Research on Multimodal 3D Face Recognition Method Based on Deep Learning

Jie Zhang, Chengqing Pan, Jinlin Huang

Abstract—**In order to improve the effectiveness and accuracy of multimodal 3D face recognition, this paper proposes a multimodal 3D face recognition method based on deep learning. Firstly, a 3D facial dataset is selected; secondly, noise is removed from the original multimodal 3D facial image by filling in holes, denoising, and removing sharp points. The denoised results in each pixel have a value between 0 and 1. Then, the multimodal 3D facial image dataset is trained using the multimodal fusion network of convolutional autoencoder in deep learning methods to achieve multimodal 3D facial image fusion; finally, mathematical relationships are used to divide facial regions, and DSC descriptors are used to extract contextual features of the 3D model shape. Based on this, facial similarity is calculated to achieve multimodal 3D face recognition based on deep learning. The results show that the algorithm proposed in this paper outperforms the other two algorithms in all cases, regardless of the threshold value. The error recognition rate is low, and the face recognition time is less than 9.6 seconds. This indicates that the method proposed in this paper can effectively improve multimodal 3D face recognition in efficiency and accuracy, with strong application performance.**

Index Terms—**Deep learning, Multimodal, Normalization processing, Convolutional autoencoder, Facial pose parameters**

I. Introduction

SINCE the 21st century, the continuous development of technology has greatly enhanced the convenience of people’s lives, and increasingly enlarged the range of people’s activities. Therefore, a trans-spatial, fast and accurate identity authentication system is particularly important. Traditional identity authentication is based on physical documents, which have disadvantages such as easy to lose and forge, difficult to carry, and cannot guarantee people’s information security in current society. Therefore, biometric based identity authentication has been favored by a large number of researchers because of its advantages such as convenience, reliability, uniqueness, universality, stability and collectability. Biometric recognition technology is a technology that verifies identity by detecting the unique, measurable, and verifiable biological features of the human body. Among them, facial recognition has become the most widely used method of identity authentication for its advantages of concealment, concurrency, and non invasiveness, which has been applied in various industries and has also become a focus of research [3, 4]. After the 21st century, with the development of hardware devices, a series of 3D facial scanning devices have emerged; subsequently, 3D facial recognition has advanced rapidly. Researchers are gradually inclined towards research on three-dimensional facial recognition. 3D facial recognition has competitive strength in responding to environmental changes. Compared with 2D facial data, 3D facial data has three-dimensional structural information of the face, which conforms to the characteristics of human vision. Therefore, the research on 3D facial recognition technology is of profound significance.

3D facial data has the characteristics of randomness and complexity, posing great difficulties for the research of recognition methods. Yet, through effective feature extraction, the specificity of facial surfaces can be well expressed, algorithm complexity can be effectively reduced, and recognition speed can be largely accelerated. Excellent feature extraction methods can not only simplify the operation in the recognition process, but also greatly enhance the recognition effect. However, the subjective initiative of facial models determines that the 3D data collected from the same object in a non-specific environment has low similarity, while the similarity in the surface structure of the face results in higher similarity among different faces. The information content of 3D facial data is intrinsically complex. However, transforming complex and chaotic original 3D data into feature vectors that are easy to calculate and visually reflect the facial surface through feature extraction can not only improve the real-time performance of the algorithm, but also ensure the recognition rate of the algorithm. At present, the most commonly used features in 3D facial recognition can be divided into local features and overall features.

Dai et al. introduced template matching method for multi-modal 3D face recognition [5], which used weighted ratio function to recognize standardized feature vectors, cluster analysis to classify and extract face image features, and template matching to recognize face features. This method effectively improves face recognition and tracking effect under the condition of uncertain target movement direction, but the accuracy is not high. Huang et al. proposed a multi-modal 3D facial recognition method based on CNN [6], collecting various forms of facial data and performing necessary preprocessing on the data. Convolutional Neural Networks (CNN) is utilized to

Manuscript received June 07, 2023; revised September 28, 2023. This work was supported by Changzhou Sci&Tech Program (CE20235043).

Jie Zhang is a Senior Experimentalist at the School of Computer Engineering, Jiangsu University of Technology, Changzhou 213001, China.(e-mail: zhangjie@jsut.edu.cn).

Chengqing Pan is a PhD student at the School of Economics and Management, China University of Mining and Technology, Xuzhou 221116, China. (e-mail: 570097528@qq.com).

