can be made from these results. Although the preprocessing methods are simple and fast, they are beneficial to FR accuracy. This is because preprocessing is helpful for alignment



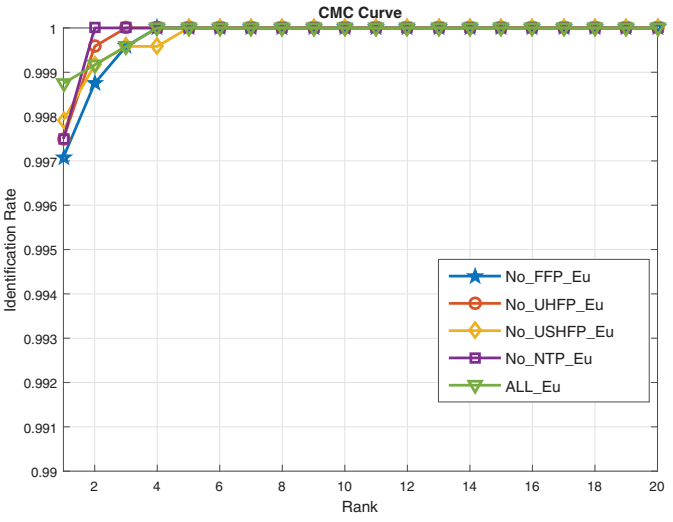
(a) ROC-BU-3DFE-NVALL-JB



(b) ROC-BU-3DFE-NVALL-EU



(c) CMC-BU-3DFE-NVALL-JB



(d) CMC-BU-3DFE-NVALL-EU

Fig. 13. ROC and CMC curves of one-patch-absent experiment on BU-3DFE.

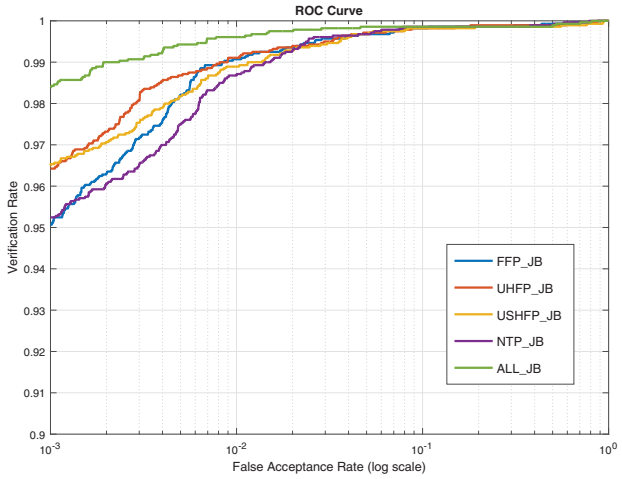
and normative range image cropping. Fig. 18 presents typical range images preprocessed by different modes, the first row of Fig. 18(a) displays the range images preprocessed using “Both_NT_P” and the second row displays the range images preprocessed using “No_NT”; the two rows correspond one to one. In the same manner, the range images in the first row of Fig. 18(b) are preprocessed using “Both_NT_P” and the range images in the second row are the corresponding images preprocessed using “Both_No_NT_P”. From Fig. 18, it is intuitionistic that the proposed preprocessing can reduce the difficulty of recognition.

4.2.4. Improvement of supervision method

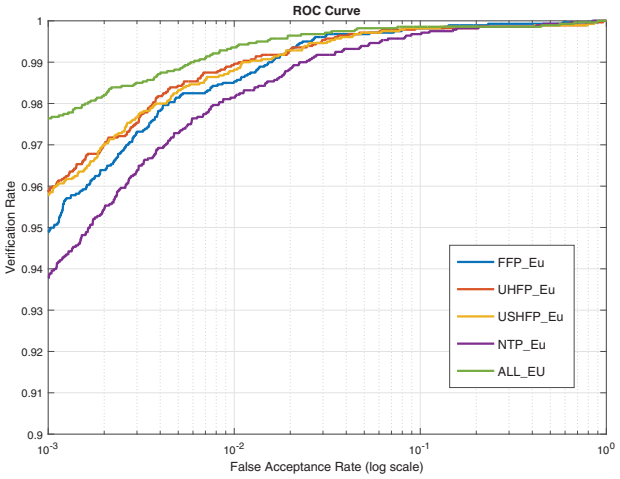
Center Loss [66] achieves a best performance on several 2D FR benchmarks; specially under the protocol of a small training set, under 20,000 individuals. To evaluate the proposed modified multiscale Triplet Loss, in this section we compared our performance with the Center Loss and traditional Triplet Loss based on the FRGC_NVALL_ours$^{eu}_{suhfp}$ experiment. We obtained 98.24% rank-1 IR and 98.01% VR at 0.1% FAR using the Center Loss as supervision; 98.12% rank-1 IR and 98.28% VR at 0.1% FAR was achieved using the traditional Triplet Loss supervision. A comparison of the results of the three different supervisions are presented in Table 12. The results are mainly caused by the fact that Center Loss suffers from a setback because many individuals in the training data have only a single real 3D model, and the instability of traditional Triplet Loss.

4.2.5. Comparison of the performance among three Pre-ResNet architectures

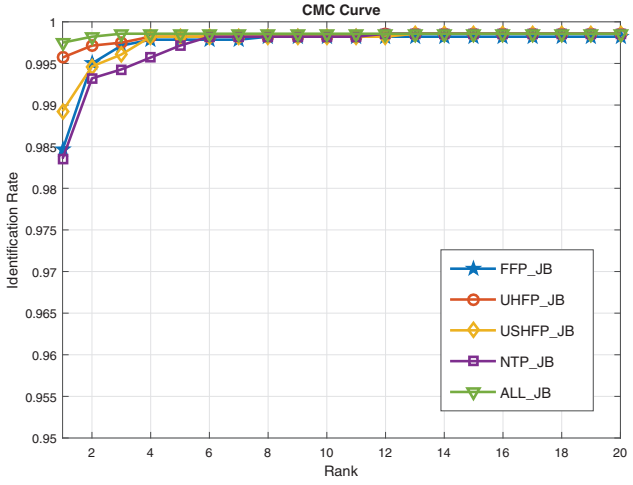
In this section, we extracted the features from the three Pre-ResNets (Pre-ResNet-14, Pre-ResNet-24, and Pre-ResNet-34 mentioned in Section 3.3) trained under the same conditions and compared their performance based on the “FRGC_ NVALL_ours$^{eu}_{suhfp}$” experiment aiming to determine a tradeoff between computational efficiency and recognition accuracy. Empirically, the Pre-ResNet-24 is more effective and demonstrates approximately a 2% improvement both on 0.1% FAR VR and rank-1 IR compared with the other two networks; the time cost in training and feature extraction processing are both satisfactory. The comparison of recognition accuracy are presented in Table 13. Further, we present a comparison of the iteration numbers when the best performance appears



