# A Deep Neural Network-Driven Feature Learning Method for Multi-view Facial Expression Recognition

Tong Zhang, Wenming Zheng, *Member, IEEE*, Zhen Cui, Yuan Zong, Jingwei Yan, and Keyu Yan

***Abstract*—In this paper, a novel deep neural network (DNN)-driven feature learning method is proposed and applied to multi-view facial expression recognition (FER). In this method, scale invariant feature transform (SIFT) features corresponding to a set of landmark points are first extracted from each facial image. Then, a feature matrix consisting of the extracted SIFT feature vectors is used as input data and sent to a well-designed DNN model for learning optimal discriminative features for expression classification. The proposed DNN model employs several layers to characterize the corresponding relationship between the SIFT feature vectors and their corresponding high-level semantic information. By training the DNN model, we are able to learn a set of optimal features that are well suitable for classifying the facial expressions across different facial views. To evaluate the effectiveness of the proposed method, two nonfrontal facial expression databases, namely BU-3DFE and Multi-PIE, are respectively used to testify our method and the experimental results show that our algorithm outperforms the state-of-the-art methods.**

***Index Terms*—Deep neural network (DNN), multi-view facial expression recognition, scale invariant feature transform (SIFT).**

## I. Introduction

FACIAL expression recognition (FER) has become a hot research topic of human-computer interaction (HCI) and drawn a lot of attention due to its great potential in multimedia applications, e.g. digital entertainment, customer service, driver monitoring [1] and so on. HCI would become more friendly and natural if computers are able to recognize affects as human beings, which can benefit from solving FER problems. FER aims to analyze and classify a given facial image into one of the six commonly used emotion types [2], where the six emotion categories are angry, disgust, fear, happy, sad and surprise. Numerous algorithms of FER have been proposed in the literatures during the past several years, including expression recognition from frontal and non-frontal facial images. Comparing to frontal FER, non-frontal FER is more challenging and more applicable in real scenarios. However, only a small part of algorithms among the proposed various methods address this challenging issue [3]–[9]. For both frontal and non-frontal FER problems, a general recognition framework appeared in most of previous works can be divided into two major steps, one is the feature extraction and the other is classifier construction. For the classifier construction, most of the popular classifiers, such as support vector machine (SVM) and Bayes classifier, together with some unsupervised learning techniques are employed in the FER problem. Nevertheless, it is notable that the FER approaches based on the aforementioned framework need to be optimized by tuning parameters within each step and no feedbacks are provided from one step to another.

To extract the facial features, various image features are employed in the previous papers, such as local binary pattern (LBP) [6], [10], local phase quantization [1], histograms of oriented gradients [11], [12] and scale-invariant feature transform (SIFT) [13]. In [6], Moore and Bowden proposed the method of using LBP features to deal with the multi-view FER issue, in which the LBP are extracted from facial images which are divided into several blocks. Then a multi-class SVM is used for classification. Rudovic *et al.* proposed another algorithm to cope with the non-frontal FER problem in [7], in which the authors used a coupled Gaussian process regression model to map the facial points in non-frontal facial image to those in the frontal one. Thus the state-of-the-art algorithms of frontal FER can be applied to dealing with the non-frontal FER problem. In [8] and [9], Zheng *et al.* propose to use SIFT features extracted from the landmark points of certain locations of each facial image, such as the landmark points near mouth and eyes, to describe the facial image for expression recognition, in which high classification accuracy was reported. Among the various facial features, SIFT has demonstrated promising performance due to its robust property to image scaling, rotation, occlusion and illumination difference.

In contrast to the aforementioned recognition frameworks, a new recognition framework based on deep learning network, such as convolution neural network (CNN) or deep belief network (DBN), was presented in recent years. This new framework

Manuscript received December 23, 2015; revised May 11, 2016 and June 30, 2016; accepted July 22, 2016. Date of publication August 3, 2016; date of current version November 15, 2016. This work was supported in part by the National Basic Research Program of China under Grant 2015CB351704, in part by the National Natural Science Foundation of China under Grant 61231002 and Grant 61572009, and in part by the Natural Science Foundation of Jiangsu Province under Grant BK20130020. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chengcui Zhang. *(Corresponding author: Wenming Zheng.)*

T. Zhang and K. Yan are with the Key Laboratory of Child Development and Learning Science (Southeast University), Ministry of Education, Research Center for Learning Science, Southeast University, Nanjing 210096, China, and also with the School of Information Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: tongzhang@seu.edu.cn; 230139081@seu.edu.cn).

W. Zheng, Z. Cui, Y. Zong, and J. Yan are with the Key Laboratory of Child Development and Learning Science (Southeast University), Ministry of Education, Research Center for Learning Science, Southeast University, Nanjing 210096, China (e-mail: wenming_zheng@seu.edu.cn; zhen.cui@seu.edu.cn; xhzongyuan@seu.edu.cn; yanjingwei1989@126.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2016.2598092

had been employed to fulfill various tasks in image processing and achieved great success. Especially, CNN has been applied for image analysis [14]–[16], face recognition [17], FER [18], action recognition [19] and so on. In CNN, convolution layers, pooling layers and some other kinds of common neural layers are usually stacked iteratively to extract some high-level semantic features. In the task of image classification, raw images are directly used as the input data and fed into a multi-layer CNN framework. After training on a large-scale dataset with a back-propagation algorithm [20], the trained CNN is able to capture discriminative features of images. In contrast to CNN, DBN uses multiple layers of feature-detecting neurons to learn the feature representation hierarchically and performs backpropagation for a global optimization [21]–[23]. As a non-linear model, DBN has been extensively applied to various problems, such as handwritten digits recognition and FER in [24], [25]. However, as mentioned in previous works, when original images are employed as inputs, the training dataset should be large enough so that the CNN or DBN model is able to learn effective image representation during the training period. Otherwise, the over-fitting problem may occur and result in lower classification accuracies. To address this problem, additional data is often introduced to train the CNN or DBN model, just as the method proposed in [26].

The recent development of neural systems for recognizing human's facial emotions revealed that the brain perception of human's facial expression could be divided into several major periods happening in different brain areas [27]. The first period is about the low level salient image feature extraction occurring in the occipitotemporal cortex, and the other periods are about the high level emotional semantic feature learning as well as the emotion perception happening in other brain areas such as frontoparietal cortex, orbitofrontal cortex and amygdala [27]. Inspired by the neural systems working mechanism on facial emotion perception, here we develop a special deep neural network (DNN) for the multi-view FER task on a relatively small dataset.