Jinlin Huang is a vice professor at the School of Electrical & Information Engineering, Jiangsu University of Technology, Changzhou 213001, China. (corresponding author to provide phone:+86-519-86953220, fax: +86-519-86953220, e-mail: dxhjl@jsut.edu.cn).

extract features, and different network structures are adopted for different data modalities and types. Then, data from different modalities is integrated, and information such as feature points and curvature features is used to construct a 3D facial model that is further standardized and optimized to improve recognition accuracy. Based on the established 3D facial model and extracted features, this method can effectively improve facial recognition performance, but the recognition speed is poor. Li et al. introduced local feature extraction for multimodal 3D facial recognition [7]. It collects facial data from multiple angles and modalities, and extracts necessary local features from different modalities of facial data based on local feature extraction methods. Then, different local features from multimodal data are integrated, and a 3D facial model through the fused local features is thus reconstructed. With a good 3D model and model matching methods, the similarity between the input facial data and existing facial models is compared for recognition. This method can efficiently advance the effectiveness of facial recognition, but the image recognition recall rate is low.

Therefore, this article introduces deep learning technology to recognize multimodal 3D faces. It selects a 3D facial dataset and utilizes the multimodal fusion network of convolutional autoencoder in deep learning methods; the multimodal 3D facial image dataset is then trained to achieve 3D facial image fusion; following this, multimodal 3D facial pose parameters are calculated, and the multimodal 3D facial recognition based on deep learning is completed, which enormously improves the effectiveness of multimodal 3D facial recognition.

II. RESEARCH ON MULTIMODAL 3D FACE RECOGNITION METHOD BASED ON DEEP LEARNING

A. 3D Face Dataset Selection

The research on deep learning is based on data, and 3D facial recognition is no exception. Table I introduces common 3D facial datasets. The dataset used in this paper is Texas3DFR, which includes 118 people from different continents, races, and ages, with varying facial expressions and light intensity. The dataset includes facial RGB images and facial depth maps [8]. The schematic diagram of some samples in the dataset is shown in Fig. 1.

TABLE I
COMMON 3D FACE DATASETS

Data set	Number of people	Number of samples	Collecting device	Color Map	Data type
FRGC v2	466	4007	Minolta vivid 910	available	Depth map
Bosphorus	105	4666	Mega Capturor	available	Point cloud
BU-3DFE	100	2500	3DMD digitizer	available	Grid
Texas3DFR	118	1150	Kinect	available	Depth map
UMB-DB	143	1473	Minolta vivid 910	available	Depth map
IIIT-D	106	4605	Kinect	available	Depth map
Lock3DFace	509	5711	Kinect V2	available	Depth map
EURECOM	52	936	Kinect	available	Depth map

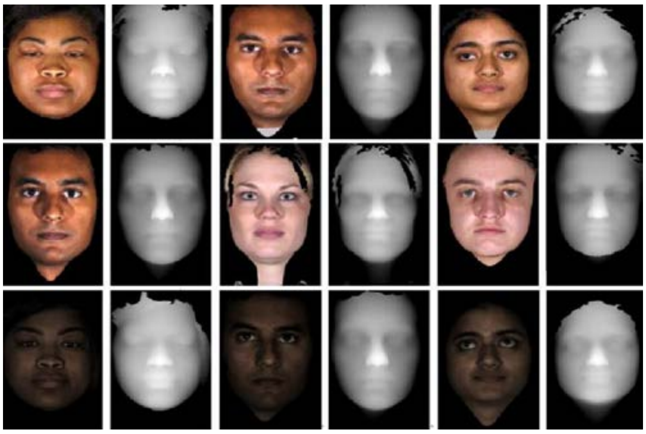


Fig. 1. Partial facial models in Texas3DFR

B. Multimodal 3D facial raw image preprocessing

The facial dataset includes multimodal 3D facial RGB images and depth maps. However, influenced by different data collection equipment, there may be voids, noises, sharp points, and other phenomena in depth maps, as well as noises in multimodal 3D facial RGB images. In order to minimize the impact on algorithm results as much as possible, it is necessary to preprocess RGB images and depth maps to eliminate these phenomena. Depth maps are particularly worth mentioning. Due to their low quality, depth maps can produce significantly negative impact, requiring operations such as filling voids, eliminating noise, and removing sharp points. This section focuses on the preprocessing of depth maps: filling voids, denoising, and removing sharp points.