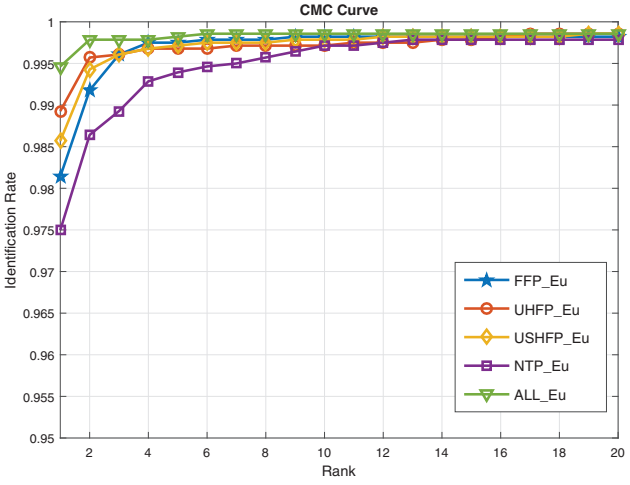
(a) ROC-Bosphorus-NVALL-JB



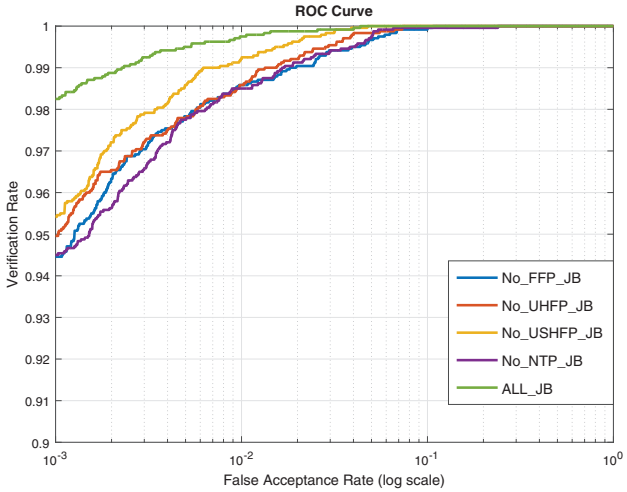
(b) ROC-Bosphorus-NVALL-EU



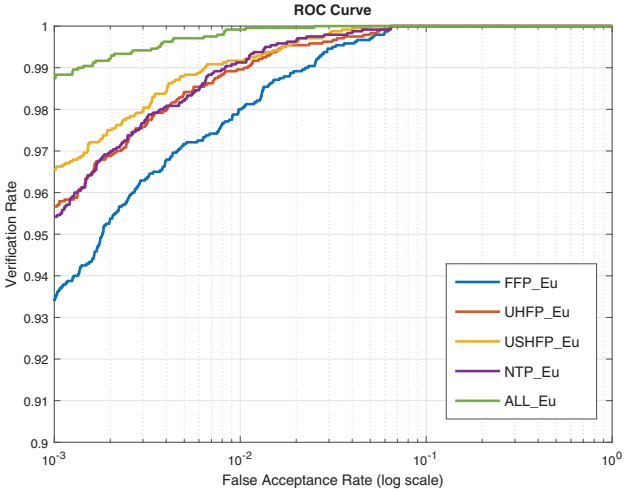
(c) CMC-Bosphorus-NVALL-JB



(d) CMC-Bosphorus-NVALL-EU

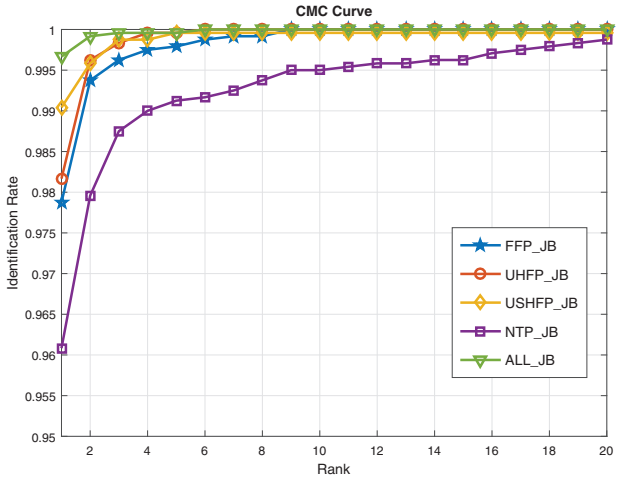


(e) ROC-BU-3DFE-NVALL-JB

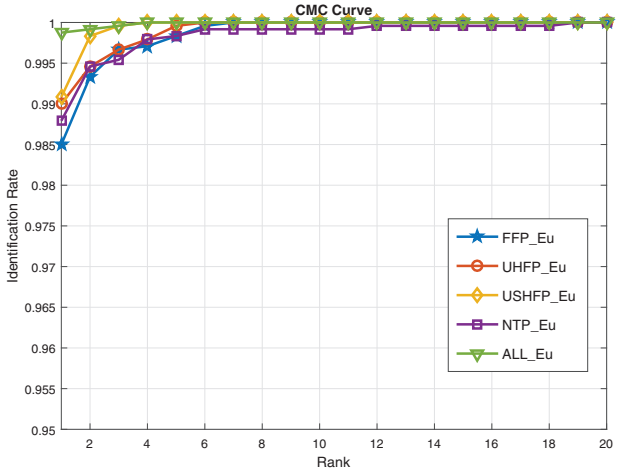


(f) ROC-BU-3DFE-NVALL-EU

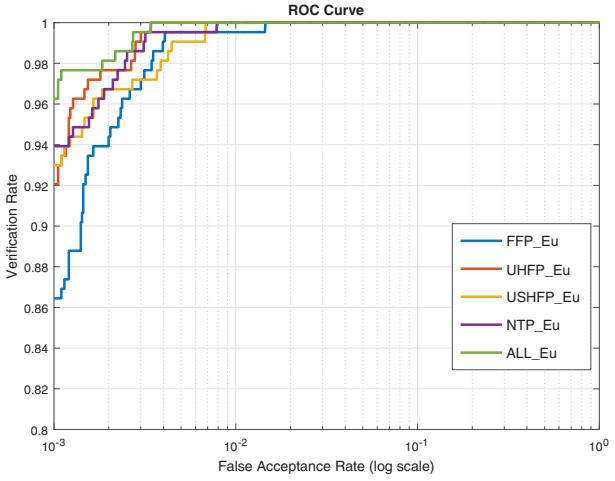
Fig. 14. ROC and CMC curves of experiments regarding performance comparison with each patch and all combination form; caption of each sub-figure consist of "curve type-dataset name-test protocol-classier type".



