Fusing Normal Vector and Curvature Features on the Mesh for 3D Facial Expression Recognition

Yu Gu*
Hefei University of Technology, China
yugu.bruce@ieee.org

Yue Fu
Hefei University of Technology, China
fuyue@mail.hfut.edu.cn

ABSTRACT

3D Facial Expression Recognition (FER) has received increasing attention in various applications (e.g., human-computer interaction and driver fatigue detection), as it can enable the machines to understand human intentions or emotions more accurately. However, the current use of 2D projection geometric features has presented limitations when distinguishing facial expressions. To address the above problem, we propose a novel solution by fusing normal vector and curvature features on a mesh. More specifically, we are the first to extract the azimuth and elevation information obtained by the triangle facet normal projection (i.e., mesh-AE descriptor). To the best of our knowledge, we are the first to fuse normal and curvature information on a 3D mesh rather than using 2D projected facial attribute maps for 3D FER. Then, we use the proposed two types of features (mesh-AE descriptor and mesh-H descriptor) to train a two-channel convolutional neural network. Principal Component Analysis (PCA) has greatly increased the efficiency of training. Finally, a linear Support Vector Machine (SVM) is used to identify six types of facial expressions. The experimental results demonstrate the well-designed system can realize accurate and generalized 3D FER.

CCS CONCEPTS

• **Human-centered computing**; • **Human computer interaction (HCI)**; • **HCI design and evaluation methods**;

KEYWORDS

3D expression recognition, Normal projecting, Curvature, Convolutional neural network

ACM Reference Format:
Yu Gu and Yue Fu. 2020. Fusing Normal Vector and Curvature Features on the Mesh for 3D Facial Expression Recognition. In *2020 The 8th International Conference on Information Technology: IoT and Smart City (ICIT 2020), December 25–27, 2020, Xi'an, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3446999.3447022

*Yu Gu, IEEE Senior Member, Hefei University of Technology, China (e-mail: yugu.bruce@ieee.org)

ICIT 2020, December 25–27, 2020, Xi'an, China

ACM ISBN 978-1-4503-8855-9/20/12…$15.00
https://doi.org/10.1145/3446999.3447022

1 INTRODUCTION

Facial Expression (FE), as a non-verbal communication form, plays a crucial role in the social communication. It reflects the mental states of a person, and it can give some clues for the unpredictable events caused by these emotions. Therefore, FER has received increasing attention from researchers, it has a wide range of applications in many areas, such as real time facial expression detection, driver fatigue, pain recognition of patient and so on [1]. Nowadays, a large amount of research on FER mainly focused on identifying these six common types of facial expressions proposed by Ekman et al. [2], namely anger, disgust, fear, happy, sadness, surprise.

Previous facial expression studies [3] have focused on 2D expression recognition. However, lots of studies show that the features which are extracted from 2D images are easily affected by the illumination changes, pose variations, and use of makeup. Moreover, these characteristics can't represent the geometry texture from the 3D face. With the improvement of 3D imaging technologies, FER using 3D face scans has received more and more attention. 3D representations are capable of overcoming the limitations mentioned above. 3D faces are naturally robust to light and changes and facial shape deformations caused by facial muscle movements, and it can provide key cues for expression prediction.

In 3D FER, the most important step is to represent shape patterns of different expressions, many researchers focus on extracting discriminative features to distinguish different expressions. Existing 3D FER approaches generally conclude into two aspects, traditional learning-based methods and deep learning-based methods [3]. Traditional solutions are usually divided into three steps: data preprocessing, face representation and classification data preprocessing, face representation, and automatic classification. For face representation, handcraft feature descriptors are very popular in FER, such as LBP-based [4, 6], SIFT-based [7]. Sun et al.[5] proposed a new FER method which combined Gabor filters and Local Binary Patterns (LBPs), the former one can represent facial shape and appearance over a broader range of scales and orientations while the latter one can capture subtle appearance details, then feature fusion was applied to combine these two vectors, finally the Support Vector Machine(SVM) was adopted to classify facial expressions. Chao et al. [6] proposed an improved facial feature, called the expression-specific local binary pattern (es-LBP), was presented by emphasizing the partial information of human faces on particular fiducial points. The work of Xue et al. [8] dealt with the problem of person-independent FER from a single 3D scan, they used Haar-like features on the depth image and AdaBoost classifier. Hela Bejaoui et al. [9] used 3D Morphable Model(3DMM) to reconstruct the face in 3D space from a 2D image, then extracted a set of features using the mesh-LBP operator on the mesh.

