IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 14, 2017, accepted December 9, 2017, date of publication December 15, 2017,
date of current version February 28, 2018.

Digital Object Identifier 10.1109/ACCESS.2017.2784096

Facial Expression Recognition Using Weighted Mixture Deep Neural Network Based on Double-Channel Facial Images

BIAO YANG[1], JINMENG CAO[1], RONGRONG NI[2], YUYU ZHANG[1]
[1]Department of Information Science and Engineering, Changzhou University, Changzhou 213164, China
[2]College of Mechanical and Electrical, Changzhou Textile Garment Institute, Changzhou 213164, China

Corresponding author: Biao Yang (yb6864171@cczu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61501060 and Grant 61703381, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20150271, and in part by the Key Laboratory for New Technology Application of Road Conveyance of Jiangsu Province under Grant BM20082061708.

ABSTRACT Facial expression recognition (FER) is a significant task for the machines to understand the emotional changes in human beings. However, accurate hand-crafted features that are highly related to changes in expression are difficult to extract because of the influences of individual difference and variations in emotional intensity. Therefore, features that can accurately describe the changes in facial expressions are urgently required. Method: A weighted mixture deep neural network (WMDNN) is proposed to automatically extract the features that are effective for FER tasks. Several pre-processing approaches, such as face detection, rotation rectification, and data augmentation, are implemented to restrict the regions for FER. Two channels of facial images, including facial grayscale images and their corresponding local binary pattern (LBP) facial images, are processed by WMDNN. Expression-related features of facial grayscale images are extracted by fine-tuning a partial VGG16 network, the parameters of which are initialized using VGG16 model trained on ImageNet database. Features of LBP facial images are extracted by a shallow convolutional neural network (CNN) built based on DeepID. The outputs of both channels are fused in a weighted manner. The result of final recognition is calculated using softmax classification. Results: Experimental results indicate that the proposed algorithm can recognize six basic facial expressions (happiness, sadness, anger, disgust, fear, and surprise) with high accuracy. The average recognition accuracies for benchmarking data sets "CK+," "JAFFE," and "Oulu-CASIA" are 0.970, 0.922, and 0.923, respectively. Conclusions: The proposed FER method outperforms the state-of-the-art FER methods based on the hand-crafted features or deep networks using one channel. Compared with the deep networks that use multiple channels, our proposed network can achieve comparable performance with easier procedures. Fine-tuning is effective to FER tasks with a well pre-trained model if sufficient samples cannot be collected.

INDEX TERMS Facial expression recognition, double channel facial images, deep neural network, weighted mixture, softmax classification.

I. INTRODUCTION

Facial expression recognition (FER) aims to predict basic facial expressions (e.g., happiness, sadness, anger, surprise, disgust, and fear) from human facial images, as shown in Fig. 1. This method helps machines understand the intention or emotion of humans by analyzing their facial images. FER elicited considerable attention because of its potential application in human-abnormal behavior detection, computer interfaces, autonomous driving, health management, and other similar tasks.

Pre-processing, such as face detection and rotation rectification, are needed for a given facial image. The former is consistently realized with cascade classifiers, such as the Adaboost [1] and the Viola-Jones frameworks [2]. Rotation rectification can be implemented with the aid of landmarks such as the eyes [3]. Facial expression features are extracted from facial regions after pre-processing. Geometric and appearance features are commonly used. For the former, the locations of many facial landmark points are extracted and subsequently combined into a feature

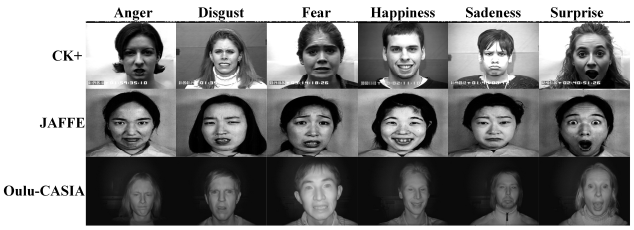


FIGURE 1. Six basic facial expressions in benchmarking datasets.

vector that encodes facial geometric information (e.g., angle, distance, and position) [4]. Appearance features are used to model the appearance variations of a particular face via a holistic spatial analysis [5]. The features of motion information are used for expression recognition in a sequence of facial images [6]. Finally, an effective classifier is used to recognize different facial expressions by learning parameters on the extracted features.