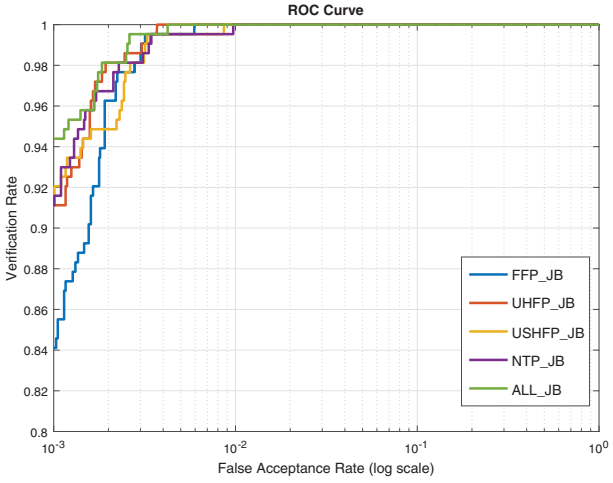
(g) CMC-BU-3DFE-NVALL-JB



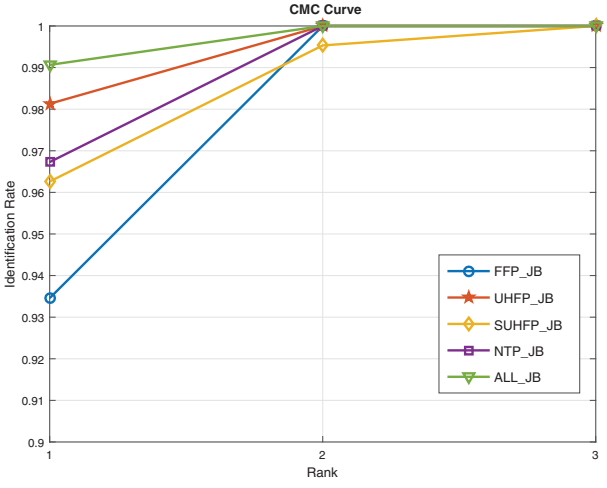
(h) CMC-BU-3DFE-NVALL-EU



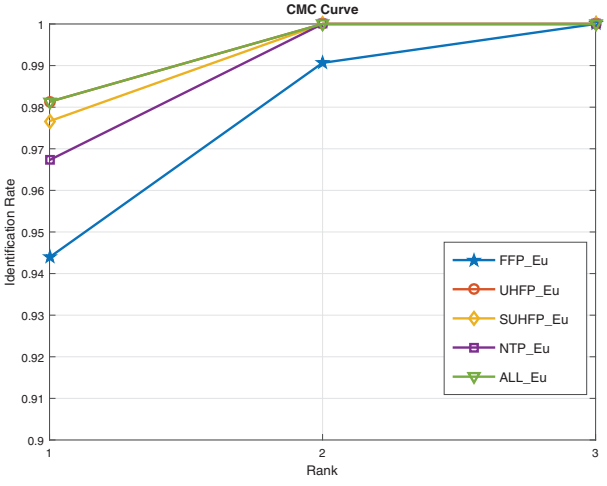
(i) ROC-3D-TEC-CASEI-JB



(j) ROC-3D-TEC-CASEI-EU



(k) CMC-3D-TEC-CASEI-JB



(l) CMC-3D-TEC-CASEI-EU

Fig. 14. Continued

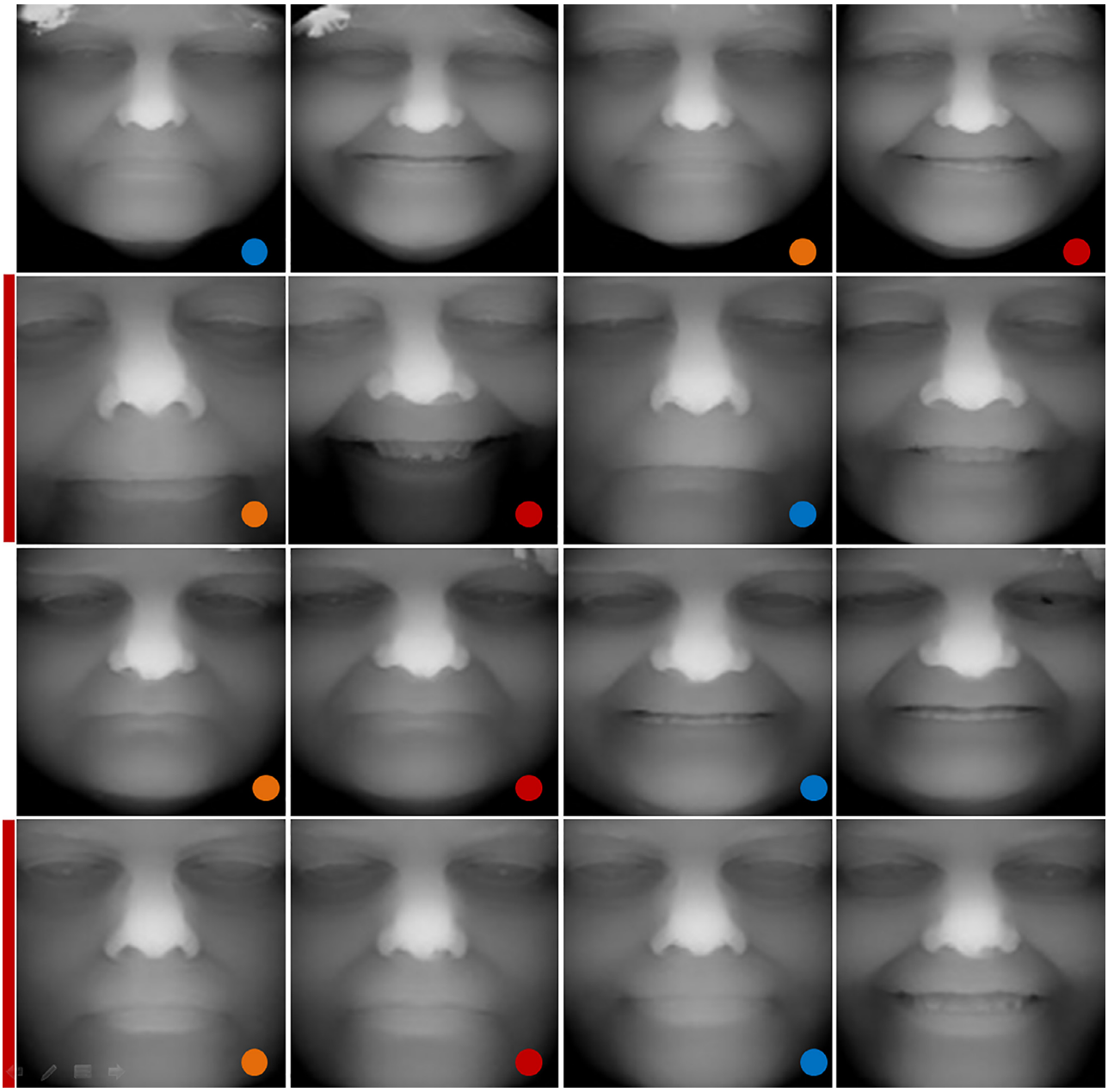


Fig. 15. Range images of 3D-TEC incorrectly identified.

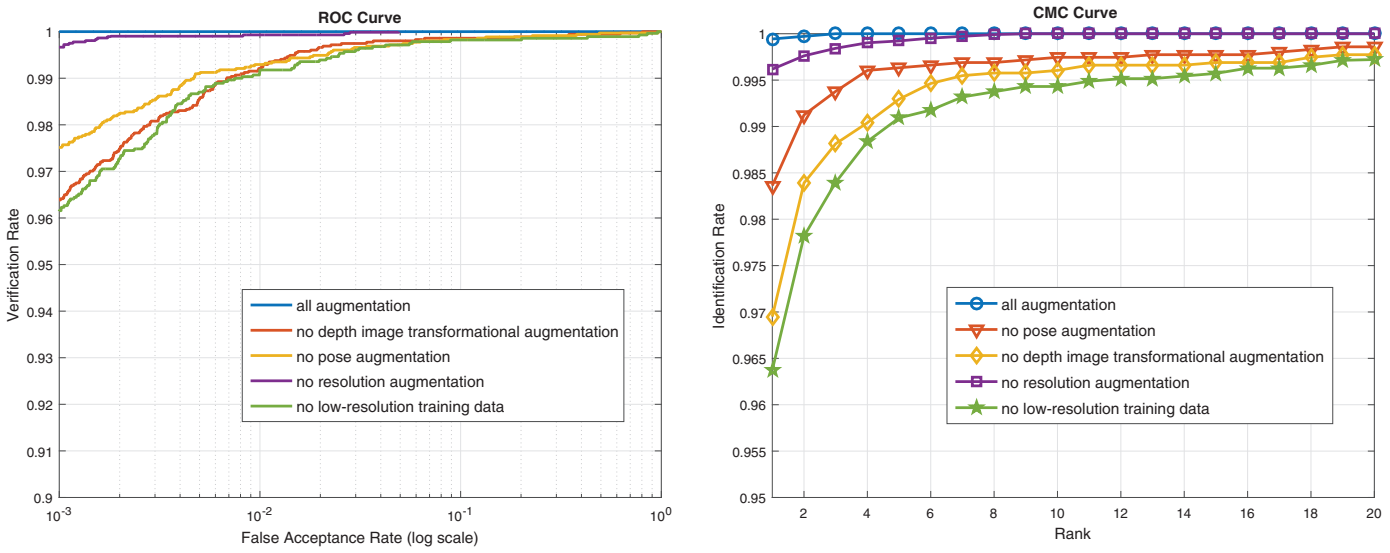
(denoted as: IterN-Best) to analyze the overfitting. One observation is that the Pre-ResNet-14 is not suitable for our case because of the limited layers and our training data are clearly not sufficient to support Pre-ResNet-34, which results in rapid overfitting. The CMC and ROC curves of these experiments are presented in Fig. 19.

4.3. Time cost analysis

Computational efficiency is an important concern for a real-world 3D FR system, which is one of the major advantages of the proposed approach. Table 14 displays the timing comparison between the proposed approach and several top-ranked 3D face recognition approaches in terms of identification. Because our approach is an order of 3–50 times faster than the competing approaches from Table 14, it is considered as one of the fastest approaches available. Furthermore, the proposed approach can easily profit from parallel processing (it currently runs in a single thread).