1) Filling voids

A void is defined as a background pixel surrounded by continuous foreground pixels [9]. Median filtering is used to fill in voids in this paper in that it can not only eliminate noise but also preserve edge information to the maximum extent possible. For example, a 5 * 5 image matrix with a total of 25 pixels is used to sort the pixel values. When the current pixel is empty, search is conducted from the current pixel, and the search range is gradually increased until 5 pixels with depth appear in the neighborhood of the current pixel, and then median filtering is performed in this neighborhood [10]. If a void appears at point (k,s) in image f, and N is a (k,s) neighborhood, the depth value of point (k,s) filled with voids through median filtering is shown in equation (1).

$$f(k,s) = G_{Median}[f(i,j)],(i,j) \in N \qquad (1)$$

2) Denoising

There are many image denoising algorithms in the field of vision. The denoising problem can be described clearly in simple mathematical language: $y = x + e$. y represents the original image with noise, x represents the image without noise, and e represents the probability distribution of noise. The denoising process is to bring the original image y to x as close as possible. In order to obtain the best noise reduction effect while preserving the integrity of information to the maximum extent, this paper uses Bilateral filter algorithm to denoise the depth map [11].

3) Removing sharp points

In facial images, sharp points are defined as parts of the image that do not belong to the face, such as ears, clothing, hair, etc. (see the boxed part in Fig. 2). Sharp points have almost no effect on color maps, but produce significant impact on depth maps. Therefore, it is necessary to remove sharp points in the depth map [12]. From Fig. 2, it can be seen that there is a boundary between the sharp point and the main facial part. Thus, the paper calculates the expected deviation and standard deviation of pixel values within its surrounding 10 neighborhoods for each pixel. If the absolute value of the difference between the pixel value of the pixel and the expected value is no less than 0.8 standard deviation, the pixel is considered as a sharp point and should be removed.

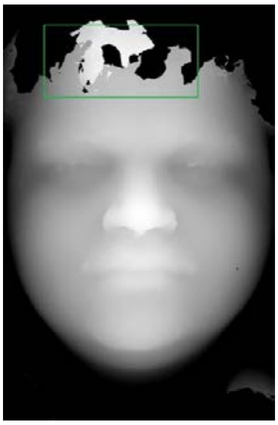
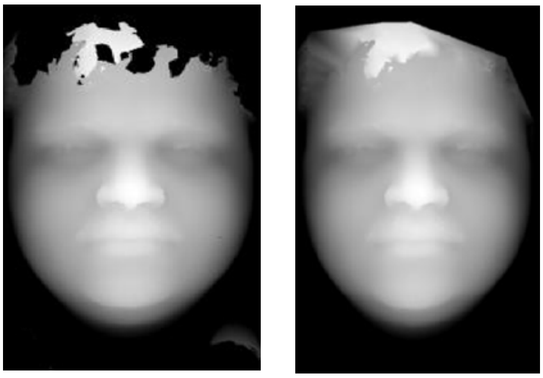


Fig. 2. Sharp Point

At the end of this section, a brief summary of the preprocessing process is provided. Firstly, median filtering is used to fill the voids; pixels from the voids are searched until 5 pixels with depth appear in the neighborhood of the current pixel point, and then median filtering is performed again in the neighborhood. Then, the depth map is denoised. In order to minimize noises to the greatest extent while preserving the integrity of information as much as possible, the paper uses Bilateral filter to denoise the image [13]. Finally, sharp points are removed in the depth map, which are not part of the face, such as hair, ears, clothing, etc., through relevant operations. After the preprocessing of the above three parts, the depth map achieves the optimal effect, as shown in Fig. 3, in which Fig. 3(a) and Fig. 3(b) represents the visual effects before and after the depth map preprocessing representatively, demonstrating a sharp contrast.



(a) Before treatment (b) After treatment

Fig. 3. Comparison of depth maps before and after processing

C. Facial 3D Image Feature Training Based on Deep Learning

The deep learning-based 3D face reconstruction network (R-net) utilizes the loss between the projected 2D face generated by the predicted 3D face model and the input 2D face image, and trains the network through backpropagation to obtain the final reconstructed fine 3D face model. The network predicts 239 3D face related parameters, such as shape coefficient, expression coefficient, and illumination coefficient, etc. Calculation of the similarity of the input image obtains a mixed level loss that is backpropagated in the network to train the network [14].

The 3DMM facial model from BFM is used in the training:

$$S = S(\alpha,\beta)(\delta) = \overline{S} + B_{id}\alpha + B_{\exp}\beta \qquad (2)$$

$$T = T(\delta) = \overline{T} + B_t\delta \qquad (3)$$

Among them, $\overline{S}$ and $\overline{T}$ represent the shape and texture of the average three-dimensional face respectively. Principal component analysis of identity, expression, and texture is represented by B_{id}, $B_{\exp}$, and B_t respectively that are all measured using standard deviation, in which α represents the identity coefficient, β the expression coefficient, and δ the texture coefficient.