Despite recent rapid developments, FER remains challenging because of several factors, such as illumination changes, partial occlusions of facial regions, and head deflection. These interferences may influence the performance of face detection and reduce the accuracy of FER. Hand-crafted features are no longer suitable for FER tasks with severe disturbances. Fortunately, deep learning may provide a satisfactory solution to these issues.

Convolutional neural network (CNN) has recently achieved rapid advances in pattern recognition, especially in face detection [7] and handwritten mathematical expression recognition [8]. CNN can automatically understand and learn the abstract signatures of the target with a deep network [9]. With deeper layers and elaborate designs, CNN or any other deep network can perfectly realize FER under wild conditions.

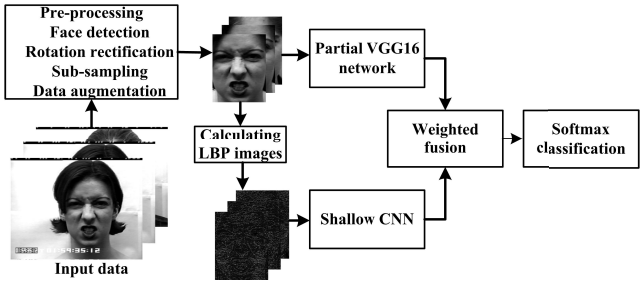


FIGURE 2. Pipeline of the proposed FER approach based on WMDNN.

A weighted mixture deep neural network (WMDNN) is proposed for FER tasks under wild conditions. The pipeline of the proposed FER approach is illustrated in Fig. 2. As shown in the figure, several pre-processing approaches, such as face detection, rotation rectification, data augmentation, are necessary for input facial images. The corresponding local binary pattern (LBP) facial image for a pre-processed facial image is calculated to focus on facial local information. LBP is a commonly used texture feature. It has some advantages such as easy calculation and small data size. LBP is widely utilized in face recognition. Thus, we argue that it may be suitable for FER. Different deep neural networks are used for different channels of facial images. A partial VGG16 network with initial parameters obtained from the VGG16 model pre-trained on ImageNet is built for facial grayscale images to automatically extract expression-related features. For LBP facial images, a shallow CNN, which refers to the construction of DeepID [10], is built for automatic feature extraction. Outputs from binary channel facial images are then fused in a weighted way and the fusion results are finally processed by softmax classification to predict current facial expression from six basic expressions(happiness, sadness, anger, surprise, disgust, and fear).

Our work focuses on the issues of feature extraction and expression recognition for facial images. The novelties of the study are threefold. First, binary channels of facial images, including grayscale images and their corresponding LBP images, are used for FER because of their complementary properties. Second, a fine-tuning strategy is utilized to make full use of a well-learned pre-trained model (VGG16 model on ImageNet). Finally, outputs of both channels are weighted fused to predict a robust result. Three benchmarking datasets and several practical facial images are used to evaluate the effectiveness of our work.

The rest of this study is organized as follows. Section 2 reviews related work on FER approaches. Section 3 provides details of the proposed weighted mixture deep networks. Section 4 shows the experimental results and analysis. The conclusions are presented in Section 5.

II. RELATED WORK

A detailed review of FER is beyond the scope of this paper, which can be referred to [1]. Here we only review some work on feature extraction, which is a significant issue of FER.