To imitate the first period of the neural cognition system on FER, and meanwhile to alleviate the aforementioned over-fitting problem, we accurately detect those salient facial landmarks covering major expression units of faces and then extract low-level SIFT descriptors from those salient facial landmarks as robust local appearance models to input the sequent network units. Such a process can benefit for characterizing subtle expression changes when facing overwhelming information such as facial pose, personal ID, etc.

According to the facial action coding system developed by Ekman *et al.* [30], different facial regions play different roles in FER. For example, those facial landmarks of the regions around eyes and mouth may contribute more to expression recognition than other landmark points. Consequently, it is reasonable to consider different contributions of different landmark points located at different facial regions. This intuition motivates us to introduce a new kind of layer in our DNN which learns discriminative facial features across different facial landmark points with multi-channel projection matrices, named projection layer. In contrast to those traditional multi-view FER approaches which need to firstly estimate the facial pose and then use the pose-specific expression recognition model to handle this problem, the proposed method can deal with the multi-view FER problem without requiring any facial pose estimation and hence is more suitable for the practical FER problem.

To further extract high-level facial expression features, we use 1D convolutional filters only on feature channels, rather than 2D convolutional filters on the mixed channels of features and spatial positions. One main reason is, the prior landmark detection has provided definite matching points, and thus it is not necessary to confuse the spatial position information by using convolutional and pooling layers, which is often done in conventional CNN networks. Considering the size of training data, we can properly stack a series of layers, including 1D convolution layer, projection layer, fully connected layer, etc., to construct a DNN for FER. As a summary, different from the previous DNN models such as CNN and DBN, the main novelties of our proposed network framework are three folds:

1) we use the 2D SIFT feature matrix consisting of salient low-level features extracted from facial landmark points as the input data of DNN to imitate the first period of the neural cognition mechanism on FER. The use of facial landmark points for feature extraction may alleviate the misalignment problem, which is very different from the pooling strategy used in conventional CNN where the pooling strategy just acts an adverse behavior.
2) we use projection layer to learn discriminative facial features across different facial landmark points. In this process, the facial features associated with all landmarks are integrated to produce an ensemble feature set that better discriminate the different expressions.
3) we employ 1D convolutional layer which is directly performed on feature channels to extract high-level features instead of 2D filters in conventional CNNs. Compared to conventional CNNs, an advantage of using both projection layer and convolutional layer is that it can significantly reduce model complexity and hence the proposed framework are more adaptive to small sample tasks.

The rest of this paper is organized as follows. In Section II, we propose the DNN-driven facial feature learning method for expression recognition. In Section III, we use two datasets to evaluate the propose method for multi-view FER problem. Section IV concludes our paper.

## II. DNN-Driven Facial Feature Learning Method

In this section, we will address the DNN-driven facial feature learning method in details. For this purpose, we firstly address the method of extracting SIFT features from each facial image.

### A. Facial Feature Extraction

To extract the SIFT features, we firstly annotate a fixed number of key points from each facial image, where the key points are located around the nose, mouth and eyes. Then, we extract a set of SIFT feature vectors associated with the key points to represent the facial image, in which each SIFT feature vector is a 128-dimensional vector. Basically, the SIFT feature extraction

procedures include the steps of scale-space extrema detection, key point localization, orientation assignment and key point description [28]. We extract SIFT features from certain positions as used in [8], in which the main orientation of each key point is set to be a certain value (usually 0) for all key points. In the key point description step, normalization is handled to make it invariant to changes in illumination.

### B. DNN Framework

The SIFT facial feature vectors extracted from landmark points of each facial image had been proven to be effective for the FER problem [8]. For multi-view FER problem, however, it is notable that the distributions of facial feature vectors may vary with the changes of facial poses. Consequently, it would be advantageous to learn a set of ensemble discriminative features from the raw SIFT facial features associated with various facial poses in order to improve the multi-view FER performance. To this end and motivated by the recent development of DNN, in this section we will propose a novel method of using DNN to learn the ensemble features for our multi-view FER problem. Fig. 1 illustrates our DNN-driven feature learning framework, which consists of six layers, i.e., two projection layers, one 1D convolution layer, two fully connected layers and a soft-max layer. Different from conventional CNN or DBN that directly uses the raw facial images to learn facial features, the proposed network adopts 2D SIFT descriptor to firstly extract salient low-level features and then learn higher-level semantic features by using 1D convolutional operation.

Assume that we have located $M$ landmark points from each facial image, and for each point we extract a $N$-dimensional SIFT feature vector. Then, we can put the $M$ SIFT feature vectors together to form a $M \times N$ feature matrix to represent each facial image, in which each row of the feature matrix corresponds to a SIFT feature vector associated with one key point of this face image. The feature matrices corresponding to the multiple facial images are finally fed into the DNN as input data to train this network. In the projection layer of this deep network, we use multiple left multiplication projection matrices to integrate the facial features associated with all landmarks to produce more discriminative features that could better discriminate the different expressions. This process could also be regarded as a spatial filtering of key points via proper linear combinations of those rows of the input data matrix. In addition, the right multiplication projection matrices in the second projection layer are further used to extract more discriminative features among the high-level features for the FER problem.

Let $\mathcal{H}_t = \{\mathbf{H}_{t,j}^{(l)} | j = 1, \ldots, C_l\}$ $(t = 1, \ldots, N_l)$ denote the $t$th multi-channel projection matrix set consisting of $C_l$ channels of projection matrices, where $\mathbf{H}_{t,j}^{(l)}$ denote the $j$th channel matrix of $\mathcal{H}_t$, $N_l$ denotes the number of multi-channel projection matrix sets, and $C_l$ denotes the number of channels in $\mathcal{H}_t$. Then, the left multiplication projection layer can be expressed as the following form:

$$\mathbf{O}_t = \sum_{j=1}^{C_l} \mathbf{H}_{t,j}^{(l)} \mathbf{I}_j, (t = 1, 2, \ldots, N_l) \tag{1}$$

where $\mathbf{O}_t$ denotes the matrix in $t$th channel of the output and $\mathbf{I}_j$ is the $j$th channel of the input matrices. If the projection layer applies right multiplication, i.e.,

$$\mathbf{O}_t = \sum_{j=1}^{C_r} \mathbf{I}_j \mathbf{H}_{t,j}^{(r)}, (t = 1, 2, \ldots, N_r) \tag{2}$$

then it projects each row of the input matrix from one feature space to another, which also results in the dimension reduction outcome simultaneously.