Since the same face presents different effects in the camera under different lighting conditions, the generated 3D face shape can be projected from 3D to 2D using a lighting model. The modified convolutional neural network is called R-Net. With the above model, the network predicts the 239 dimensional vector x to generate a 3D facial model.

1) Reconstruction of network loss function

The loss of R-Net is divided into three parts. The first part is the loss of image level, the second part is the loss of perception level, and the third part is the loss of regularization. Next, we will introduce these three losses.

2) Image level loss

The loss at the image level includes two parts: skin color loss and key point loss.

3) Skin color loss

$$L_{photo}(x) = \frac{\sum_{i\in M} A_i \left\| I_i - I'_i(x) \right\|_2}{\sum_{i\in M} A_i} \qquad (4)$$

In the above equation, I represents the color of a pixel in the input 2D image; I' represents the color of the corresponding pixel on the generated projection map; M represents the estimated facial area for calculation; $\|\cdot\|$ represents finding the L2 norm; A represents the skin confidence of the pixel.

The calculation formula for skin confidence level A is as follows:

$$A = \begin{cases} 1 & P_i > 0.5 \\ P_i & other \end{cases} \qquad (5)$$

Among them, P_i is the probability of a certain pixel as skin color value.

4) Key point loss

$$L_{lan}(x)=\frac{1}{N}\sum_{n=1}^{N}\omega_n\left\|q_n-q'_n(x)\right\|^2 \quad (6)$$

In the above equation, N is the number of key points; ω is the weight of the key points, with the key points for the lips and nose set to 20 and the others set to 1; q represents the key point coordinates of the original image; q' represents the coordinates of the projected key points.

5) Perceived level loss

In order to reduce the impact of low lighting on the reconstruction model, the network inputs the input image and the reconstructed projection image through MTCNN for face detection and key point extraction, and then inputs them into the pre-trained recognition network to obtain two feature vectors. The more similar the two are, the smaller the loss. This loss is the perceptual level loss, which is expressed in the following formula:

$$Lper(x)=1-\frac{\langle f(I),f(I'(x))\rangle}{\|f(I)\|\cdot\|f(I'(x))\|} \quad (7)$$

In the above equation, $f(\cdot)$ represents the facial feature vector, and $\langle\cdot\rangle$ represents the inner product operation of the vector.

6) Regularization loss

To prevent overfitting, R-net uses the following common regular constraints:

$$L_{coef}(x)=\omega_\alpha\|\alpha\|^2+\omega_\beta\|\beta\|^2+\omega_\gamma\|\delta\|^2 \quad (8)$$

In the formula, ω_α represents the weight of identity keys, ω_β represents the weight of expression keys, and ω_γ represents the weight of texture keys.

7) Texture loss

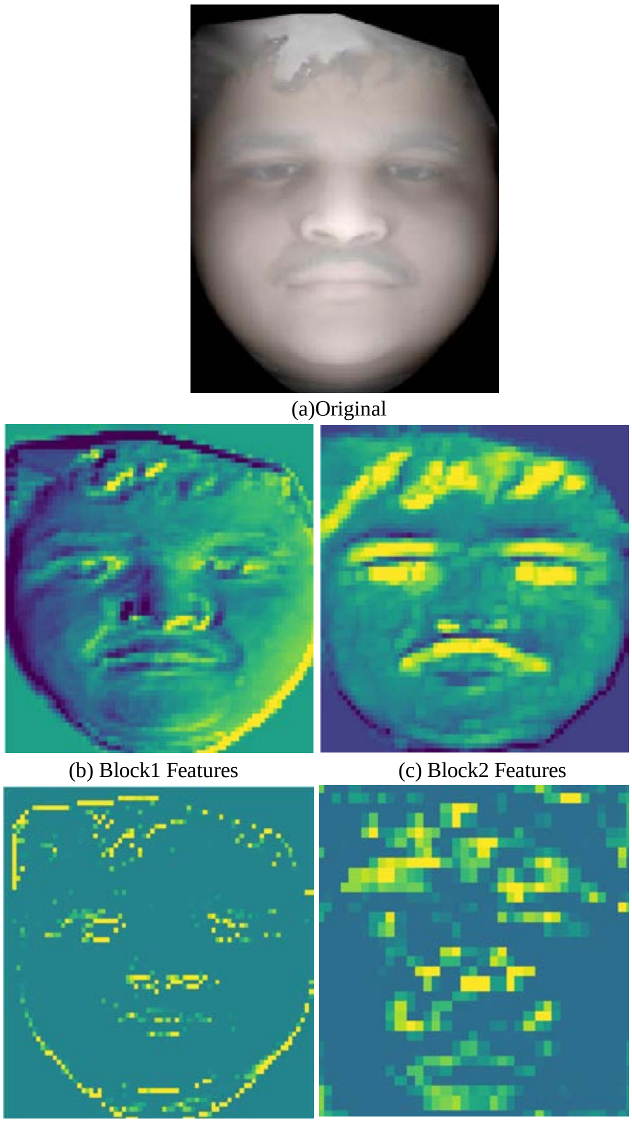
$$L_{tex}(x)=\sum_{R\in\{r,g,b\}}\mathrm{var}(T_R(x)) \quad (9)$$