A. FER APPROACHES BASED ON HAND-CRAFTED FEATURES

A FER task focuses on extracting facial expression features from facial RGB (or grayscale) images and recognizing different facial expressions with a trained classifier. Traditional FER tasks depend on hand-crafted features. The three main types of features are appearance, geometric, and motion features. Commonly used appearance features include pixel intensity [11], Gabor texture [12], LBP [13], and histogram of oriented gradients (HOG) [14]. These features capture global and detailed information from facial images and can thus reflect an individual's expression. However, these features are extracted from the entire facial region, and local regions that are highly related to expression changes, such as the eyes, nose, and mouth, are ignored. Therefore, geometric features, which are represented by the geometric relationships of facial landmark points detected from local regions that are highly related to expression changes, are used for FER tasks [15]. Moreover, the combination of different features is a promising trend. For example,

a two-stage multi-task framework to study FER was proposed by Zhong *et al.* [16]. Key facial regions were effectively detected through multi-task learning, and features were extracted from these regions through a sparse coding strategy. Afterward, SVM was used as a classifier to recognize different expressions. Zhang *et al.* [17] extracted texture and landmark features from facial images. These two features are complementary and can catch subtle expression changes.

These FER tasks mainly involve still facial images. With the development of FER for video analysis, an increasing number of researchers have focused on motion features, such as optical flow [18], motion history images (MHI) [6], and volume LBP [19]. Dynamic models of FER tasks have also been widely studied. Walecki *et al.* used a conditional random field (CRF) framework to recognize different facial expressions and motion units on faces [20]. They argued that temporal variations in facial expressions could improve the accuracy of FER. Jain *et al.* [4] combined linear chain CRF, hidden CRF, and the additional variables of the hidden layer to build a dynamic model. This model can describe expression changes through a similarity analysis.

B. FER APPROACHES BASED ON DEEP LEARNING

Existing FER approaches based on hand-crafted features demonstrate a limited recognition performance. Efforts should be exerted to manually extract effective features related to expression changes. Many studies have recently investigated FER issues based on deep learning in consideration of FER's great success in pattern recognition, especially with the development of the Emotion Recognition in the Wild Challenge (EmotiW) [21]. A thorough review of deep learning is beyond the scope of this study; however, readers may refer to [9], [22]. This work mainly discusses a few deep networks that can be used to implement FER tasks. Zhao *et al.* proposed deep belief networks (DBNs) to automatically learn facial expression features, and a multilayer perceptron (MLP) was trained to recognize different facial expressions based on the learned features. They argued that MLP outperforms SVM and the RF classifier [23]. Boughrara *et al.* presented a constructive training algorithm for MLP applied to FER applications [24]. Aside from MLP, CNN is also commonly used to extract features simultaneously and classify expressions. Lopes *et al.* presented a CNN for FER and reported its satisfactory performance in the "CK+" dataset. A data augmentation strategy was proposed to address the lack of labeled samples for CNN training. Several pre-processing technologies were also used to preserve expression-related features in facial images. Later, Yu *et al.* combined several CNNs to study FER [25]. These CNNs were fused by learning the set weights of the network response. Kim *et al.* also trained multiple deep CNNs for robust FER [26]. The committee of deep CNNs was improved by varying the network architecture and random weight initialization. To learn improved features specific for expression representation, Liu *et al.* proposed AU-inspired deep networks (AUDNs) inspired by the psychological theory that expressions can be decomposed into multiple facial action units [27]. However, the recognition ability of AUDN is restricted due to the single modality of input facial images. Mollahosseini *et al.* attempted to learn improved features specific for expression representation through a very deep neural network [28]. The network consisted of two convolutional layers, each followed by a max pooling layer and four inception layers. However, this network is difficult to train without using powerful machines (especially powerful GPUs). In short, recent FER approaches based on deep learning outperform traditional FER approaches based on hand-crafted features. However, only a few studies on deep learning employ facial depth images as an input of deep networks.

III. PROPOSED METHOD

A. PRE-PROCESSING

1) FACE DETECTION

Face detection is the key issue in FER. Excessive background information that is uncorrelated to expression recognition exists in a facial image, even when the image is selected from a benchmarking facial expression dataset. Thus, precise FER depends on the accuracy of the results of face detection, which should exclude uncorrelated background information as much as possible. The commonly used Viola–Jones framework [2] is used for face detection in the present study. Certain results of face detection (represented by a yellow rectangle) are illustrated in Fig. 3.

FIGURE 3. Illustration of detected faces with different facial.