The convolutional layer uses a group of filters to process small local parts of the input and is always followed by the max-pooling layer. Different from most DNN approaches such as CNN, the filters here are 1D sequences and applied to the input matrix only along the row direction according to the structure of feature matrix. These filters may evoke strong responses when dealing with some parts of the input feature matrix while the values of the other parts are suppressed, and thus capture those crucial local structures. As the convolution layer follows the first projection layer, it is able to extract high-level features from the ensemble features of the projection layer. In addition, it is interesting to see that the left multiplication projecting transform and the right one actually factorize the classic 2D convolutional matrix in CNN networks, thus the number of connections in the proposed network can be largely reduced and the over-fitting problem can be alleviated [29]. After the filtering operation, the results are passed through a max-pooling layer, which yields multiple new feature maps. Assuming $V_{i,j}^{r,l}$ represents the value of the unit at position $(r, l)$ of the $j$th feature map in the $i$th layer, then the convolution operation can be given by

$$V_{(i,j)}^{(r,l)} = \sum_{k \in \mathcal{S}} \sum_{p=0}^{d_i - 1} W_{i,j,k}^{p} V_{i-1,k}^{r,l+p} \tag{3}$$

where $W_{i,j,k}^{p}$ denotes the value of the unit at position $p$ in the $j$th channel of the convolution kernels in the $i$th layer, $d_i$ represents the length of the kernels in the $i$th layer, and $\mathcal{S}$ denotes the set that contains the indexes of the feature maps in the $(i-1)$th layer which are connected with the current feature map.

The max-pooling layer is realized by down-sampling the filtering results of the convolution layer just along the row direction by taking the maximum filter activation within a specified window. As a result, it transforms the filtering result to a lower resolution version, which would make the network be robust to the minor variations of positions.

The output of the second projection layer ("P4" in Fig. 1), denoted by $\mathbf{Q} = [Q_{c,i,j}]_{C \times I \times J}$, passes through a nonlinear activation function before it is set to the first full connection layer ("F5" in Fig. 1), where $C$, $I$, $J$ denote the numbers of the channels, rows and columns, respectively. This process can be described as follows:

$$F_{c,i,j} = \tanh(b_c + Q_{c,i,j}) \tag{4}$$

where the result of the activation function is denoted as $\mathbf{F} = [F_{c,i,j}]_{C \times I \times J}$ and $F_{c,i,j}$ denotes the element in the $i$th row and $j$th column of the $c$th channel, $\tanh(\cdot)$ is the hyperbolic tangent

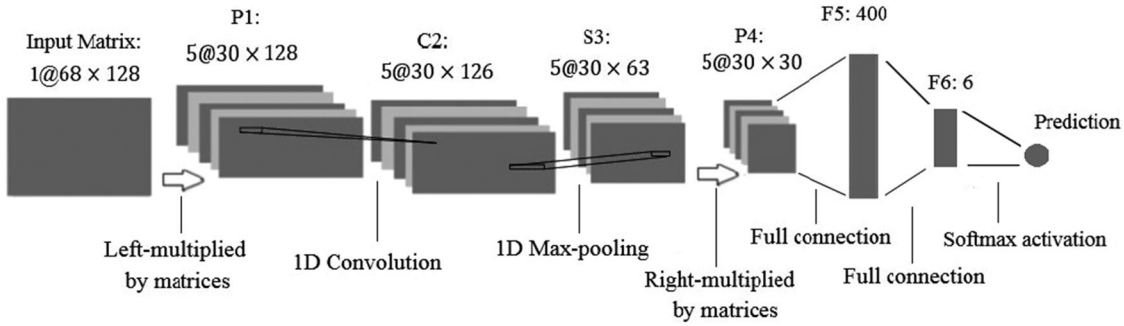



Fig. 1. Structure of the DNN-driven feature learning framework for FER. This framework consists of two projection layers, two full connected layers, one convolution layer, one max-pooling layer, and one softmax layer. Detailed descriptions are given in the text.

function and $b_c$ is the bias for the matrix of the $c$th channel. $\mathbf{F}$ has the same size as $\mathbf{Q}$.

The fully connected layer and the softmax layer are used as the same as those of CNN and DBN. The fully connected layers combine inputs from all positions, and finally the classification is done by the softmax layer.

### C. DNN Training

To train the proposed DNN, we divide the training data set into some subsets and the subsets are sequently sent to the network for training. Then a loss function defined as

$$L(\mathbf{m}, \mathbf{H}^{(l)}, \mathbf{H}^{(r)}) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{Y}\tau(y_n, i) \times m_n \times \log P_{n,i} + \lambda_1 \sum_{t=1}^{N_l}\sum_{j=1}^{C_l}||\mathbf{H}_{t,j}^{(l)}||_1 + \lambda_2 \sum_{q=1}^{N_r}\sum_{k=1}^{C_r}||\mathbf{H}_{q,k}^{(r)}||_1 \tag{5}$$

is calculated to evaluate the difference between the predicted results and the labels, where

$$\tau(y_n, i) = \begin{cases} 1, & \text{if } y_n = i \\ 0, & \text{otherwise} \end{cases}$$

where $N$ denotes the number of the training samples, $Y$ is the number of expression types, $y_n$ is the label of $n$th training sample, $\lambda_1$, $\lambda_2$ are the weights of the two $l_1$ norm regularization terms, $\mathbf{m} = [m_n]_{N\times 1}$ is a vector consisting of different weights for the training samples where $m_n$ means the value of the weight for the $n$th training sample, and $P_{n,i}$ represents the value of the prediction that the $n$th training sample is predicted to be the $i$th class.