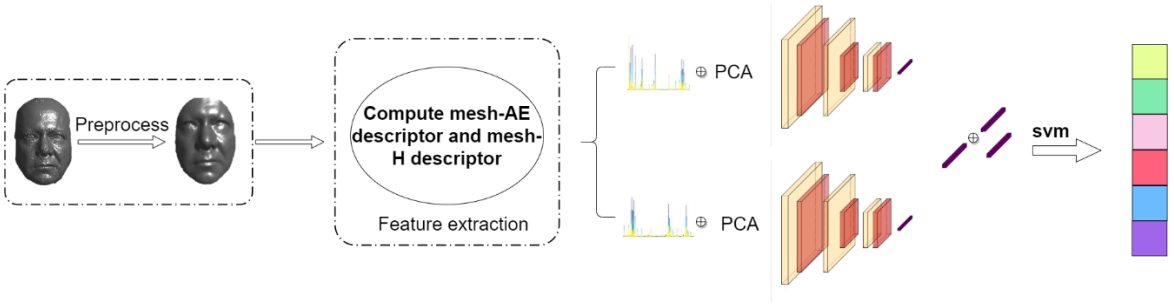

Figure 1: Pipeline of the proposed method for 3D FER

Traditional methods focus on face representation. However, the methods based deep-learning learn focus on the particularization of a deep network. Li et al.[10] presented a novel and efficient Deep Fusion Convolutional Neural Network (DF-CNN), each textured 3D face scan was represented as six types of 2D facial attribute maps, then all of them were fed into DF-CNN for feature learning and fusion learning, expression prediction was performed by a linear SVM classifier or softmax prediction. Uddin et al. [11] performed the extraction of Modified Local Direction Pattern (MLDP) features prior to deep learning and recognition. Yang et al. [12] proposed to recognize facial expressions by extracting information of the expressive component through a de-expression learning procedure, called De-expression Residue Learning (DeRL), expressive information was filtered out by the generative model and recorded in the intermediate layers, finally, they learned the residue of the generative model.

Studies mentioned above mainly used 2-D projected geometric features such as depth maps, curvature maps, normal maps. Depth images were one of the most commonly used imaging modalities, thus many computer vision and pattern recognition solutions can be used to analyze the photometric information in 2D images. Though the idea of extending 2D techniques is attractive, this modality losses the full 3D geometry by reducing it to a 2.5D projection. To address these problems, and preserve full geometric features of 3D face scans on the mesh, we propose a method that extracts normal vector projection (mesh-AE descriptor) and the mean curvature information (mesh-H descriptor) of every triangle mesh facet. Then, we use PCA for feature dimensionality reduction. Finally, we fuse two aspects of information through a two-channel convolutional neural network, we carry out experiments on the Bosphorus database, which is one of the most popular databases to evaluate 3D FER approaches. The results clearly illustrate the effectiveness of the proposed method.

The contributions of this paper are as follows:

• We are the first to use the azimuth information and elevation information obtained by the triangle facet normal projection for 3D FER.

• To the best of our knowledge, our method is the first to fusing normal and curvature information on a 3D mesh rather than use 2D projected facial attribute maps for 3D FER.

• The extracted features are fed into a two-channel convolutional neural network for feature learning and fusion learning after dimensionality reduction and achieve good performance.

2 METHOD

2.1 Overview of proposed method

In this section, we give a brief description of the 3D face expression recognition pipeline. As shown in Figure 1. The whole framework can be summarized in 3 steps:

[Data preprocessing]. To improve the feature extraction, we process the original mesh to obtain a more regular tessellation. The preprocessing techniques applied to 3D faces in our work are: (i) face crop, (ii) smoothing, (iii) holes filling:

[Feature extraction]. We compute the elevation and azimuth value of the normal vector on the mesh and combine the curvature information, we call it the mesh-AE descriptor and the mesh-H descriptor respectively. It represents the geometry features of different expressions. Then a grid of points has been defined and projected on the mesh manifold. We partition the facial surface into a grid of regions. A histogram of each region is extracted and later concatenated to form a global description of the face.

[Facial expression recognition]. To reduce data dimensionality, we use PCA before expression classification. Then whole features are fed into a two-channel convolutional neural network for feature learning and fusion learning, the fused deep features are then classified by linear SVM to obtain final result.