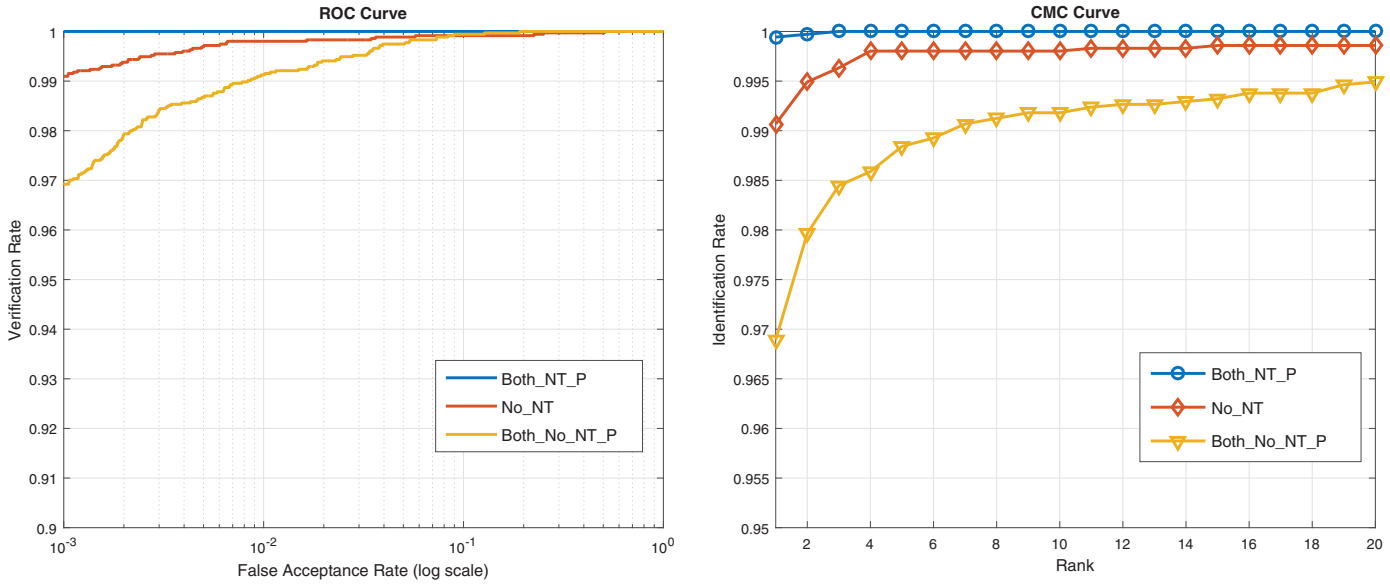
The timing experiments were conducted on a PC with 3.2 GHz Intel Core processors and a NVIDIA GTX750 GPU for training and testing using Caffe implementation [89]. The literature [20] evaluated their method on a PC with 2.6 GHz dual-processors and an NVIDIA K40 GPU. Our CPU performance is marginally superior to the others; however, our GPU performance is slightly weaker than [20]; The calculational efficiencies of proposed method and [20] are comparable.

The two performed experiments were “neutral vs. all” experiment on the FRGC V2.0 dataset based on Euclidean distance and Joint Bayesian. To analyze the computational complexity of the propose approach, we measured two computational costs including preprocessing time (the time cost of a probe scan from raw data to features) and matching time (consumed by features matching). The facial data processing step required 0.76 s of which the most time-consuming part was the PCA pose correction and interpolation. During the features matching, the proposed approach



(a) ROC curves (b) CMC curves

Fig. 16. CMC and ROC curve of data augmentation experiments in "FRGC_NVALL_ours$^{eu}_{suhfp}$" experiments.



(a) ROC curves (b) CMC curves

Fig. 17. CMC and ROC curve of three different preprocessing modes in "FRGC_NVALL_ours$^{eu}_{suhfp}$" experiments.

required 0.08 s based on Euclidean distance to identify each probe scan when the gallery size was 466 faces. The dimension of the mapping matrix of the Joint Bayesian model was set to 100 in this paper. We provide a comparison with the state-of-the-art methods in descending order of computational efficiency.

Spreeuwers [90] proposed a computationally efficient 3D FR approach. Their approach required 2.54 s to identify a probe in a gallery of 466 faces. Their approach is three times slower than the proposed approach, which remains the fastest among the existing approaches Specifically, they can conduct 11,150 ~ 22,300 comparisons per second consisting of matrix multiplications in the feature matching, and 2.5 s were consumed on registration in the preprocessing. Note that the literature [20] based on a deep learning technique achieved the second highest recognition results in the majority of the experiments, which required approximately 3 s to register a probe because of the rigid-ICP in preprocessing. Conversely, the proposed method required only 0.76 s for the alignment based on only three facial landmarks. Wang et al. [17], based on boosting technology, proposed a relatively computational efficient approach that required 0.66 s for matching and 2.64 s for preprocessing. Li et al. [82] introduced an approach that required 3.55 s to identify a probe in a gallery of 466 faces. However, they leveraged an optimization (l_0 minimization) where the computation time depends on the size of the gallery. Their time cost would be difficult to estimate when the gallery size increases. Lei et al. [12] proposed a 3D PFR approach using local Keypoint-based Multiple Triangle Statistics. They required 6.67 s in preprocessing (including face normalization, local feature extraction, and first-phase

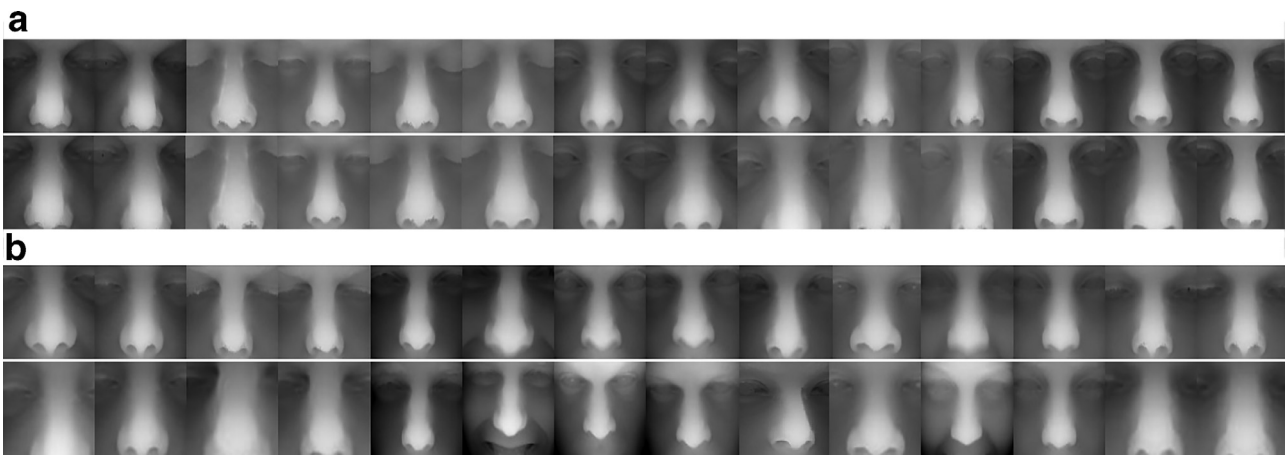


Fig. 18. Typical examples of range images preprocessed by different modes. (a) Comparison between "Both_NT_P" and "No_NT". (b) Comparison between "Both_NT_P" and "Both_No_NT_P".

Table 8
Comparison of performance of four face patches and their fusion form.