The parameters of each layer are updated according to the value of the loss function through backpropagation algorithm. In the loss function, the first term calculates the mean negative logarithm value of the prediction probability of the training samples. The second and third terms ensure the sparse structure of the matrices in the two projection layers. These sparse matrices in projection layer are able to learn discriminative facial features across different facial landmark points, where the elements of the projection matrix consist of different weights indicating the importance of the corresponding landmark points to learn the discriminative features. The parameters $\lambda_1$, $\lambda_2$ are constant which means that the values of them may not change for all the training samples once they are set. However, the values of elements in $\mathbf{m}$ are not always the same for all the training samples. The elements of $\mathbf{m}$ determine the weights of the gradients during the process that the parameters of the network are tuned. They are set according to the orientation angles during extracting SIFT descriptors, which is illustrated in detail in Section III. If some values of elements in $\mathbf{m}$ are set to be larger for a certain part of the training samples, the adjustment of the parameters in the network will be more affected by these samples.

## III. Experiments

In this section, we conduct experiments on both Multi-PIE [31] and BU-3DFE [32] facial expression databases to evaluate the proposed multi-view FER method. The Multi-PIE facial expression database contains six facial expressions which are disgust, neutral, scream, smile, squint and surprise. These facial expressions are performed by 337 subjects under 15 view points and 19 illumination conditions. The 337 subjects consist of 235 males and 102 females from different areas in the world. The BU-3DFE database contains 100 people of different ethnicities, including 56 females and 44 males. Six universal facial expressions (anger, disgust, fear, happiness, sadness and surprise) are elicited by various manners, and each of them includes 4 levels of intensities which yields 2400 facial expression models. These models are described by both 3D geometrical shapes and color textures with 83 feature points (FPs) identified on each model. Some image samples of Multi-PIE and BU-3DFE datasets are shown in Fig. 2.

For Multi-PIE database, we use the same data set as [9], i.e., images of 100 subjects are selected from all the subjects, which contain all the six facial expressions under seven views ($0^\circ$, $15^\circ$, $30^\circ$, $45^\circ$, $60^\circ$, $75^\circ$ and $90^\circ$) of one certain illumination condition. Each facial image is annotated with 68 key points. Thus, 4200 facial images are chosen from Multi-PIE database. For BU-3DFE dataset, we project the models of 100 people to facial images of $0^\circ$, $30^\circ$, $45^\circ$, $60^\circ$ and $90^\circ$ yaw angles. At the same time, the 83 FPs are also projected onto the corresponding 2D faces. Thus we get 12 000 facial images under five angles with 83 annotated points on each image. These images are transformed into gray color space and 83 SIFT descriptors are extracted on the annotated points of each face.

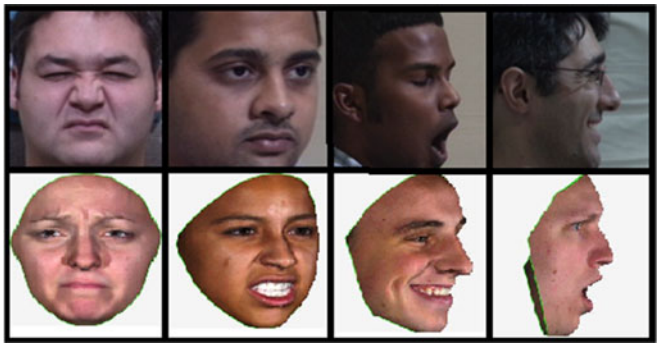


Fig. 2. Facial images in Multi-PIE and BU-3DFE datasets of different expressions. Images in the first row are from Multi-PIE and the second are from BU-3DFE. Images from the left column to the right column are under views $0°$, $30°$, $60°$, and $90°$, respectively.

### A. Experiments on Multi-PIE

For our algorithm, the parameters of DNN are set as what follows. In the input layer, the feature matrices are in size $68 \times 128$ of one channel as we extract SIFT descriptors in gray scale and the dimension of each SIFT descriptor is 128. The projection matrices in left projection layer are in size $5 \times 1 \times 30 \times 68$ which means that there are five 1-channel matrices in the left projection layer and each matrix of one channel contains 30 rows and 68 columns. The filters in convolution layer have the size $5 \times 5 \times 1 \times 3$ where the number of channel is set corresponding to the number of the 1-channel matrices in the left projection layer. The projection matrices in the right projection layer are in size $5 \times 5 \times 63 \times 30$. The first fully connected layer combines the input matrix of this layer into a long vector, and the transformation matrix in this layer has the size $4500 \times 400$ which transforms the dimension of the feature from 4500 to 400. The size of the transformation matrix in the second fully connected layer is $400 \times 6$, in which 6 is set according to the number of the emotion types.

The experiment is carried out with a cross-validation strategy. As we have 4200 images of 100 people, we randomly divide them into a training set of 80 subjects and a testing set of 20 subjects and there's no overlap between the subjects of the two sets. Thus we get 3360 facial images in the training set and 840 facial images in the testing set. As the number of the images in training set is not so large in this experiment, we extract SIFT descriptors not only on the original images, but also on the transformed images as it is illustrated in Section II-C. For the given training set, we firstly extract SIFT descriptors on the original images, then we perform mirror transformation to these images and also extract SIFT descriptors on them. After this, the original facial images are rotated ten degrees clockwise and counterclockwise respectively so that we can get another two sets of SIFT descriptors. Thus totally 13 440 training samples are created. This process is shown in detail in Fig. 3. For the testing data, the SIFT descriptors are extracted only on the original testing images. As it is illustrated in Section II-C, we want the adjustment of the parameters to be mainly affected by the SIFT descriptors extracted on the original training images, so we set some of the weights in $\mathbf{m}$ of (5) to be 2.5 for this part of descriptors and 1 for others. The values of $\lambda_1$, $\lambda_2$ are both set to be 0.0000135 according to the experimental results.

Fig. 3. Image rotation and mirror transformation in SIFT extraction process. The first column contains original images. The images in the second column are under mirror transformation. The images in the third and fourth columns are rotated ten degrees clockwise and counterclockwise respectively.