In equation (9), R represents a predefined skin area, including the cheeks, nose and forehead, with the ear and neck removed to reduce their impact on reconstruction results, $\{r,g,b\}$ represents three channels of skin color, $T_R(x)$ represents skin region texture, and $\mathrm{Var}(\cdot)$ represents loss function.

Finally, all the losses above added equal to loss $L(\hat{x})$. By backpropagation of this loss, the network achieves the purpose of training and can ultimately reconstruct the corresponding 3D face model from a 2D image.

Next, implement feature extraction of facial 3D images. The purpose of this step is to input facial RGB images only for the effect of 3D facial recognition, which still has strong discrimination in situations of facial occlusion and dim lighting.

The underlying features are particularly important for image classification, so it’s key to extract the parameters of the convolutional layer with distinct features from the pre-trained model. Take the fused image as an example. The feature map obtained from the original image through the network model is shown in Fig. 4, where 4 (a) is the original image. In order to verify the ability to extract features and the applicability of migrating to new tasks of the network model as proposed in the paper, the feature maps output from block 1 to block 4 of the original image after passing through the network model were observed. Fig. 4 shows that the features extracted from block1 to block3 are relatively general, while those extracted from block4 are relatively abstract. Specifically, 4 (b) represents the feature map extracted by block1, which can roughly distinguish the facial texture; 4 (c) represents the feature map extracted by block2, highlighting some details such as eyes and the mouth; 4 (d) represents Extract feature maps obtained by block3, demonstrating facial contours and other information; and 4 (e) is the feature map extracted by block4.



(a)Original

(b) Block1 Features

(c) Block2 Features

(d) Block3 Features

(e) Block4 Features

Fig. 4. Original image and feature maps

It can be seen from Fig. 4 that although the number of layers of the network proposed in the paper is small, the ability in feature learning is strong, so transfer learning is feasible. Freeze block1, block2, and block3 without training. These network weights can be directly applied to new tasks for image feature extraction, realizing fast loss reduction, improving training efficiency, and reducing training time.

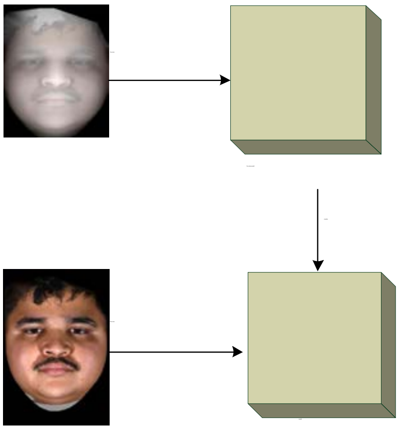


Fig. 5. Transfer learning process

III. Research on multimodal 3D face recognition

A. Facial region division

The distribution pattern of facial organs is relatively uniform. In the front view of the face, the mouth is located below the human eye area, and the midpoint of the mouth is perpendicular to the line connecting the eyes. Five eyes refer to the length of five eyes from the left hairline to the right hairline. According to this mathematical relationship, , the mouth occupies the middle three regions with “one eye” as a unit in the horizontal direction; while in the vertical direction, the mouth area occupies a lower area with “one court” as a unit.

The “Three Courts and Five Eyes” roughly describes the distribution of facial organs of the human body. If rough positioning of the mouth is carried out directly in this way, the mouth area can be generally determined when the face is facing forward. However, when facial expressions change sharply, it is often difficult to accurately divide the area. Therefore, region division is carried out based on the results of nose point detection. Extending downwards from the y-coordinate of the nose point to y-20 as the mouth reference line, the area below this reference line is marked as the mouth area as the MO area; extending upwards to y+100 as the forehead reference line based on the y-coordinate of the nose point, the area above this reference line is marked as the forehead area as the HE area, and the area upwards to y+30 as the eye reference line. The area stuck between the two reference lines is marked as the eye area as the EY area; the nose area NO is between the eye reference line and the mouth reference line. In this way, these three reference lines divide the face into the forehead area, the eye area, nose area, and mouth area.