(a) NVALL comparing experiments of each patch on Bosphorus

Patch name	IRs-JB	IRs-EU	VRs-JB	VRs-EU
FFP	98.46%	98.14%	95.06%	94.89%
UHFP	99.57%	98.93%	96.42%	95.85%
USHFP	98.93%	98.57%	96.49%	95.78%
NTP	98.36%	97.50%	95.24%	93.78%
ALL	**99.75**%	99.46%	**98.39**%	97.64%

(b) NVALL comparing experiments of each patch on BU-3DFE

Patch name	IRs-JB	IRs-EU	VRs-JB	VRs-EU
FFP	97.88%	98.50%	94.46%	93.38%
UHFP	98.17%	99.00%	94.96%	95.67%
USHFP	99.04%	99.08%	95.42%	96.54%
NTP	96.08%	98.79%	94.5%	95.42%
ALL	99.67%	**99.88**%	98.25%	**98.71**%

(c) Case I comparing experiments of each patch on 3D-TEC

Patch name	IRs-JB	IRs-EU	VRs-JB	VRs-EU
FFP	93.45%	94.39%	84.11%	86.45%
UHFP	98.13%	98.13%	91.12%	92.06%
USHFP	96.26%	97.66%	91.59%	92.99%
NTP	96.73%	96.72%	91.12%	93.93%
ALL	**99.07**%	98.13%	94.39%	**96.26**%

Table 9
NVALL experiments with one patch absence on BU-3DFE .

Patch combination	IRs-JB	IRs-EU	VRs-JB	VRs-EU
No FFP	99.08%	99.71%	98.13%	97.25%
No UHFP	99.63%	99.75%	98.42%	97.96%
No USHFP	99.33%	99.79%	98.50%	97.83%
No NTP	99.54%	99.75%	98.50%	98.17%
No absence	99.67%	99.88%	98.25%	98.71%

Table 10
Comparison of data augmentation experiments. (a) Comparison about the advantage of the multiple data augmentation on FRGC_NVALL_ours$^{eu}_{suhfp}$ experiments. (b) Comparison with the absence of resolution augmentation and low-resolution training data on BU-3DFE "neutral vs. all" experiment using Ours$^{eu}_{allp}$.

(a)

Augmentation mode	IRs	0.1% FAR VRs
All augmentation. FRGC_NVALL_ours$^{eu}_{suhfp}$	99.94%	100%
No transformational augmentation. FRGC_NVALL_ours$^{eu}_{suhfp}$	96.39%	96.95%
No 3D pose augmentation. FRGC_NVALL_ours$^{eu}_{suhfp}$	98.36%	97.51%
No resolution augmentation. FRGC_NVALL_ours$^{eu}_{suhfp}$	99.62%	99.67%
No low-resolution training data. FRGC_NVALL_ours$^{eu}_{suhfp}$	96.37%	96.12%

(b)

Augmentation mode	"neutral vs. all" IRs	"neutral vs. all" 0.1% FAR VRs
All augmentation. Ours$^{eu}_{allp}$	99.88%	98.71%
No resolution augmentation. Ours$^{eu}_{allp}$	99.21%	98.01%
No low-resolution training data. Ours$^{eu}_{allp}$	95.16%	94.63%

Table 11
Comparison of three preprocessing modes in "FRGC_NVALL_ours$^{eu}_{suhfp}$" experiments.

Preprocessing modes	n vs. all IRs	n vs. all VRs
Both_NT_P. FRGC_NVALL_ours$^{eu}_{suhfp}$	99.94%	100%
No_NT. FRGC_NVALL_ours$^{eu}_{suhfp}$	99.07%	99.10%
Both_No_NT_P. FRGC_NVALL_ours$^{eu}_{suhfp}$	96.89%	96.89%

Table 12
Comparison of proposed multi-scale Triplet Loss, traditional Triplet Loss,and Center Loss on the FRGC V2.0 "neutral vs. all" experiment.

Supervision loss	IRs	VRs
Center Loss. FRGC_NVALL_ours$^{eu}_{suhfp}$	98.24%	98.01%
Traditional triplet loss. FRGC_NVALL_ours$^{eu}_{suhfp}$	98.12%	98.28%
Modified Multi-scale triplet loss. FRGC_NVALL_ours$^{eu}_{suhfp}$	99.94%	100%

classification) and 1.82 s to match a probe in a gallery of 466 faces. Kakadiaris et al. [74] required 15 s for facial data preprocessing and they can perform 1000 one-to-one face matchings per second. Aly et al. [91] adopted LDA to reduce the dimension of the features, which resulted in a time cost of distance calculation less than 3 ms per probe for the entire FRGC v2.0 gallery (466 faces). This was the best reported existing results in terms of matching. However, they require 36 s in preprocessing (including detecting landmarks, registration to the AvFM, and ICP registrations). Other algorithms such as [19,25,76,78,92] require more than 150 s to match a probe when the size of the gallery is 466 faces. The above analysis indicates that the majority of the existing 3D FR approaches are subject to improvement of the computational efficiency when considered for real-world applications.

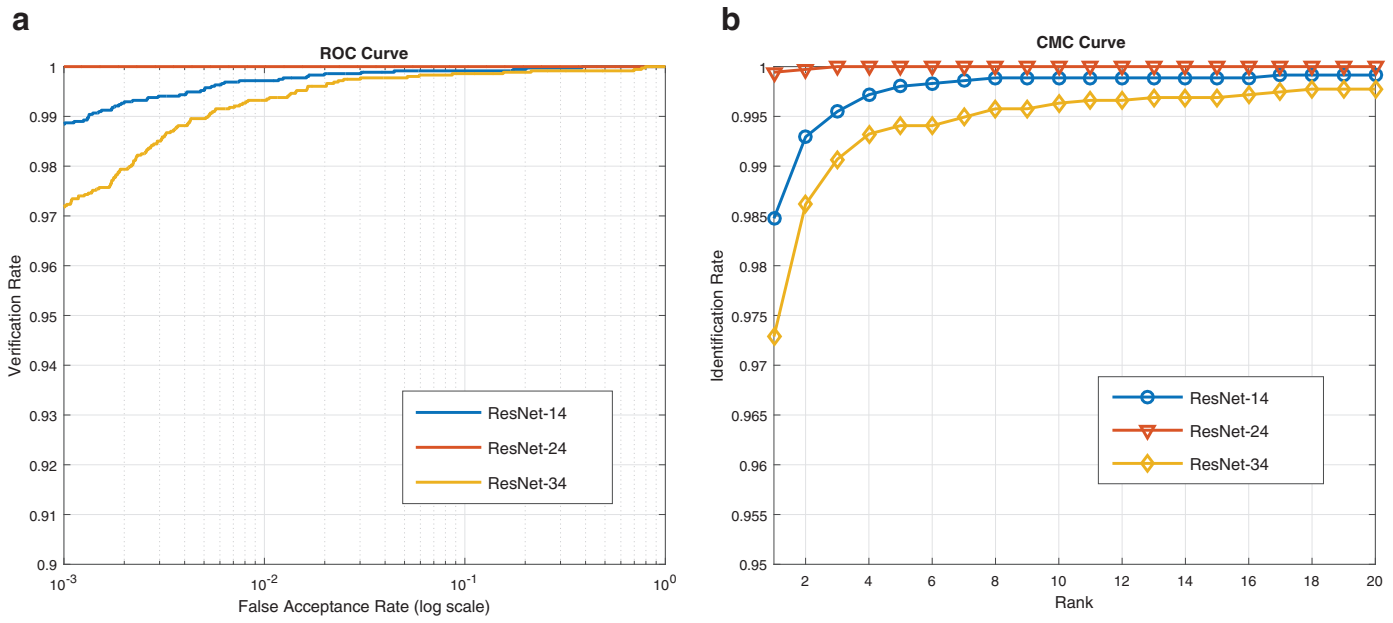


Fig. 19. CMC and ROC curves of network architectures compared on FRGC V2.0 dataset "neutral vs. all" experiments. (a) ROC curves. (b) CMC curves.

Table 13
Comparison of Pre-ResNet-14, Pre-ResNet-24, and Pre-ResNet-34 on the FRGC V2.0 "neutral vs. all" experiment.