2) ROTATION RECTIFICATION

Facial images in benchmarking datasets and real environments vary in rotation, even for images of the same subject. These variations are unrelated to facial expressions and may thus affect the recognition accuracy of FER. To address this issue, the facial region is aligned via rotation rectification by means of a rotation transformation matrix defined as follows:

$$(Lx', Ly', 1) = [Lx, Ly, 1] \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}. \quad (1)$$

where (Lx, Ly) represents the original coordinate in the facial image and (Lx', Ly') represents the coordinate (x, y) after rotation transformation. θ represents the rotation angle formed by the line segment that moves from one eye center to the other. The horizontal axis is zero. We use the DRMF proposed by Cheng *et al.* to detect both eyes in facial images with high accuracy and speed [29]. After rotation rectification, all detected facial regions are rescaled to 72×72 to reduce the dimension. A smaller size of facial region can further accelerate the speed of FER, but it may also lead to losing

of facial information, especially for the information acquired from facial LBP images.

3) CALCULATING LOCAL BINARY PATTERN FACIAL IMAGES

LBP is a commonly used descriptor to capture texture information of the given target. The LBP coding of a given pixel is calculated by comparing its value with adjacent pixels [13]. As shown in Fig. 4, the left part illustrates all pixel values of a local region, whereas the right part provides the LBP coding of the center pixel in the way of binary coding.

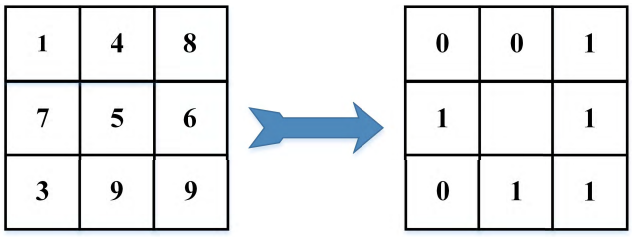


FIGURE 4. Illustration of LBP coding.

After the pixel is effectively encoded by LBP coding, its LBP value can be calculated as follows:

$$LBP = \sum_{n=1}^{N} S(g_n - g_c) * 2^n. \tag{2}$$

where $S(*)$ represents the signature function and N represents number of adjacent pixels. g_c and g_n indicate the values of center pixel and adjacent pixels, respectively. A LBP facial image can be obtained by calculating LBP value of each pixel.

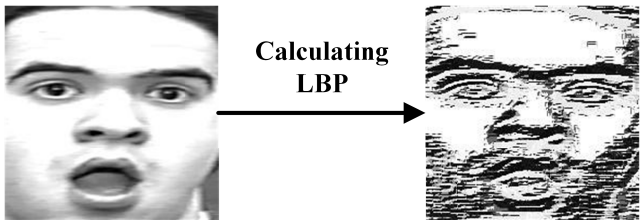


FIGURE 5. Illustration of calculating facial LBP image.

Fig. 5 shows a LBP facial image of surprise expression. Expressions related facial regions, such as mouths, eyes, and eyebrows, are more remarkable in LBP images than in gray scale images.

B. FEATURE EXTRACTION FROM GRAYSCALE FACIAL IMAGES

Lack of sufficient training samples limits the performance of CNN-based FER approach. Data augmentation can partly handle this issue at the risk of over-fitting. Thus, fine-tuning is used to extract expression related features from facial grayscale images by referring to the deep neural network that attained high success in similar tasks.

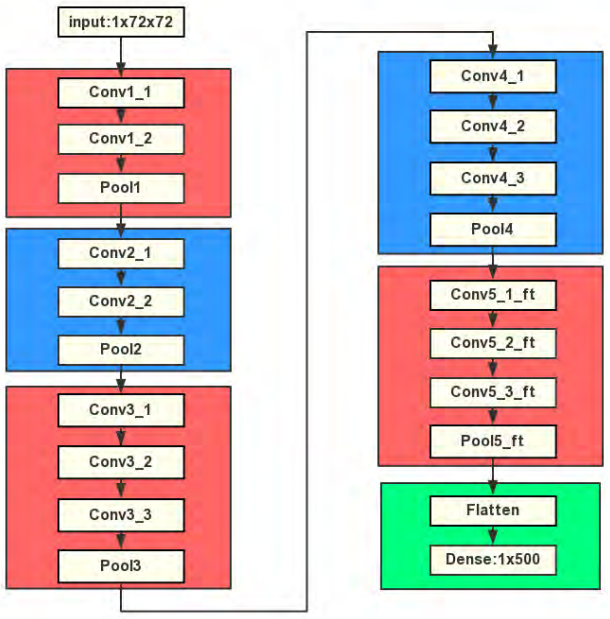


FIGURE 6. Structure of the partial VGG16 network used to extract expression related features from facial grayscale images.