2.2 Data preprocessing

In 3D FER, raw data generally consists of pose variations, presence of non-facial regions, as well as the presence of holes on the face surface. To improve the performance of recognition, preprocessing is an undeniably important procedure. Because one of the advantages of mesh-AE descriptor and mesh-H descriptor is that it does not require normalization, thus pose correction is no need to consider. Our preprocessing steps include:

(i) Face crop: To remove noisy and redundant parts of the 3D face, such as hair, ears, neck, we usually discard the regions that are located beyond a spherical neighborhood of the nose tip. A size of 70 mm to 90 mm is usually used as the radius of the crop.
(ii) Smoothing: To reduce the noise in the data acquisition process, a smooth process is needed. The main measure is to replace each vertex with the average of its neighbors.
(iii) Holes filling: Due to 3D sensors may cause holes in the 3D face, thus we used cubic interpolation to fill holes.

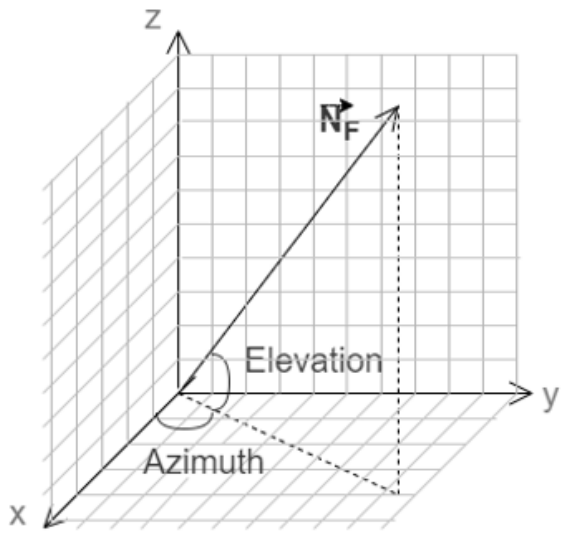

Figure 2: The normal vector N_F is converted to a spherical coordinate system to produce two angeles, one is azimuth, the other is elevation.

2.3 Mesh-AE descriptor and Mesh-H descriptor

3D facial shape deformations caused by facial muscle movements contain important cues to distinguish different expressions, how to catch shape deformations better? We use two kinds of geometric information, one is azimuth and elevation which is obtained by converting the normal vector of mesh facet into spherical coordinates, the other one is the mean curvature. The curvature value of each triangular facet is weighted by the curvature values of the three points that make up the tessellation.

2.3.1 Normal vector projecting. The normal vector is commonly used to measure the degree of curvature of a 3D face. The more the surface is curved, the more the normal vectors at each point are dispersed, and the degree to which these normal vectors are dispersed reflects the degree of curvature of the surface at that point. Let F be a triangle facet on a face mesh and P_1, P_2, P_3 be its vertexes. The unit normal vector N_F of F is computed by:

$$\vec{N}_F = \frac{(P_1 - P_2) \wedge (P_2 - P_3)}{(P_1 - P_2) \wedge (P_2 - P_3)} \tag{1}$$

As a normal vector has three components in x, y, z directions, $\vec{N}_F = (x, y, z)$. The normal vector of the face is then converted to a spherical coordinate system, and this conversion is one-to-one, so there is no loss of information.

$$\begin{aligned} r &= \sqrt{x^2 + y^2 + z^2} \\ \varphi &= \tan^{-1}\left(\frac{y}{x}\right) \\ \theta &= \tan^{-1}\left(\frac{z}{\sqrt{x^2+y^2}}\right) \end{aligned} \tag{2}$$

where φ is azimuth, azimuth is the counterclockwise angle in the x-y plane measured in radians from the positive x-axis, where θ is elevation angle in radians from the x-y plane. As show in Figure 2. Where r is the distance from the origin to point. For the unit normal vector, r = 1.

2.3.2 The mean curvature. In differential geometry, the two principal curvatures at a given point on a surface measure the degree to which this point is bent in different directions. The mean curvature is quantized by two principal curvatures and the mean curvature value at point p is defined as:

$$H(p) = \frac{k_1(p) + k_2(p)}{2} \tag{3}$$

where $k_1(p)$ and $k_2(p)$ denote respectively maximum and minimum principal curvatures at point p. The two principal curvatures at a vertex of a face mesh can be estimated by cubic-order surface fitting [13]. It measures the degree of curvature of a surface in space, and the mean curvature value of each triangle facet is the average of the curvature values at the points that make up the tessellation.