Network	IRs	0.1% FAR VRs	IterN-Best
Pre-ResNet-14. FRGC_NVALL_ours$^{eu}_{suhfp}$	98.47%	98.84%	276500
Pre-ResNet-24. FRGC_NVALL_ours$^{eu}_{suhfp}$	99.94%	100%	99500
Pre-ResNet-34. FRGC_NVALL_ours$^{eu}_{suhfp}$	97.29%	97.17%	45000

Table 14
Comparison of computation time (s) of one scan feature extraction and matching probe scan against a gallery of 466 faces.

Approaches	Processing	Matching	Total
Kakadiaris et al. [74] (2007)	15	0.5	15.5
Aly et al. [91] (2010)	36	**0.02**	36.02
Wang et al. [17] (2010)	2.64	0.66	3.3
Spreeuwers et al. [90] (2011)	2.5	0.04	2.54
Li et al. [82] (2014)	3.05	0.5	3.55
Li et al. [14] (2015)*	69.5	5.5	75
Lei et al. [12] (2016)	6.67	1.82	8.49
Kim et al. [20] (2017)	3.16	0.09	3.25
FRGC_NVALL_ours$^{eu}_{suhfp}$	**0.76**	0.08	**0.84**
FRGC_NVALL_ours$^{jb}_{suhfp}$	**0.76**	0.16	0.92

*The gallery size for computation times is 105.

5. Conclusions

A novel systematic work based on deep learning techniques for 3D FR was proposed in this work. It can significantly improve the computational efficiency and recognition performance compared to the state-of-the-art methods. Specifically, the proposed approach requires only 0.84 s to identify a 3D face with a gallery of 466 faces, which is three times faster than the best-reported existing results. Moreover, the highest rank-1 IRs (99.75% and 99.07%) and 0.1% FAR VRs (98.92% and 96.26%) were achieved, respectively, on the Bosphorus and 3D-TEC datasets. Three observations can be drawn based on the experimental results.

Firstly, the proposed approach is computationally efficient and can be easily applied in real-world systems. Secondly, an optimized network and supervision technology of deep learning can significantly improve the robustness of the 3D features and extraction efficiency. Lastly, combining multiple data augmentation and additional facial data collection is a reasonable solution addressing the training data size problem of 3D FR.

In the future, a more effective facial pose augmentation technique can be developed to address large pose variations. Furthermore, a self-established 3D facial dataset will soon be available to the public, which can further boost research in the field of 3D FR.

Declarations of interest

None.

Acknowledgment

This work is supported by Chinese Universities Scientific Fund (Grant No. 2019NYB05), National Natural Science Foundation of China (Grant No. 61702350, 61402307, 61403265, 71774134, and 71373216), Sichuan Science and Technology Program (Grant No. 18YYJC1287, 2015SZ0226), Sichuan University (Grant No. 2018SCUH0042), National Key Research and Development Program of China (Grant No. 2016YFC0801100) and National Key Scientific Instrument and Equipment Development Project of China (Grant no. 2013YQ49087903).

References

[1] G.B. Huang, M. Mattar, T. Berg, E. Learned-Miller, Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments, Technical Report 07–49, University of Massachusetts, Amherst, 2008. 2007
[2] C. Creusot, N. Pears, J. Austin, A machine-learning approach to keypoint detection and landmarking on 3D meshes, Int. J. Comput. Vis. 102 (1–3) (2013) 146–179.
[3] S. Malassiotis, M.G. Strintzis, Robust real-time 3D head pose estimation from range data, Pattern Recognit. 38 (8) (2005) 1153–1165.
[4] Y. Cai, M. Yang, Z. Li, Robust head pose estimation using a 3D morphable model, Math. Probl. Eng. 2015 (5) (2015) 1–10.
[5] A. Danelakis, T. Theoharis, I. Pratikakis, P. Perakis, An effective methodology for dynamic 3D facial expression retrieval, Pattern Recognit. 52 (C) (2016) 174–185.
[6] X. Xiang, H.A. Le, P. Dou, Y. Wu, I.A. Kakadiaris, Evaluation of a 3D-aided pose invariant 2D face recognition system, in: Proceedings of the IEEE International Joint Conference on Biometrics, 2018.
[7] P. Koppen, Z. Feng, J. Kittler, M. Awais, W. Christmas, X. Wu, H. Yin, Gaussian mixture 3D morphable face model, Pattern Recognit. 74 (2018) 617–628.
[8] G. Hu, F. Yan, J. Kittler, W. Christmas, C.H. Chan, Z. Feng, P. Huber, Efficient 3D morphable face model fitting, Pattern Recognit. 67 (C) (2017) 366–379.

[9] K. Kollreider, H. Fronthaler, J. Bigun, Verifying liveness by multiple experts in face biometrics, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPRW '08, 2008, pp. 1–6.

[10] K.A. Nixon, V. Aimale, R.K. Rowe, Spoof detection schemes, in: Handbook of Biometrics, Springer, Boston, MA, 2008, pp. 403–423.

[11] X. Yu, Y. Gao, J. Zhou, Sparse 3D directional vertices vs continuous 3D curves: efficient 3D surface matching and its application for single model face recognition, Pattern Recognit. 65 (2017) 296–306.

[12] Y. Lei, Y. Guo, M. Hayat, M. Bennamoun, X. Zhou, A two-phase weighted collaborative representation for 3D partial face recognition with single sample, Pattern Recognit. 52 (C) (2016) 218–237.

[13] S. Berretti, N. Werghi, A.D. Bimbo, P. Pala, Selecting stable keypoints and local descriptors for person identification using 3D face scans, Vis. Comput. 30 (11) (2014) 1275–1292.

[14] H. Li, D. Huang, J.M. Morvan, Y. Wang, L. Chen, Towards 3D face recognition in the real: a registration-free approach using fine-grained matching of 3D keypoint descriptors, Int. J. Comput. Vis. 113 (2) (2015) 128–142.

[15] F. Al-Osaimi, M. Bennamoun, A. Mian, An expression deformation approach to non-rigid 3D face recognition, Int. J. Comput. Vis. 81 (3) (2009) 302–316.

[16] Y. Wang, C.S. Chua, Y.K. Ho, Facial feature detection and face recognition from 2D and 3D images, Pattern Recognit. Lett. 23 (10) (2002) 1191–1202.

[17] Y. Wang, J. Liu, X. Tang, Robust 3D face recognition by local shape difference boosting., IEEE Trans. Pattern Anal. Mach. Intell. 32 (10) (2010) 1858–1870.

[18] Y. Lei, M. Bennamoun, M. Hayat, Y. Guo, An efficient 3D face recognition approach using local geometrical signatures, Pattern Recognit. 47 (2) (2014) 509–524.

[19] L. Ballihi, B.B. Amor, M. Daoudi, A. Srivastava, D. Aboutajdine, Boosting 3-d–geometric features for efficient face recognition and gender classification, IEEE Trans. Inf. Forensics Secur. 7 (6) (2012) 1766–1779.

[20] D. Kim, M. Hernandez, J. Choi, G. Medioni, Deep 3D face identification, IEEE International Joint Conference on Biometrics (IJCB) (2017) 133–142.

[21] O.M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in: Proceedings of the British Machine Vision Conference, 1(3), 2015, p. 6.

[22] S. Zulqarnain Gilani, A. Mian, Learning from millions of 3d scans for large-scale 3d face recognition, IEEE Conference on Computer Vision and Pattern Recognition (2018) 1896–1905.

[23] D. Chen, X. Cao, L. Wang, F. Wen, J. Sun, Bayesian face revisited: a joint formulation, in: Proceedings of the European Conference on Computer Vision, 2012, pp. 566–579.

[24] P.J. Neugebauer, Reconstruction of real-world objects via simultaneous registration and robust combination of multiple range images, Int. J. Shape Model. 03 (01n02) (1997) 71–90.