The proposed deep neural network for feature extraction is based on the VGG16 network of Simonyan and Zisserman [30]. VGG16 is chosen because of its effective performance in visual detection and fast convergence. Fig. 6 shows the primary module of the network. Compared with the traditional VGG16, our partial VGG16 network is simplified by removing two dense layers. The dimension of the input data is $1 \times 72 \times 72$. We then fix the structures of the first four blocks. For the fifth block, we change the names of each layer by adding "ft" (ft means fine-tune) at the end of its original name. We also change the structure of Conv5_1_ft. The parameters of layers that belong to this block are shown in Table. 1. Only one dense layer is preserved and its dimension is set to 1×500. We decrease the learning rates of layers that belong to the fifth block by 10 time (0.001 used for layers of the fifth block) than their original values (0.01 used for layers of other blocks) to guarantee that they can learn more effective information. Finally, the initial portion of the network is initialized with weights from a VGG16 model trained on the ImageNet dataset. Rectified Linear Unit (ReLu) activations are applied after each convolutional layer.

TABLE 1. Parameter of layers belonging to block five.

	Conv5_1_ft	Conv5_2_ft	Conv5_3_ft	Pool5_ft
Number	256	256	512	
Size	7×7	3×3	3×3	2×2
Stride	1	1	1	2
Pad	3	0	0	0

C. FEATURE EXTRACTION FROM LBP FACIAL IMAGES

To the best of our knowledge, no elaborate model is trained on LBP images. Thus, we construct a shallow CNN model

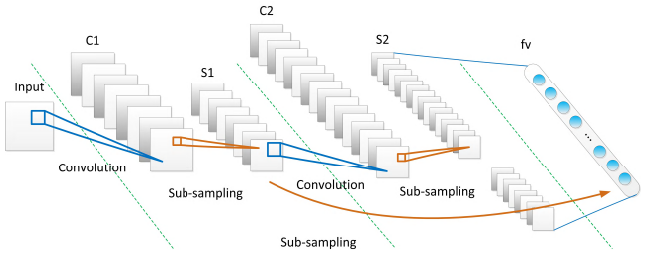


FIGURE 7. Structure of the shallow CNN used to extract features from LBP facial images.

that is similar to DeepID to automatically extract expression-related features from LBP facial images. Fig. 7 illustrates its structure, which comprises an input layer, two convolution ("C") and sub-sampling ("S") layers, and a feature vector ("fv") layer. Sixty-four filters are used in the first convolution layer ("C1") for the input facial LBP images, which focus on the detailed information of facial expressions. This layer uses a convolution kernel of 7×7 and outputs 64 images of 72×72 pixels. This layer is followed by a sub-sampling layer ("S1"), which uses optional max pooling (with kernel size 2×2) to reduce the image to half its size. A new convolution layer ("C2") performs 256 convolutions with a 3×3 kernel to map the previous layer and is followed by another sub-sampling layer ("S2") with a 2×2 kernel. All parameters used in the shallow CNN are listed in detail in Table 2. Then, the output is given to a fully connected hidden layer ("fv") with 500 neurons. The "fv" layer is connected with sub-sampling layers "S1" and "S2" to guarantee the scale invariance of the extracted features. The ability to handle nonlinear data is guaranteed by adding "Relu" activations after sub-sampling layers "S1" and "S2". Data augmentation is used to synthetically increase the number of LBP facial images. Thus, over-fitting can be handled by using the "dropout" operation [31] (parameter was set to 0.5) between the "S" layers ("S1" and "S2") and "fv" layer.

TABLE 2. Parameters of the shallow CNN.