Table I compares the average recognition accuracy of our neural network with the results achieved by various algorithm on Multi-PIE, including deep learning methods and algorithms based on hand-crafted features. For deep learning methods, we compare the results with the networks including DBN, CNN and the joint fine-tuning DNN (JFDNN) proposed in [18]. To set the parameters of these algorithms as optimal as possible, we conduct experiments for CNN, DBN and JFDNN by traversing in a large range of the numbers of nodes in each layer. Concretely, for convolutional layers in CNN and JFDNN, the range of convolution kernels number in each convolutional layer is [10, 35]. We build these networks with theano [33], [34], which is a public and popular tool that stochastic gradient descent can be done automatically. And in all our experiments no additional samples from other datasets are used. The DBN used in our experiment contains two hidden layers which contain 1200 and 200 nodes respectively. The input of the DBN is an 8704 dimensional vector assembled by the SIFT descriptors of the 68 key points in each facial image. The accuracy of DBN is $76.1\%$. The CNN in our experiment contains three convolution layers followed by three max-pooling layers respectively and two fully connected layers. The filters are in size $23 \times 1 \times 5 \times 5$ in the first convolution layer, $15 \times 23 \times 3 \times 3$ in the second convolution layer and $23 \times 15 \times 3 \times 3$ in the third convolution layer respectively. As CNN employs image as input directly, images which contain only facial areas are cropped from the original images of Multi-PIE dataset and normalized to the size $64 \times 64$. After hundreds rounds' training, the highest accuracy it achieves is $77.8\%$. For JFDNN, as video sequences are considered as its inputs, it cannot directly be applied to FER based on static images. For this reason, we transform their framework into a 2D based one. We build a 2D JFDNN consisting of a 2D CNN of three convolution layers and a deep geometry network of two hidden

TABLE I
COMPARISON OF THE AVERAGE RECOGNITION ACCURACIES AMONG THE STATE-OF-THE-ART AND OUR NETWORK ON MULTI-PIE DATABASE

| Methods | Subjects | Expression number | Pose number | Features | Overall(%) |
|---|---|---|---|---|---|
| DBN (2 hidden layers) | 100 (80 train, 20 test) | 6 | 7 | SIFT (3360 training samples) | 76.1 |
| CNN (3 convolution layers) | 100 (80 train, 20 test) | 6 | 7 | augmented images (13440 training images) | 77.8 |
| 2D JFDNN [18] | 100 (80 train, 20 test) | 6 | 7 | augmented images and geometry features of landmarks (13440 training images) | 82.9 |
| Moore and Bowden [6] | 100 (80 train, 20 test) | 6 | 7 | $LBP^{\mathrm{ms}}$ | 73.3 |
| Moore and Bowden [6] | 100 (80 train, 20 test) | 6 | 7 | LGBP | 80.4 |
| Zheng GSRRR [9] | 100 (80 train, 20 test) | 6 | 7 | SIFT | 79.3 |
| Zheng GSRRR [9] | 100 (80 train, 20 test) | 6 | 7 | $LBP^{u2}$ | 81.7 |
| Our Method | 100 (80 train, 20 test) | 6 | 7 | SIFT (3360 training samples) | 82.0 |
| Our Method | 100 (80 train, 20 test) | 6 | 7 | SIFT (13440 training samples) | **85.2** |

TABLE II
RESULTS OF AVERAGE RECOGNITION ACCURACIES OF EACH EXPRESSION VERSUS THE DIFFERENT FACIAL VIEWS ON THE MULTI-PIE DATABASE

| Features | Results(%) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $0^\circ$ | $15^\circ$ | $30^\circ$ | $45^\circ$ | $60^\circ$ | $75^\circ$ | $90^\circ$ | Average |
| 2D JFDNN [18] | 85.1 | 83.3 | 85.8 | 84.2 | 80.8 | 80.8 | **80.0** | 82.9 |
| Our Method (13 440 samples) | **88.2** | **88.0** | **87.3** | **85.3** | **83.8** | **83.3** | 78.8 | **85.2** |

layers, where the filters in three convolution layers are with size $15 \times 1 \times 5 \times 5$, $17 \times 15 \times 3 \times 3$ and $23 \times 17 \times 3 \times 3$ and the numbers of hidden nodes are both set to be 50. The input of the 2D CNN are static images and the input of the deep geometry network are concatenated coordinates of landmark points of each face. After jointly fine-tuning the two networks, we get accuracy of $82.9\%$. Besides the deep learning methods, our result is compared with the previous works based on hand-drafted descriptors. Multiple kinds of descriptors are employed in [6] to fulfill the FER task, and the best result of them is $80.4\%$ by using the LGBP descriptor. The accuracies of the method in [9] are $79.3\%$ by using SIFT descriptors and $81.7\%$ with $LBP^{u2}$. For our method, if the features are not augmented by performing mirror transformation and rotation to the images, the result is $82.0\%$ which is competitive to the result of [9]. However, if we use the augmented 13440 samples, the accuracy of our network can achieve the accuracy of $85.2\%$, which is the highest among these methods. The comparison between the results of our method by using the original 3360 training samples and the augmented 13440 samples demonstrates the effectiveness of adding features extracted from transformed images.

Table II shows the accuracies of the comparison between 2D JFDNN transformed from [18] and our neural network corresponding to the seven yaw head poses ($0^\circ$, $15^\circ$, $30^\circ$, $45^\circ$, $60^\circ$, $75^\circ$ and $90^\circ$). From this table we can see that the result of each view in our method is higher than 2D JFDNN except the result under $90^\circ$, which is $1.2\%$ lower. High accuracies are achieved in $0^\circ$, $15^\circ$ and $30^\circ$ which are $88.2\%$, $88.0\%$ and $87.3\%$ separately, and the view corresponds to the lowest accuracy is $90^\circ$. The highest accuracy of 2D JFDNN appears under the view $30^\circ$ while the highest accuracy of our algorithm appears under the view $0^\circ$.

Fig. 4 shows the confusion matrices of the different expression recognition results under different views and the result of overall expression recognition. As it is shown, the two expressions of scream and surprise are much easier to be recognized than others, which is most likely due to their relatively large muscle deformations. And for these two kinds of expressions, all the accuracies of the seven yaw head poses are higher than $90\%$. Followed are the recognition results of disgust, neutral and smile, which are more than $80\%$. The lowest accuracy is $72\%$ of the expression squint. It should be noticed that the main error classification comes from the confusion between disgust and squint. There are 17 percent of disgust samples misclassified to be squint and 19 percent of squint samples misclassified to be disgust. The high confusion may be caused by the fact that the expressions of disgust and squint have similar muscle deformations around eyes, which is pointed out by Moore and Bowden in [6].