[25] K.W.B. T. C. Faltemier, P.J. Flynn, A region ensemble for 3-d face recognition, 3, 2008, pp. 62–73.

[26] A. Mian, M. Bennamoun, R. Owens, An efficient multimodal 2D-3D hybrid approach to automatic face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 29 (11) (2007) 1927.

[27] H. Mohammadzade, D. Hatzinakos, Iterative closest normal point for 3D face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2) (2013) 381–397.

[28] P. Liu, Y. Wang, D. Huang, Z. Zhang, L. Chen, Learning the spherical harmonic features for 3D face recognition, IEEE Trans. Image Process. 22 (3) (2013) 914–925.

[29] S.Z. Gilani, A. Mian, P. Eastwood, Deep, dense and accurate 3D face correspondence for generating population specific deformable models, Pattern Recognit. 69 (2017) 238–250.

[30] S. Soltanpour, B. Boufama, Q.M.J. Wu, A survey of local feature methods for 3D face recognition, Pattern Recognit. 72 (2017) 391–406.

[31] Y. Lei, M. Bennamoun, A.A. El-Sallam, An efficient 3D face recognition approach based on the fusion of novel local low-level features, Pattern Recognit. 46 (1) (2013) 24–37.

[32] S. Elaiwat, M. Bennamoun, F. Boussaid, A. El-Sallam, A curvelet-based approach for textured 3D face recognition, Pattern Recognit. 48 (4) (2015) 1235–1246.

[33] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, J. Wan, N.M. Kwok, A comprehensive performance evaluation of 3D local feature descriptors, Int. J. Comput. Vis. 116 (1) (2016) 66–89.

[34] F.R. Al-Osaimi, M. Bennamoun, A. Mian, Integration of local and global geometrical cues for 3D face recognition, Pattern Recognit. 41 (3) (2008) 1030–1040.

[35] D. Huang, G. Zhang, M. Ardabilian, Y. Wang, 3D face recognition using distinctiveness enhanced facial representations and local feature hybrid matching, in: Proceedings of the IEEE International Conference on Biometrics: Theory Applications & Systems, 2010, pp. 1–7.

[36] X. Lu, A.K. Jain, D. Colbry, Matching 2.5d face scans to 3D models, IEEE Trans. Pattern Anal. Mach. Intell. 28 (1) (2006) 31–43.

[37] R.B. Rusu, S. Cousins, 3D is here: point cloud library (PCL), in: Proceedings of the IEEE International Conference on Robotics and Automation, 2011, pp. 1–4.

[38] M. Kaiser, X. Xu, B. Kwolek, S. Sural, G. Rigoll, Towards using covariance matrix pyramids as salient point descriptors in 3D point clouds, Neurocomputing 120 (10) (2013) 101–112.

[39] M. Draelos, Q. Qiu, A. Bronstein, G. Sapiro, Intel realsense = real low cost gaze, in: Proceedings of the IEEE International Conference on Image Processing, 2015, pp. 2520–2524.

[40] S. Bu, Z. Liu, J. Han, J. Wu, R. Ji, Learning high-level feature by deep belief networks for 3-d model retrieval and recognition, IEEE Trans. Multimed. 16 (8) (2014) 2154–2167.

[41] X. Jin, G. Dai, Z. Fan, E. Wong, F. Yi, Deepshape: deep-learned shape descriptor for 3D shape retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 39 (7) (2017) 1335–1345.

[42] B. Shi, S. Bai, Z. Zhou, X. Bai, Deeppano: deep panoramic representation for 3-d shape recognition, IEEE Signal Process. Lett. 22 (12) (2015) 2339–2343.

[43] A. Sinha, J. Bai, K. Ramani, Deep learning 3D shape surfaces using geometry images, in: Proceedings of the European Conference on Computer Vision, Springer International Publishing, 2016, pp. 223–240.

[44] E. Kalogerakis, M. Averkiou, S. Maji, S. Chaudhuri, 3d shape segmentation with projective convolutional networks, IEEE Conference on Computer Vision and Pattern Recognition (2017) 3779–3788.

[45] L.I. Huibin, J. Sun, X.U. Zongben, L. Chen, Multimodal 2D+3D facial expression recognition with deep fusion convolutional neural network, IEEE Trans. Multimed. 19(12) (2017) 2816–2831.

[46] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, J. Xiao, 3D ShapeNets: a deep representation for volumetric shapes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1912–1920.

[47] X. Xu, S. Todorovic, Beam search for learning a deep convolutional neural network of 3D shapes, in: Proceedings of the IEEE 23rd International Conference on Pattern Recognition (ICPR), 2016, pp. 3506–3511.

[48] Y. Li, S. Pirk, H. Su, C.R. Qi, L.J. Guibas, FPNN: Field probing neural networks for 3D data, in: Proceedings of the Advances in Neural Information Processing Systems., 2016, pp. 307–315..

[49] Z. Han, Z. Liu, J. Han, C.M. Vong, S. Bu, C.L. Chen, Mesh convolutional restricted Boltzmann machines for unsupervised learning of features with structure preservation on 3-D meshes., IEEE Trans. Neural Netw. Learn. Syst. 28 (10) (2016) 2268–2281.

[50] Z. Han, Z. Liu, J. Han, C.M. Vong, S. Bu, X. Li, Unsupervised 3D local feature learning by circle convolutional restricted Boltzmann machine, IEEE Trans. Image Process. 25 (11) (2016) 5331–5344.

[51] C.R. Qi, H. Su, K. Mo, L.J. Guibas, Pointnet: Deep learning on point sets for 3d classification and segmentation, IEEE Conference on Computer Vision and Pattern Recognition (2017) 652–660.

[52] Y. Taigman, M. Yang, Marc, L. Wolf, Deepface: Closing the gap to human-level performance in face verification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1701–1708.

[53] L. Wolf, T. Hassner, I. Maoz, Face recognition in unconstrained videos with matched background similarity, in: Proceedings of the Computer Vision and Pattern Recognition, 2011, pp. 529–534.

[54] Y. Sun, X. Wang, X. Tang, Deep learning face representation from predicting 10,000 classes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1891–1898.

[55] Y. Sun, X. Wang, X. Tang, Deep learning face representation by joint identification-verification, Adv. Neural Inf. Process. Syst. 27 (2014) 1988–1996.

[56] Y. Sun, X. Wang, X. Tang, Deeply learned face representations are sparse, selective, and robust, in: Proceedings of the Computer Vision and Pattern Recognition, 2015, pp. 2892–2900.

[57] Y. Sun, D. Liang, X. Wang, X. Tang, Deepid3: face recognition with very deep neural networks, arXiv:1502.00873 (2015).

[58] F. Schroff, D. Kalenichenko, J. Philbin, FaceNet: a unified embedding for face recognition and clustering (2015) 815–823.

[59] C. Lu, X. Tang, Surpassing human-level face verification performance on LFW with GaussianFace, in: Proceedings of the AAAI., 2015, pp. 3811–3819.

[60] J. Sivic, M. Everingham, A. Zisserman, Who are you? – Learning person specific classifiers from video, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2009, pp. 1145–1152.

[61] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions (2014) 1–9.

[62] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[63] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: Proceedings of the International Conference on Machine Learning., 2015, pp. 448–456.

[64] K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks, arXiv:1603.05027 (2016) 630–645.

[65] Y. Wen, Z. Li, Y. Qiao, Latent factor guided convolutional neural networks for age-invariant face recognition, in: Proceedings of the Computer Vision and Pattern Recognition, 2016, pp. 4893–4901.

[66] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, in: Proceedings of the European Conference on Computer Vision, 47, 2016, pp. 499–515.

[67] D. Miller, E. Brossard, S. Seitz, I. Kemelmachershlizerman, Megaface: a million faces for recognition at scale, arXiv:1505.02108 (2015).

[68] D.E. King, Dlib-ml: A Machine Learning Toolkit, JMLR.org, 2009. http://dlib.net/face_landmark_detection_ex.cpp.html.