	C1	S1	C2	S2
Number	64		256	
Size	7×7	2×2	3×3	2×2
Stride	1	2	1	2
Pad	3	0	0	0

D. WEIGHTED FUSION OF DIFFERENT OUTPUTS

Fig. 8 shows the proposed weighted fusion network. Expression-related feature vectors fv1 is extracted from facial grayscale images using the partial VGG16 network with the fine-tuning strategy. Feature vector fv2 is extracted from LBP facial images using the shallow CNN. Each feature vector is connected with two cascaded full connect layers for dimension reduction. These full connect layers are fc1_1 = $\{s_1, s_2, \ldots, s_m\}$ (m is experimentally to 100), fc1_2 = $\{s_1, s_2, \ldots, s_6\}$ for fv1 and fc2_1 = $\{l_1, l_2, \ldots, l_m\}$ (m = 100), fc2_2 = $\{l_1, l_2, \ldots, l_6\}$ for fv2. Distances between different

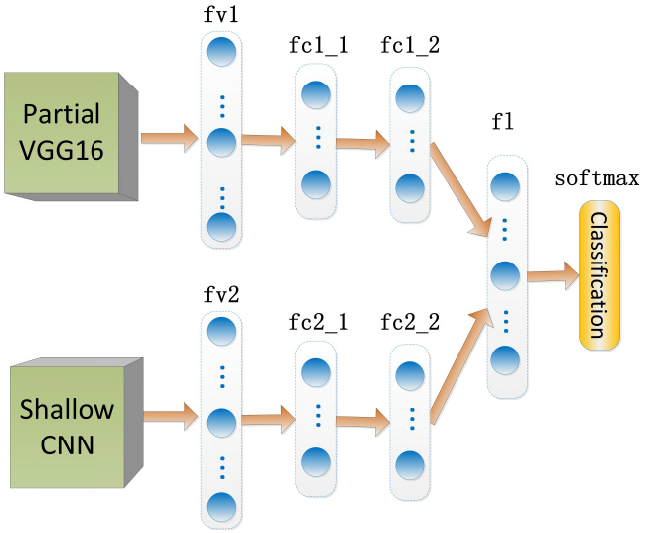


FIGURE 8. Weighted fusion network of binary outputs.

facial features are automatically captured by the network and are revealed through fc1_2 and fc2_2. Further, fc1_2 and fc2_2 are fused in a weighted way to construct a fused vector $fl = \{p_1, p_2, \ldots, p_6\}$. The i^{th} element p_i can be calculated as follows:

$$p_i = \alpha \cdot s_i + (1 - \alpha) \cdot l_i. \tag{3}$$

where α weights the contributions of facial grayscale images and LBP facial images to FER tasks; α is calculated experimentally by cross validation. Softmax classification with a dimension of 6(6 basic expressions) is used to recognize the given expression based on the fused feature vector.

The softmax function produces a categorical probability distribution, when the input is a set of multi-class logits as

$$y_i = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}}. \quad for\ i = 1, \ldots, K \tag{4}$$

where input x is K-dimensional vector ($x = \{x_i\}_{i=1}^{K}$)and output y is also a K-dimensional vector (($y = \{y_i\}_{i=1}^{K}$))with real values in the range (0,1) and that add up to 1 as normalization occurs via the sum of exponent terms that divide the actual exponentiation term. The probability that the class $y = k$ for a given input x and with $j = 1, \ldots, K$ ($K = 6$ in our work) can be written in a matrix form as follows:

$$\begin{bmatrix} P(y=1|x) \\ \vdots \\ P(y=K|x) \end{bmatrix} = \begin{bmatrix} \varsigma(x)_1 \\ \vdots \\ \varsigma(x)_K \end{bmatrix} = \frac{1}{\sum_{j=1}^{K} e^{x_j}} \begin{bmatrix} e^{x_1} \\ \vdots \\ e^{x_K} \end{bmatrix} \tag{5}$$

where $P(y = k|x)$ is the probability that the class is k given that the input is x. The cross entropy is used as the cost function, which is defined as

$$Loss(y, z) = -\sum_{i=1}^{K} Z_i \cdot log(y_i). \tag{6}$$

where z_i indicates the true label and y_i represents the output of softmax function. In the present study, we use the back-propagation (aka backprops) based on gradient descendant optimization algorithm to minimize Eq.6.