### *B. Experiments on BU3D-FE*

In this experiment, the parameters of the neural network are set almost the same as those on Multi-PIE. The main difference locates in the input layer and the first projection layer because 83 key points are annotated in each facial image in BU3D-FE instead of 68 points. So in the input layer, the feature matrix contains 83 rows and 128 columns of one channel. The projection matrices in the left projection layer are in size $5 \times 1 \times 30 \times 83$. The convolution layer contains five $1 \times 3$ filters of 5 channels. In the right projection layer, the projection matrices are in size $5 \times 5 \times 63 \times 30$. The sizes of the transformation matrices in the first and the second fully connected layer are $4500 \times 400$ and $400 \times 6$ respectively.

The experiment is also carried out with a cross-validation strategy, which is similar to the aforementioned section. This time we have 12 000 images of 100 people, and we still randomly divide them into a training set of 80 subjects and a testing set of 20 subjects without overlapping between the subjects of the two sets. Thus we get 9600 facial images in the training set and 2400 facial images in the testing set. As the training set is much larger than Multi-PIE this time, we only employ the original training samples without any augmentation (e.g., rotation, mirror transformation, etc.). This also means that in this experiment the weights of the gradients in (5) are set to be 1 for all the training samples. The values of $\lambda_1$, $\lambda_2$ are still both set to be 0.0000135 according to the experimental results.

The average recognition accuracy of our neural network is compared with the results achieved by DBN, CNN, 2D JFDNN,

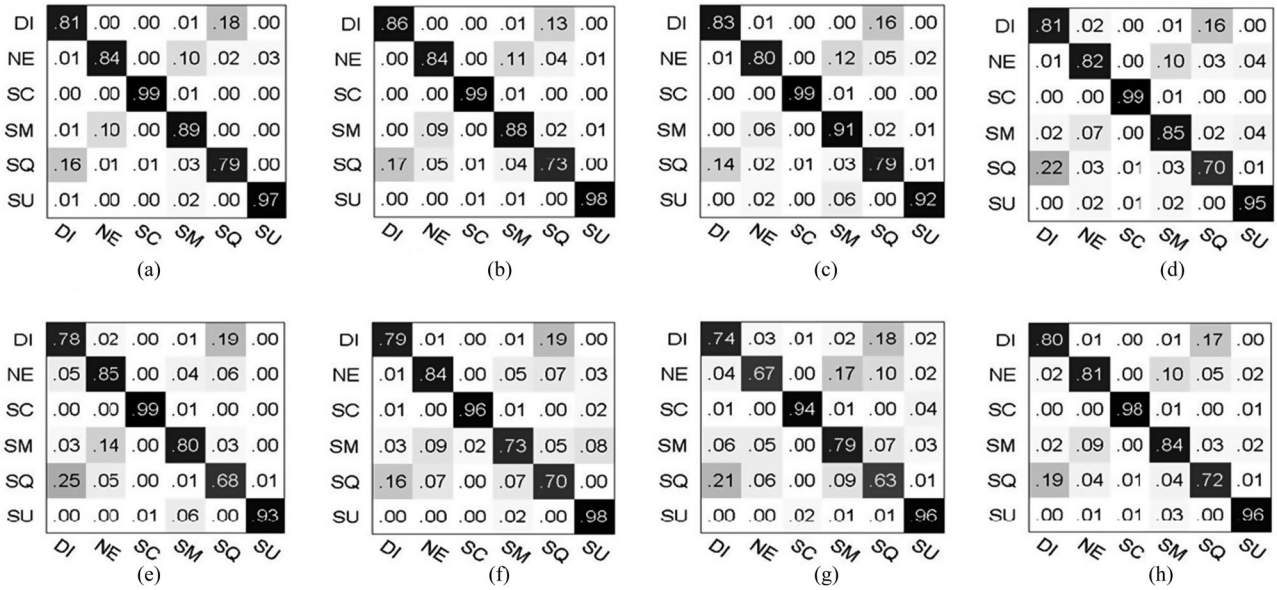



Fig. 4. Experimental results of confusion matrices on the Multi-PIE database. (a)-(g) The confusion matrices corresponding to seven facial views. (h) The overall recognition rates with respect to different facial expressions.

TABLE III
COMPARISON OF THE AVERAGE RECOGNITION ACCURACIES AMONG THE STATE-OF-THE-ART AND OUR NETWORK ON BU3D-FE DATABASE

| Methods | Expression number | Pose number | Features | Overall(%) |
|---|---|---|---|---|
| DBN (2 hidden layers) | 6 (4 levels of intensities) | 5 | SIFT | 73.5 |
| CNN (3 convolution layers) | 6 (4 levels of intensities) | 5 | original images | 68.9 |
| 2D JFDNN [18] | 6 (4 levels of intensities) | 5 | original images and geometry features of landmarks | 72.5 |
| Moore and Bowden [6] | 6 (4 levels of intensities) | 5 | $LGBP/LBP^{\mathrm{ms}}$ | 71.1 |
| Rudovic, Patras, and Pantic [7] | 7 (2 levels of intensities) | 35 train, 247 test | 39 landmarks | 76.5 |
| Zheng *et al.* [8] | 6 (4 levels of intensities) | 5 | sparse SIFT | 78.4 |
| Zheng GSRRR [9] | 6 (4 levels of intensities) | 5 | sparse SIFT | 78.9 |
| Our Method | 6 (4 levels of intensities) | 5 | SIFT | **80.1** |

[6]–[8] and [9] in Table III. As the experimental settings of [7] is different from others, we mainly focus on the methods in DBN, CNN, 2D JFDNN, [6], [8] and [9]. The DBN used in this experiment contains two hidden layers which contain 4000 and 400 nodes respectively. The input of the DBN is a 10 624 dimensional vector assembled by the 128 dimensional SIFT descriptors of the 83 key points in each facial image. The accuracy of DBN is $73.5\%$. The CNN in our experiment contains three convolution layers followed by three max-pooling layers respectively and two fully connected layers. The filters are in size $24 \times 1 \times 5 \times 5$ in the first convolution layer, $28 \times 24 \times 3 \times 3$ in the second convolution layer and $31 \times 28 \times 3 \times 3$ in the third convolution layer respectively. The input images are also normalized to the size $64 \times 64$ which is the same as the experiment on Multi-PIE. As it is shown in Table III, the accuracy of the CNN on BU3D-FE dataset is $68.9\%$. 2D JFDNN achieves $72.5\%$ which contains three convolution layers and two hidden layers, where the filters in the three convolution layers are in size $16 \times 1 \times 5 \times 5$, $24 \times 16 \times 3 \times 3$ and $28 \times 24 \times 3 \times 3$ and the numbers of the hidden nodes are set to be 50. The accuracy of the method in [6] using $LGBP/LBP^{\mathrm{ms}}$ is $71.1\%$ and higher accuracies are achieved with SIFT using algorithms proposed in [8] and [9], which are $78.4\%$ and $78.9\%$ separately. Our method achieves $80.1\%$ which is competitive to the methods above.