2.3.3 LBP description on the mesh. The calculated angle and curvature information need to be encoded into higher-order feature information like LBP operator, since the tessellations are disordered, we first need to establish the interconnections between them. Naoufel Werghi et al. [14] proposed the concept of mesh-LBP which compute LBP directly on the mesh surface. The method is first to construct sequences of facets ordered in a circular fashion around a central facet, for the triangular mesh representation S = (V, F), where V and F are the sets vertices and facets of the mesh respectively. For any facet f_c, f_{out} facet shares an edge with f_c. Generally, there are three ordered facets fout_1, fout_2, and fout_3 that are adjacent to the central facet f_c, a sequence of f_{gap} facets located between each pair of f_{out} sets, and shares a single vertex with f_c, they look like filling the gap between the f_{out} facets. This procedure produces the first ring around the central facets f_c. Based on the first ordered ring, nine new f_{out} facets can be defined. Repeat the above process to obtain a second ring of ordered facets. Continuing the process iteratively, a multilayer ordered ring of facets around the central facet f_c can be obtained. Figure 3 shows the process of ordered ring construction. The mesh-LBP operator at the facet f_c is defined as follows,

$$\begin{aligned} mesh\,LBP^r_m(f_c) &= \sum_{k=0}^{m-1} s(h(f_k^r) - h(f_c))\ \alpha(k) \\ s(x) &= \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \end{aligned} \tag{4}$$

where r and m are the ring number and the number of facets per ring respectively, s(x) is the step function, h(x) is a scalar function on the mesh, α(k) is a discrete function, and $\alpha(k) = 2^k$ that multiplies single digit by a power of 2, where k represents the facet position. The obtained ordered ring structure forms an adequate support for encoding angle and curvature values, we let h(f) be azimuth value and elevation value respectively, the outputs of mesh-LBP operators of equation 1) are accumulated into a discrete histogram. Then concatenate the two histograms together (mesh-AE descriptor). Also, let h(f) be the mean curvature as the scalar function to compute mesh-LBP (mesh-H descriptor).

2.4 Feature Dimensionality Reduction

The obtained two features result in a large feature vector, use this feature vector can slow down the training period of the system, thus

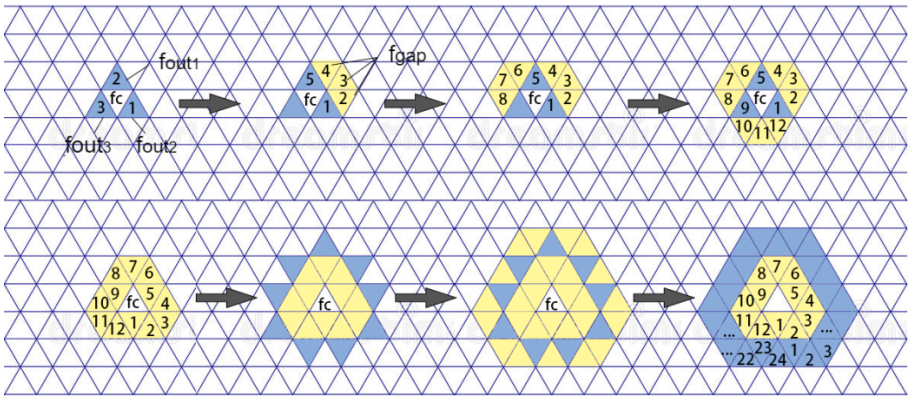

Figure 3: The process of ordered ring construction. Steps 1,2,3,4 show the construction process of the first ordered faceted ring, and the yellow part in step 5 is the first ring obtained. Steps 6,7 are the construction process of the second ordered faceted ring, the blue part of step 8 is the second ring obtained.