B. DSC Descriptor

With the widespread application of shape contextual features in 2D shape matching, many researchers attempt to extract shape contextual features from 3D models. Andrea Frome proposed a three-dimensional shape context, where 3DSC is directly extended from 2DSC. As for the latter, it uses the sampling point P as the center and R as the radius, while for 3DSC, it takes P as the center of a spherical region and R as the radius. Similar to 2DSC, 3DSC also has concentric spheres of different sizes, with J representing a radius of $R=\{R_0,...,R_j\}$, K representing an elevation of $\theta=\{\theta_0,...,\theta_k\}$, and L representing an azimuth of $\phi=\{\phi_0,...,\phi_L\}$. The entire sphere area centered around P is divided into $J\times K\times L$ spaces. Then each space corresponds to one element of $J\times K\times$An feature vectors. Formula (10) presents the radius calculation method.

$$R_j=\exp\{\ln(r\min)+\frac{j}{J}\ln(\frac{r_{\max}}{r_{\min}})\} \qquad (10)$$

In equation (6), the minimum radius $r_{\min}$ is the first radius R_0 , and the maximum radius $r_{\max}$ is R_j .

C. Multimodal 3D Face Recognition Based on Similarity Calculation

The 3D shape context is similar to the 2D shape context in matching. Take two faces A and B as examples. The ridge point set is extracted in the sub domain as shown in equations (11) to (16):

$$A_{HE}=\{A_{H1},A_{H2},...,A_{Hn}\} \qquad (11)$$

$$A_{NO}=\{A_{N1},A_{N2},...,A_{Nn}\} \qquad (12)$$

$$A_{MO}=\{A_{M1},A_{M2},...,A_{Mn}\} \qquad (13)$$

$$B_{HE}=\{B_{H1},B_{H2},...,B_{Hn}\} \qquad (14)$$

$$B_{NO}=\{B_{N1},B_{N2},...,B_{Nn}\} \qquad (15)$$

$$B_{MO}=\{B_{M1},B_{M2},...,B_{Mn}\} \qquad (16)$$

Calculate the three-dimensional shape context of each ridge point set separately. Divide it into 8 parts in the longitudinal direction, 6 parts in the latitudinal direction, and 3 parts in the radial direction. Divide the spherical space into 144 spaces, and each point set can be divided into n×144.

Similar to the two-dimensional histogram matching, the similarity calculation also uses the 2X coefficient test as the matching cost. The matching cost of all points in the point set A and B is calculated circularly in the subregion to generate the cost matrix, that is, the shape distance between the two shapes. The Hungarian algorithm is used to match the best points to avoid the mixing of false points and minimize the matching cost.

The decision level weighted fusion is performed on the matching costs of each region to obtain the final matching cost C_s . The calculation formula is shown in equation (17), and the face model with the lowest C_s values in the library set, i.e. the face model with the highest similarity, is selected according to the nearest neighbor rule as the recognition result.

$$C_s=C_Hw_H+C_Ew_E+C_Nw_N+C_Mw_M \qquad (17)$$

In equation (17), C_H,C_E,C_N,C_M represents the matching costs of HE, EY, NO, MO regions respectively, and w_H,w_E,w_N,w_M represents the corresponding weights of each region respectively, thus achieving multimodal 3D face recognition based on deep learning.

IV. Experiment

A. Experimental Environment

The algorithm described in the paper is programmed and tested in a GPU server environment based on ubuntu16.04.

The GPU consists of six GTX1080Ti blocks, with 8G graphics memory, CUDA version 8.6, CPU model i78700, and Tensorflow 2.2 for neural network training. After the construction is completed, a series of operations such as manipulating nodes and updating variables can be performed to calculate the graph, ultimately obtaining the data.

B. Experimental Analysis

To test the proposed method, it is necessary to verify whether it has the function of 3D facial recognition, that is, whether the method can correctly recognize faces in cases of occlusion and illumination. The processed data is used as input for the test set, and under different threshold T conditions, the method proposed in this paper is compared and analyzed with other 3D face recognition related methods in three evaluation indicators: accuracy, error recognition rate, and recognition time.

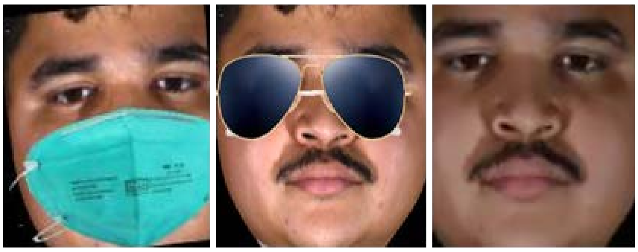


Fig. 6. Image processed with occlusion and light

C. Experimental Result

1) Accuracy of multimodal 3D face recognition in different environments

To compare the effectiveness of the proposed method in multimodal 3D face recognition in different environments, the methods mentioned in reference [5], reference [6], and the proposed method in this paper are used to verify the accuracies of multimodal 3D face recognition. The results are shown in Table II-VII.