TABLE IV
RESULTS OF AVERAGE RECOGNITION ACCURACIES OF EACH EXPRESSION VERSUS THE DIFFERENT FACIAL VIEWS ON THE BU3D-FE DATABASE

| Features | Results(%) | | | | | |
|---|---|---|---|---|---|---|
| | $0^\circ$ | $30^\circ$ | $45^\circ$ | $60^\circ$ | $90^\circ$ | Average |
| Zheng GSRRR (sparse SIFT) [9] | 78.9 | 80.1 | 80.1 | 78.4 | 77.0 | 78.9 |
| Our Method | **79.7** | **80.7** | **81.0** | **80.5** | **79.5** | **80.1** |

Table IV shows the accuracy of each view of our method. As the GSRRR method proposed in [9] achieves high accuracy, our result is also compared with the results of GSRRR under all five views. From this table we can see that the accuracies under all the views of our method are higher than those of GSRRR. The highest accuracies of the two methods are both achieved under the view $45^\circ$, which are $81.0\%$ of our method and $80.1\%$ of GSRRR. And the lowest accuracies appear both under the view $90^\circ$.

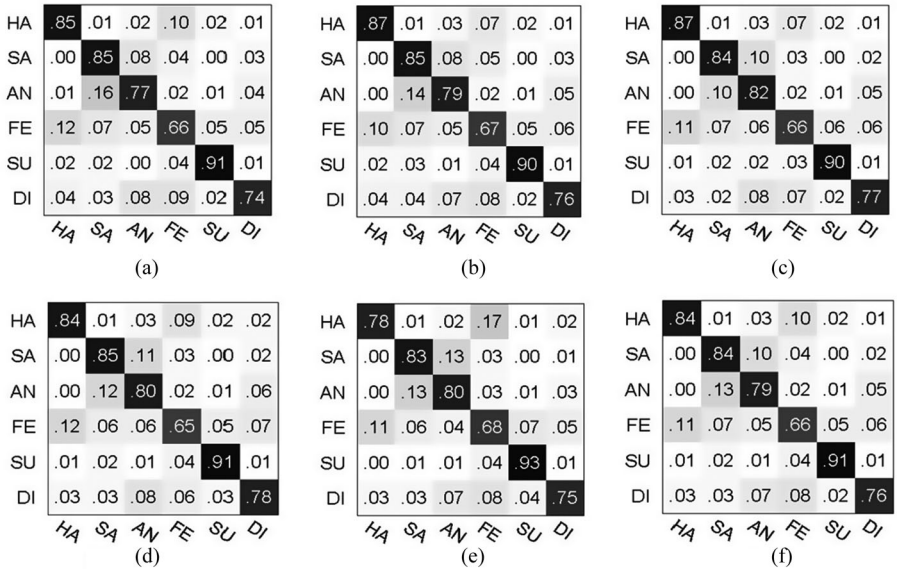



Fig. 5. Experimental results of confusion matrices on the BU3D-FE database. (a)-(e) The confusion matrices corresponding to five facial views. (f) The overall recognition rates with respect to different facial expressions.

Fig. 5 shows the confusion matrices under different views on BU3D-FE dataset. As it is shown, the highest accuracy is $91\%$ of surprise which is consistent with the result in Fig. 4 because of large muscle deformations. Followed are the recognition results of happy and sad, which are more than $80\%$. The lowest accuracy is $66\%$ of fear. Relatively high confusions appear between two pairs of expressions, which are angry versus sad and fear versus happy.

## IV. Conclusion

In this paper, a DNN-driven feature learning method is proposed to deal with the multi-view FER problem by borrowing the visual mechanism of FER. The SIFT descriptors are firstly extracted from those accurate detected landmarks to imitate the salient low-level visual feature detection of the first period in the neural cognition system. In sequent, two novel layers including the projection layer and convolutional layer are designed based on the structure of the low-level input feature to adaptively learn spatial discriminative information as well as extract more robust high-level features, which is very different from those conventional CNNs and DBNs. As a factorization on 2D convolutional matrix, the two layers can largely reduce the space complexity of parameters and further alleviate the overfitting phenomenon especially on those small dataset. The extensive experiments on two different facial expression databases demonstrate that our proposed framework is more competitive over state-of-the-arts under the same experimental environments.

## References

[1] A. Tawari and M. M. Trivedi, "Face expression recognition by cross modal data association," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1543–1552, Nov. 2013.

[2] P. Ekman and W. V. Friesen, *Pictures of Facial Affect*. Palo Alto, CA, USA: Consulting Psychologists Press, 1976.

[3] M. Pantic and I. Patras, "Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 36, no. 2, pp. 433–449, Apr. 2006.

[4] S. Moore and R. Bowden, "The effects of pose on facial expression recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2009, pp. 1–11.

[5] S. Kumano, K. Otsuka, J. Yamato, E. Maeda, and Y. Sato, "Pose-invariant facial expression recognition using variable-intensity templates," *Int. J. Comput. Vis.*, vol. 83, no. 2, pp. 178–1941, 2009.

[6] S. Moore and R. Bowden, "Local binary patterns for multi-view facial expression recognition," *Comput. Vis. Image Understand*. vol. 115, no. 4, pp. 541–558, 2011.

[7] O. Rudovic, I. Patras, and M. Pantic, "Coupled Gaussian process regression for pose-invariant facial expression recognition," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 350–363.