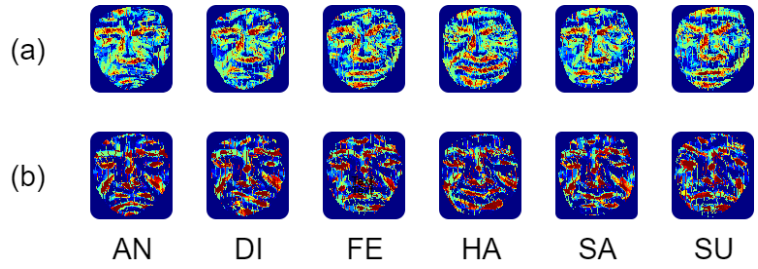

Figure 4: (a) Examples of mesh-AE descriptor which computed using the normal vector. (b) Examples of mesh-H descriptor which computed using the mean curvature.

we use Principal Component Analysis (PCA) to reduce the dimension of feature vector to increase the speed of the learning process. The purpose of the PCA technique is to reduce the dimensionality of the input feature vector, whilst retaining most of the intrinsic information content of the original data. After the PCA process, we obtain smaller yet more efficient vectors for facial expression recognition.

2.5 Facial Expression Classification

After data downscaling, the feature dimension of the mesh-AE descriptor is reduced from 168 * 2250 to 168 * 128, the feature dimension of the mesh-AE descriptor is reduced from 168 * 1125 to 168 * 64. In this paper, a two-channel Convolutional Neural Network (CNN) is designed to classify facial expressions. The network structure is composed of two sub-convolutional neural networks. Each CNN contains 3 convolutional layers and 3 pooling layers, due to the different data dimensions of the two parallel inputs, our two sub-CNN parameter settings are slightly different. The parameter settings are shown in Table1. Two types of features are fed into a two-CNNs for feature learning, after a series of convolution, pooling, and finally dropout operations, the data is flattened. Then concatenate the flatten layer data to generate a single high-dimensional representation of each 3D face scan, and the characteristic dimension of fusion is 3040 dimensions. The feature is further fused by the dense layer, then a 32-dimensional fused deep feature is generated. Finally, a linear SVM classifier is trained by fused 32-dimensional features to recognize six expressions.

3 EXPERIMENTS AND RESULTS

3.1 The Bosphorus database

The Bosphorus database [15] contains 105 subjects, it includes a total of 4,666 pairs of 3D face models and 2D face images. The Bosphorus database is the second most commonly used database in 3D FER. It contains not only neutral and six fundamental expressions but also different types of occlusions and systematic variations of poses. Figure 5 shows the six types of expressions, namely, anger, disgust, fear, happiness, sadness, surprise for the same person in the Bosphorus dataset.

3.1.1 Face Representation. After a series of face slicing, smoothing and hole filling, the optimized 3D mesh is generated. Then compute the mesh-AE descriptor and mesh-H descriptor respectively. If the feature description of the whole face is fed into the classifier for classification, facial spatial information will be lost, resulting in poor classification performance, so the whole 3D mesh faces are divided into a certain number of local regions. This way our descriptors contain spatial information. Firstly, we use the nose tip and perform simple geometric calculations to get a set of fiducial points. These

Table 1: : Network architecture of a two-channel CNN model.

Layer Type	Filters	Filter Size	Output Dimension
Input_1	-	-	168*64*1
Convolution_1	32	5*5	168*60*32
Maxpooling	-	2*2	82*60*32
Convolution_2	16	3*3	80*28*16
Maxpooling	-	2*2	40*14*16
Convolution_3	8	3*3	32*12*8
Maxpooling	-	2*2	19*6*8
Dropout	-	-	19*6*8
Flatten_1	-	-	912
Input_2	-	-	168*128*1
Convolution_4	64	5*5	168*124*64
Maxpooling	-	2*2	82*62*64
Convolution_5	8	3*3	80*60*32
Maxpooling	-	2*2	19*14*8
Convolution_6	8	3*3	38*28*8
Maxpooling	-	-	19*14*8
Dropout	-	-	19*14*8
Flatten_2	-	-	2128
Concatenate	-	-	3040
Dense	-	-	32

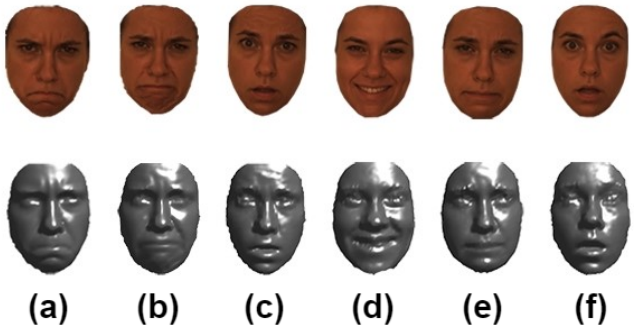

Figure 5: Face models for one person with six expressions in the Bosphorus dataset:(a) Anger (b) Disgust (c) Fear (d) Happiness (e) Sadness (f) Surprise. The first row represents 2D texture images, the second row represents processed 3D mesh faces.