TABLE II
ACCURACIES IN DIFFERENT ENVIRONMENTS AT T=0.85

Algorithm	Mask	Sunglasses	Light
Ref. [5] Method	0.9291	0.9603	0.9412
Ref. [6] Method	0.9621	0.9683	0.9707
Our Method	0.9709	0.9897	0.9751

TABLE III
ERROR RECOGNITION RATE IN DIFFERENT ENVIRONMENTS AT T=0.85

Algorithm	Mask	Sunglasses	Light
Ref. [5] Method	0.0531	0.0365	0.0408
Ref. [6] Method	0.0293	0.0158	0.0216
Our Method	0.0137	0.0169	0.0124

TABLE IV
ACCURACIES IN DIFFERENT ENVIRONMENTS AT T=0.90

Algorithm	Mask	Sunglasses	Light
Ref. [5] Method	0.9151	0.9427	0.9389
Ref. [6] Method	0.9512	0.9590	0.9623
Our Method	0.9678	0.9709	0.9683

TABLE V
ERROR RECOGNITION RATE IN DIFFERENT ENVIRONMENTS AT T=0.90

Algorithm	Mask	Sunglasses	Light
Ref. [5] Method	0.0506	0.0337	0.0374
Ref. [6] Method	0.0268	0.0134	0.0193
Ours	0.0101	0.0116	0.0185

TABLE VI
ACCURACIES IN DIFFERENT ENVIRONMENTS AT T=0.95

Algorithm	Mask	Sunglasses	Light
Ref. [5] Method	0.9128	0.9401	0.9357
Ref. [6] Method	0.9501	0.9571	0.9605
Our Method	0.9578	0.9707	0.9628

TABLE VII
ERROR RECOGNITION RATE IN DIFFERENT ENVIRONMENTS AT T=0.95

Algorithm	Mask	Sunglasses	Light
Ref. [5] Method	0.0474	0.0321	0.0352
Ref. [6] Method	0.0236	0.0109	0.0159
Our Method	0.0091	0.0103	0.0155

In the above experiment, the threshold T was set to 0.85, 0.90, and 0.95 respectively, and corresponding accuracy and error rate of the three algorithms were then recorded. Through experimental records, it can be seen that the features extracted by the algorithm in this article are more discriminative and can achieve better recognition performance than other methods, with the best stability performance in three environments. Regardless of the experimental condition or threshold value, the accuracy of the method proposed in this article is higher than that of the other two methods, while the error recognition rate is the lowest among the three algorithms, demonstrating the superiority of the method.

2) Multimodal 3D face recognition takes time

In order to compare the efficiency of multimodal 3D face recognition by the proposed method, the methods of reference [5], reference [6], and the proposed method are used to verify the time consumption of multimodal 3D face recognition. The results are shown in Table VIII.

TABLE VIII
TIME CONSUMPTION OF MULTIMODAL 3D FACE RECOGNITION

Number of images/frames	Multimodal 3D face recognition takes time/s		
	Ref. [5] Method	Ref. [6] Method	Proposed Method
1000	56.8	66.3	2.8
2000	68.9	87.9	3.9
3000	70.9	99.1	5.2
4000	89.3	156.0	6.8
5000	129.8	189.5	8.3
6000	185.0	228.2	9.6

According to Table VIII, when the number of multimodal 3D facial images is 1000, the multimodal 3D facial recognition time of reference [5] method is 56.8 seconds, that of reference [6] method is 66.3 seconds, and that of this method is 2.8 seconds; when the number of multimodal 3D facial images is 3000, the multimodal 3D facial recognition time of reference [5] method is 70.9 seconds, that of reference [6] method is 99.1 seconds, and that of this method is 5.2 seconds; when the number of multimodal 3D facial images is 6000, the multimodal 3D facial recognition time of reference [5] method is 185.0 seconds, that of reference [6] method is 228.2 seconds, and that of this method is 9.6 seconds. It shows clearly that time consumption for multimodal 3D face recognition by this method is consistently lower than that of the other two methods, indicating that this method can effectively improve the efficiency of multimodal 3D face recognition.