[8] W. Zheng, H. Tang, Z. Lin, and T. Huang, "A novel approach to expression recognition from non-frontal face images," in *Proc. Int. Conf. Comput. Vis.*, Sep.-Oct. 2009, pp. 1901–1908.

[9] W. Zheng, "Multi-view facial expression recognition based on group sparse reduced-rank regression," *IEEE Trans. Affective Comput.*, vol. 5, no. 1, pp. 71–85, Jan.–Mar. 2014.

[10] J. Whitehill, G. Littlewort, I. Fasel, and J. Movellan, "Toward practical smile detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 2106–2111, Nov. 2009.

[11] M. Dahmane and J. Meunier, "Emotion recognition using dynamic grid-based HoG features," in *Proc. IEEE Int. Conf. Automat. Face Gesture Recog. Workshops*, Mar. 2011, pp. 884–888.

[12] M. Dahmane and J. Meunier, "Prototype-based modeling for facial expression analysis," *IEEE Trans. Multimedia*, vol. 16, no. 6, pp. 1543–1552, Apr. 2014.

[13] W. Zheng, H. Tang, Z. Lin, and T. Huang, "Emotion recognition from arbitrary view facial images," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 490–503.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1106–1114.

[15] Z. Wu, Y. Huang, and L. Wang, "Learning representative deep features for image set analysis," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1960–1968, Nov. 2015.

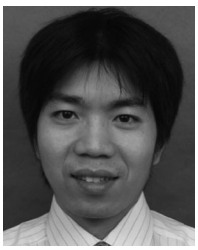
[16] L. Tran, D. Kong, H. Jin, and J. Liu, "Privacy-CNH: A framework to detect photo privacy with convolutional neural network using hierarchical features," in *Proc. Nat. Conf. Artif. Intell.*, 2016, pp. 1317–1323.

[17] C. Ding and D. Tao, "Robust face recognition via multimodal deep face representation," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2049–2058, Nov. 2015.

[18] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2982–2991.
[19] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.
[20] Y. LeCun *et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989.
[21] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
[22] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.
[23] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proc. Annu. Int. Conf. Mach. Learn.*, 2009, pp. 609–616.
[24] J. M. Susskind, A. K. Anderson, G. E. Hinton, and J. R. Movellan, *Generating Facial Expressions With Deep Belief Nets*. Rijeka, Croatia: INTECH Open Access Pub., 2008.
[25] M. Ranzato, J. Susskind, V. Mnih, and G. E. Hinton, "On deep generative models with applications to recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 2857–2864.
[26] H. Chen *et al.*, "Standard plane localization in fetal ultrasound via domain transferred deep neural networks," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 5, pp. 1627–1636, Sep. 2015.
[27] R. Adolphs, "Neural systems for recognizing emotion," *Current Opinion Neurobiol.*, vol. 12, no. 2, pp. 169–177, 2002.
[28] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[29] H. Mobahi, R. Collobert, and J. Weston, "Deep learning from temporal coherence in video," in *Proc. Annu. Int. Conf. Mach. Learn.*, 2009, pp. 737–744.
[30] P. Ekman, W. V. Friesen, and J. C. Hager, "Facial action coding system: The manual on CD-ROM. Instructor's Guide," Salt Lake City, UT, USA: Netw. Inform. Res. Co., 2002.
[31] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010.
[32] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D facial expression database for facial behavior research," in *Proc. IEEE Int. Conf. Automat. Face Gesture Recog. Workshops*, 2006, pp. 211–216.
[33] F. Bastien, P. Lamblin, R. Rascanu *et al.*, "Theano: new features and speed improvements," *CoRR*, 2012. [Online]. Available: http://arxiv.org/abs/1211.5590.
[34] J. Bergstra *et al.*, "Theano: A CPU and GPU math compiler in Python," in *Proc. 9th Python Sci. Conf.*, 2010, pp. 1–7.

**Tong Zhang** received the B.S. degree in information science and technology from Southeast University, Nanjing, China, in 2011, the M.S. degree from the Research Center for Learning Science, Southeast University, in 2014, and is currently working toward the Ph.D. degree in information and communication engineering at Southeast University.

His interests include pattern recognition, machine learning, and computer vision.



**Wenming Zheng** (M'08) received the B.S. degree in computer science from Fuzhou University, Fuzhou, China, in 1997, the M.S. degree in computer science from Huaqiao University, Quanzhou, China, in 2001, and the Ph.D. degree in signal processing from Southeast University, Nanjing, China, in 2004.

Since 2004, he has been with the Research Center for Learning Science, Southeast University. He is currently a Professor and the Director of the Key Laboratory of Child Development and Learning Science of Ministry of Education, Research Center for Learning Science, Southeast University. His research interests include affective computing, neural computation, pattern recognition, machine learning, and computer vision.

**Zhen Cui** received the B.S. degree from Shandong Normal University, Jinan, China, in 2004, the M.S. degree from Sun Yat-sen University, Guangzhou, China, in 2006, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2014.

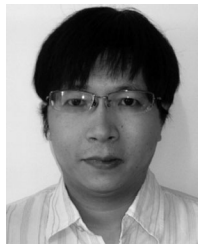
He was a Research Fellow with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, from 2014 to 2015. He is currently an Associate Professor with Southeast University, Nanjing, China. His research interests include sparse coding, manifold learning, deep learning, face detection, alignment and recognition, and image super resolution.

**Yuan Zong** received the B.S. and M.S. degrees in electronics engineering from Nanjing Normal University, Nanjing, China, in 2011 and 2014, respectively, and is currently working toward the Ph.D. degree at the Research Center for Learning Science, Southeast University, Nanjing, China.

His research interests include affective computing, pattern recognition, and speech signal processing.

**Jingwei Yan** received the B.S. degree in instrument science and engineering from Southeast University, Nanjing, China, in 2011, and is currently working toward the Ph.D. degree at the Key Laboratory of Child Development and Learning Science of the Ministry of Education, Research Center for Learning Science, Southeast University.

His research interests include deep learning and affective computing.



**Keyu Yan** is currently working toward the Ph.D. degree in information and communication engineering at Southeast University, Nanjing, China.

His current research interests include computer vision, pattern recognition, and machine learning.