predefined 24 points are then projected onto the face along the nose tip. As shown in Figure 6. We extract the values of neighborhood facets around each point. Calculating the histogram of each block according to a uniform pattern [14]. The dimension of each piece is 1125 * 7. Finally, the 24 blocks are concatenated together to form a histogram with a dimension of 1125 * 168. Azimuth and elevation are obtained by normal vector projecting, so we use equation 1) to get histogram respectively and then concatenated vertically, the obtained dimension is 2250 * 168.

3.1.2 Implementation Details. We evaluate our experiment on the Bosphorus dataset. We select all the subjects containing six types

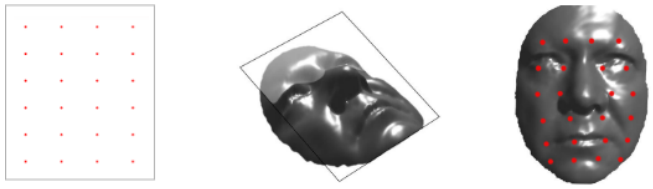

Figure 6: Construction of the face grid on the mesh.

of facial expressions, for a total of 453 3D models. We adopt the 90%/10% partition into training and test sets randomly. Our network structure uses the Keras framework. We use Adam optimizer with a batch size of 30, momentum of 0.9, and dropout of 0.5 for the fully connected layers during training. We use 300 epochs to train the two-channel CNN model. After training learning and fusing learning, we extracted the penultimate layer of data from the convolutional neural network, a fused 32-dimensional feature. Finally, this fused feature is then classified with a linear SVM classifier to evaluate our method.

3.2 Performance and Analysis

We conduct three experiments on the Bosphorus dataset, mesh-H features, mesh-AE features, and a combination of the above two (mesh-AE + mesh-H) are fed into the CNN separately. From the Figure 7, we see that mesh-AE features get better recognition results than mesh-H features, with a recognition rate of 81%. The recognition of the mesh-H descriptor is only 79.8%. From the results of our comparison of a single feature, we can conclude that angular information is more descriptive of facial expression characteristics than curvature information. And the recognition results are optimal

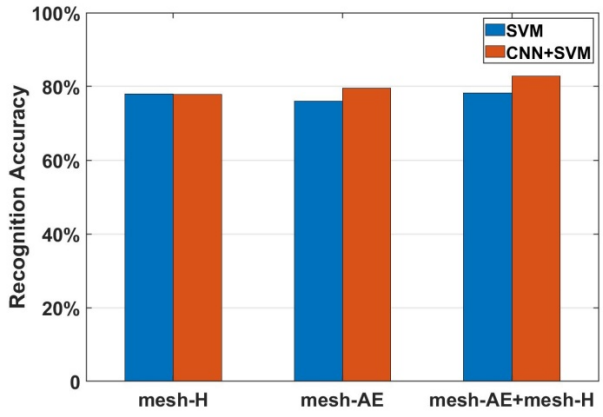

Figure 7: Comparison of classification methods.

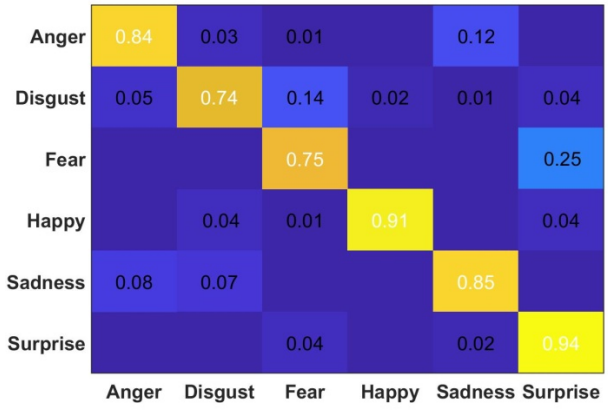

Figure 8: The confusion matrix of six expressions.