V. CONCLUSION

In order to improve the effectiveness of multimodal 3D face recognition, this paper proposes a multimodal 3D face recognition method based on deep learning. Firstly, a 3D facial dataset is selected; secondly, noise is removed from the original multimodal 3D facial image through methods such as filling holes, denoising, and removing sharp points. The multimodal 3D facial image dataset is then trained using a multimodal fusion network of convolutional autoencoders to achieve multimodal 3D facial image fusion; finally, mathematical relationships are used to divide facial regions, and DSC descriptors are used to extract contextual features of the 3D model shape, calculate facial similarity, and achieve multimodal 3D face recognition based on deep learning. Experimental results show that the recognition accuracy of the method proposed in this article is superior to that of the other two methods, with largely low error recognition rate and fast face recognition time of less than 9.6 seconds. The statistics indicate that the proposed method can effectively improve the efficiency of multimodal 3D face recognition, demonstrating the superiority of this algorithm.

REFERENCES

[1] Z. Gao, W. Diao, Y. Huang, R. Xu, H. Lu, J. Zhang, “Identity authentication based on keystroke dynamics for mobile device users,” Pattern Recognition Letters, vol. 148, no. 7, pp. 123-136, 2021.

[2] J. W. Lee, W. K. Lee, S. Y. Sohn, “Patenting trends in biometric technology of the Big Five patent offices,” World Patent Information, vol. 65, no. 6, pp. 102040-102046, 2021.

[3] P. Pankaj, P. K. Bharti, B. Kumar, “A New Design of Occlusion-Invariant Face Recognition Using Optimal Pattern Extraction and CNN with GRU-Based Architecture,” International Journal of Image and Graphics, vol. 26, no. 22, pp. 156-162, 2022.

[4] B. Nassih, A. Amine, M. Ngadi, Y. Azdoud, D. Naji, N. Hmina, “An efficient three-dimensional face recognition system based random forest and geodesic curves,” Computational geometry: Theory and applications, vol. 97, no. 1, pp. 189-193, 2021.

[5] Z. Y. Dai, K. D. Yan, S. W. Yang, Y. M. Liu, L. Li, “Research on face recognition and tracking methods based on template matching,” Journal of Shanghai Electric Power University, vol. 37, no. 1, pp. 83-88+93, 2021.

[6] Y. Z. Huang, Z. H. Yuan, J. R. Chen, W. J. Lin, “Research on face recognition methods based on CNN under the PyTorch framework,” Information and Computer (Theoretical Edition), vol. 34, no. 10, pp. 193-195, 2022.

[7] Q. J. Li, H. C. Yang, “Optimization of face recognition methods based on local feature extraction,” Financial Technology Times, vol. 29, no. 4, pp. 46-52, 2021.

[8] S. Yang, N. Cao, B. Guo, G. Li, “Depth map super-resolution based on edge-guided joint trilateral upsampling,” The Visual Computer, vol. 12, no. 3, pp. 38-45, 2022.

[9] J. Dong, T. Fu, Y. lin, Q. Deng, J. Fan, H. Song, “Hole-filling based on content loss indexed 3D partial convolution network for freehand ultrasound reconstruction,” Computer Methods and Programs in Biomedicine, vol. 211, no. 31, pp. 106421-106429, 2021.

[10] P. Singh, A. K. Bhandari, R. Kumar, “Naturalness balance contrast enhancement using adaptive gamma with cumulative histogram and median filtering,” Journal for Light-and Electronoptic, vol. 18, no. 10, pp. 251-259, 2022.

[11] D. Esan, P. A. Owolawi, C. Tu, “Anomalous Detection in Noisy Image Frames using Cooperative Median Filtering and KNN,” IAENG International Journal of Computer Science, vol. 49, no. 1, pp. 1-10, 2022.

[12] C. Shi, C. Tan, L. Wang, “A Facial Expression Recognition Method Based on a Multibranch Cross-Connection Convolutional Neural Network,” IEEE Access, vol. 18, no. 9, pp. 3-15, 2021.

[13] Y. Zhang, J. Wang, F. Zhang, S. Lv, L. Zhang, M. Jiang, Q. Sui, “Intelligent fault diagnosis of rolling bearing using the ensemble self-taught learning convolutional auto-encoders,” IET Science, Measurement & Technology, vol. 16, no. 22, pp. 168-172, 2022.

[14] Z. Y. Deng, “Research and Application of Webpage Information Recognition Method Based on KNN Algorithm,” IAENG International Journal of Applied Mathematics, vol. 52, no. 3, pp. 725-731, 2022.

[15] B. Bai, H. Ci, H. Lei, Y. Cui, “A local integral-generalized finite difference method with mesh-meshless duality and its application,” Engineering Analysis with Boundary Elements, vol. 139, no. 6, pp. 14-31, 2022.