[69] T.C. Faltemier, K.W. Bowyer, P.J. Flynn, Using a multi-instance enrollment representation to improve 3D face recognition, in: Proceedings of the IEEE International Conference on Biometrics: Theory, Applications, and Systems, 2007, pp. 1–6.

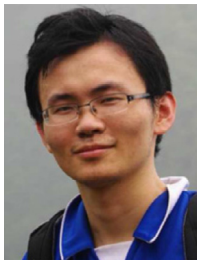
[70] J. Liu, Y. Deng, T. Bai, Z. Wei, C. Huang, Targeting ultimate accuracy: face recognition via deep embedding, arXiv:1506.07310 (2015).

[71] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W. Worek, Overview of the face recognition grand challenge, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, pp. 947–954.

[72] S. Berretti, A.D. Bimbo, P. Pala, 3D Face recognition using isogeodesic stripes, IEEE Trans. Pattern Anal. Mach. Intell. 32 (12) (2010) 2162–2177.
[73] T. Maurer, D. Guigonis, I. Maslov, B. Pesenti, A. Tsaregorodtsev, D. West, G. Medioni, Performance of geometrix activeid tm̂ 3D face recognition engine on the FRGC data, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, p. 154.
[74] I.A. Kakadiaris, G. Passalis, G. Toderici, M.N. Murtuza, Y. Lu, N. Karampatziakis, T. Theoharis, Three-dimensional face recognition in the presence of facial expressions: an annotated deformable model approach., IEEE Transactions on Pattern Analysis & Machine Intelligence 29 (4) (2007) 640–649.
[75] A.S. Mian, M. Bennamoun, R. Owens, Keypoint detection and local feature matching for textured 3D face recognition, Int J Comput Vis 79 (1) (2008) 1–12.
[76] D. Huang, M. Ardabilian, Y. Wang, L. Chen, 3-D face recognition using eLBP-based facial description and local feature hybrid matching, IEEE Trans. Inf. Forensics Secur. 7 (5) (2012) 1551–1565.
[77] J.A. Cook, V. Chandran, C.B. Fookes, 3D Face recognition using log-Gabor templates, in: Proceedings of the British Machine Vision Conference, 2006, p. 83.
[78] C.C. Queirolo, L. Silva, O.R.P. Bellon, M.P. Segundo, 3D face recognition using simulated annealing and the surface interpenetration measure, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2) (2010) 206–219.
[79] A. Savran, L. Akarun, Bosphorus database for 3D face analysis, in: Proceedings of the Biometrics and Identity Management, 2008, pp. 47–56.
[80] O. Ocegueda, G. Passalis, T. Theoharis, S.K. Shah, I.A. Kakadiaris, Ur3D-c: linear dimensionality reduction for efficient 3D face recognition, in: Proceedings of the International Joint Conference on Biometrics, 2011, pp. 1–6.
[81] D. Smeets, J. Keustermans, D. Vandermeulen, P. Suetens, Meshsift: local surface features for 3D face recognition under expression variations and partial data, Comput. Vis. Image Underst. 117 (2) (2013) 158–169.
[82] H. Li, D. Huang, J.M. Morvan, L. Chen, Y. Wang, Expression-robust 3D face recognition via weighted sparse representation of multi-scale and multi-component local normal patterns, Neurocomputing 133 (14) (2014) 179–193.
[83] S. Berretti, N. Werghi, A. Del Bimbo, P. Pala, Special section on 3D object retrieval: matching 3D face scans using interest points and local histogram descriptors, Comput. Graph. 37 (5) (2013) 509–525.
[84] O. Ocegueda, T. Fang, S.K. Shah, I.A. Kakadiaris, 3D Face discriminant analysis using Gauss–Markov posterior marginals, IEEE Trans. Pattern Anal. Mach. Intell. 35 (3) (2013) 728–739.
[85] L. Yin, X. Wei, Y. Sun, J. Wang, M.J. Rosato, A 3D facial expression database for facial behavior research, in: Proceedings of the International Conference on Automatic Face and Gesture Recognition, 2006, pp. 211–216.
[86] V. Vijayan, K.W. Bowyer, P.J. Flynn, D. Huang, Twins 3D face recognition challenge, in: Proceedings of the International Joint Conference on Biometrics, 2011, pp. 1–7.
[87] D. Huang, M. Ardabilian, Y. Wang, L. Chen, A novel geometric facial representation based on multi-scale extended local binary patterns, in: Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition and Workshops, 2011, pp. 1–7.
[88] D. Huang, W. Ben Soltana, M. Ardabilian, Y. Wang, Textured 3D face recognition using biological vision-based facial representation and optimized weighted sum fusion, in: Proceedings of the Computer Vision and Pattern Recognition Workshops, 2011, pp. 1–8.
[89] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, Proceedings of the 22nd ACM international conference on Multimedia, ACM (2014) 675–676.
[90] L. Spreeuwers, Fast and accurate 3D face recognition, Int. J. Comput. Vis. 93 (3) (2011) 389–414.
[91] Aly, Z. Ne, B. Kberk, L. Akarun, Regional registration for expression resistant 3-d face recognition, IEEE Trans. Inf. Forensics Secur. 5 (3) (2010) 425–440.
[92] H. Drira, B. Ben Amor, A. Srivastava, M. Daoudi, R. Slama, 3D face recognition under expressions, occlusions, and pose variations., IEEE Trans. Pattern Anal. Mach. Intell. 35 (9) (2013) 2270–2283.

Ying Cai received the M.S. degree from Kunming University of Science and Technology, Kunming, China, in the area of Artificial intelligence and pattern recognition and the Ph.D. degree from Sichuan University, Chengdu, china, in the area of Electronics and information, in 2007 and 2017, respectively. She is currently an associate professor of the College of Electrical and Information Engineering, Southwest Minzu University, and an engineer in Wisesoft Co. Her research interests include pattern recognition, deep learning and 3D data processing.

Yinjie Lei received his MS degree from Sichuan University, Chengdu, China, in the area of image processing, and the Ph.D. degree in Computer Vision from University of Western Australia, Crawley, Australia. He is currently a Lecturer at Sichuan University, Chengdu, China. His research interests include image and text understanding, 3D face processing and recognition, 3D modeling, machine learning and statistical pattern recognition.



Menglong Yang received the B.S. degree in the College of Chemical Engineering and M.S. degree in the College of Computer Science and Engineering from Sichuan University, in 2005 and 2008, respectively. He is currently an associate professor of the School of Aeronautics and Astronautics, Sichuan University. From 2010–2011, he worked in Center for Biometrics and Security Research (CBSR), Institute of Automation, Chinese Academy of Sciences. Now he is an engineer in Wisesoft Co. His research interests include computer vision, pattern recognition and machine learning.

Zhisheng You received his M.S. degree in Digital Signal Processing from the Sichuan University, Chengdu, China, in1981. From 1981 to 1983, he was a Visiting Scholar at the Michigan State University, East Lansing. He is currently a Professor and Doctoral Supervisor of the School of Computer Science and Engineering, and the Director of the Institute of Image and Graphics, of the Sichuan University. He is the author or coauthor of more than 100 journal papers, and his current research interests include image processing and machine vision, pattern recognition,data fusion,neural networks and transportation engineering. Professor You is an Associate Director of China Society of Image and Graphics, an Editor of the Journal of Computer Applications (in Chinese) and the Journal of Sichuan University (Natural Science Edition)(in Chinese).

Shiguang Shan received Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2004. He has been a full Professor of this institute since 2010 and now the deputy director of CAS Key Lab of Intelligent Information Processing. His research interests cover computer vision, pattern recognition, and machine learning. He has published more than 200 papers, with totally more than 11,000 citations. He served as Area Chairs for many international conferences including ICCV11, ICPR12, ACCV12, FG13, ICPR14, ICASSP14, ACCV16, and FG18. And he is Associate Editors of several journals including IEEE T-IP, Neurocomputing, CVIU, and PRL.