when both features simultaneously input to the neural network for fusion learning. The recognition rate for fusion learning is 83.8%. We compare the results of this experiment with an experiment using the only SVM for classification. The parameters of the Linear SVM are the same in three descriptors. A 10-fold cross-validation method is used. The classification effect of the mesh-AE descriptor is slightly lower than that of the mesh-H descriptor. Also, the fusion recognition rate doesn't improve significantly. The confusion matrix of six expressions for the best one experiment results is shown in Figure 8. Surprise and happiness are significantly better identified than other expressions. The reason is the muscle movements of happiness and surprise have a very distinctive character, especially the mouth. Disgust has the worst recognition and 14% of them are misclassified as fear because they have a similar frowning motion. 25% of fear expressions are mistaken for surprise due to the very similar flexing of their mouth muscles movements.

4 CONCLUSION

In this paper, we propose a novel approach for 3D facial expression recognition, solve the problem of information loss caused by mapping 3D faces to 2D images, and obtain good performance. The advantage of this method is that it does not require normalization and at the same time provides a good description of expression features. We construct a two-channel CNN to train the proposed two types of features (mesh-AE descriptor and mesh-H descriptor). The use of dimensionality reduction methods has greatly increased the efficiency of training. Finally, a linear SVM is used to identify six types of expressions and good results are obtained. Our experiments show that normal vectors have a good ability to distinguish between different expressions. It is also shown that feature fusion with deep learning method can yield better results.

ACKNOWLEDGMENTS

This research is funded by the NSFC Grant No. 61432004 and 61772169.

REFERENCES

[1] Kumari, Jyoti, R. Rajesh, and K. M. Pooja. "Facial expression recognition: A survey." Procedia Computer Science 58.1 (2015): 486-491.
[2] Ekman, Paul. "Facial expressions of emotion: New findings, new questions." (1992): 34-38
[3] Alexandre, Gilderlane Ribeiro, José Marques Soares, and George André Pereira Thé. "Systematic review of 3D facial expression recognition methods." Pattern Recognition 100 (2020): 107108
[4] Nigam, Swati, and Ashish Khare. "Multiscale local binary patterns for facial expression-based human emotion recognition." Computational Vision and Robotics. Springer, New Delhi, 2015. 71-77
[5] Sun, Yuechuan, and Jun Yu. "Facial expression recognition by fusing Gabor and local binary pattern features." International Conference on Multimedia Modeling. Springer, Cham, 2017.
[6] Chao, Wei-Lun, Jian-Jiun Ding, and Jun-Zuo Liu. "Facial expression recognition based on improved local binary pattern and class-regularized locality preserving projection." Signal Processing 117 (2015): 1-10.
[7] Soltanpour, Sima, QM Jonathan Wu, and Mohammad Anvaripour. "Multimodal 2D-3D face recognition using structural context and pyramidal shape index." (2015): 2-6.
[8] Xue, Mingliang, *et al.* "Fully automatic 3D facial expression recognition using local depth features." IEEE Winter Conference on Applications of Computer Vision. IEEE, 2014
[9] Bejaoui, Hela, Haythem Ghazouani, and Walid Barhoumi. "Fully automated facial expression recognition using 3D morphable model and mesh-local binary pattern." International Conference on Advanced Concepts for Intelligent Vision Systems. Springer, Cham, 2017
[10] Li, Huibin, *et al.* "Multimodal 2D+ 3D facial expression recognition with deep fusion convolutional neural network." IEEE Transactions on Multimedia 19.12 (2017): 2816-2831
[11] Uddin, Md Zia, *et al.* "A facial expression recognition system using robust face features from depth videos and deep learning." Computers & Electrical Engineering 63 (2017): 114-125
[12] Yang, Huiyuan, Umur Ciftci, and Lijun Yin. "Facial expression recognition by de-expression residue learning." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018
[13] Goldfeather, Jack, and Victoria Interrante. "A novel cubic-order algorithm for approximating principal direction vectors." *ACM Transactions on Graphics (TOG)* 23.1 (2004): 45-63
[14] Werghi, Naoufel, Stefano Berretti, and Alberto Del Bimbo. "The mesh-lbp: a framework for extracting local binary patterns from discrete manifolds." IEEE Transactions on Image Processing 24.1 (2014): 220-235
[15] Alyuz, Nese, Berk Gokberk, and Lale Akarun. "A 3D face recognition system for expression and occlusion invariance." 2008 IEEE Second International Conference on Biometrics: Theory, Applications and Systems. IEEE, 2008.