IV. EXPERIMENTS RESULTS

A. DATASETS AND CONFIGURATIONS

We evaluate the performance of our FER method based on the Keras framework on the Linux platform. All experiments are performed using a standard NVIDIA GTX 1080 GPU (8 GB), a NVIDIA CUDA framework 6.5, and a cuDNN library. To facilitate fair and effective evaluations, three benchmarking datasets are used, which are composed of facial RGBD images. Descriptions of the used datasets are listed below.

1) CK+ [32]

This fully annotated dataset includes 593 sequences that represent seven expressions (happiness, sadness, surprise, disgust, fear, anger, and neutral) of 123 subjects (males and females). We only use the six basic expressions, including happiness, sadness, surprise, disgust, fear, and anger. For each sequence, we select the last frame because each sequence in this dataset begins with a neutral expression and proceeds to a peak expression. Thus, roughly 80 to 120 samples are selected for each expression. Data augmentation (by using simple operations such as rotation, translation, and skewing) is used to increase the samples of each expression by 50 times. Finally, 10-fold cross validation is used for evaluation.

2) JAFFE [33]

This fully annotated dataset includes 213 samples of 10 Japanese females. The dataset also contains the six basic expressions and a neutral expression. However, we only use the samples of six basic expressions. For each expression, we select all facial images (approximately 30 images) belonging to it. Data augmentation is used to increase the samples of each expression by 100 times. Finally, 10-fold cross validation is used for evaluation.

3) Oulu-CASIA [33]

A total of 10,800 labeled samples are captured from 80 subjects (a mix of male/female and glasses/without glasses). We also use six basic expressions for evaluation. For each expression of each subject, a sequence of facial images is provided. We select the second half of the image sequence and then flip these facial images. Thus, there are approximately 1800 samples for each expression. We do not implement data augmentation for this dataset to avoid the probable over-fitting. Finally, 10-fold cross validation is used for evaluation.

Except for the benchmarking datasets used for quantitative evaluations, we also capture practical facial images for qualitative evaluations. These images are gathered using the Kinect 2.0 sensor (1920×1080). These images contain seven expressions, including happiness, sadness, surprise, disgust, fear, anger, and neutral expressions. We only use six basic expressions in this work. A total of 1960 labeled samples are captured from 28 subjects (a mix of male/female and glasses/without glasses), with 280 samples for each expression. Data augmentation is used to increase the samples of each expression by 10 times and then 10-fold cross validation is used for evaluation. We capture all samples under constant illumination conditions but with heavy occlusion and drastic head deflection to test the robustness of the proposed FER approach. Notably, each practical facial image is down-sampled to 480×270 to reduce calculating amount. Table 3 lists other configurations of the proposed approach, such as learning rate, learning policy, and weight decay.

TABLE 3. Parameter setting of the proposed WMDNN.

Parameters	Value
Learning rate	0.0001
Learning policy	"inv"
Power	0.75
Gamma	0.001
Momentum	0.001
Weight decay	0.005
Performance evaluation	Accuracy

The convergences of the proposed approach are evaluated in three benchmarking datasets, and the results are illustrated in Figs. 9(a), 9(b), and 9(c). Each sub-figure shows the trends of accuracy (red curve) and loss (green curve) with the increase in epochs. For each dataset, the tendencies of accuracy and loss stabilize after 40C50 epochs.

B. ANALYSIS OF THE FUSION WEIGHT

We evaluate the influences of the fusion weight α to the recognition accuracies on three benchmarking datasets. The step to increase α is set as 0.1. $\alpha = 0$ equals to the case, wherein only LBP facial images are used for FER and $\alpha = 1$ indicates another extreme case that uses only facial grayscale images for FER. As shown in Fig. 10, the blue solid curve, the green chain curve, and the red dotted curve represent the results of "CK+," "JAFFE," and "Oulu-CASIA" datasets, respectively. The accuracy of $\alpha = 1$ is higher than that of $\alpha = 0$, thereby indicating that the contribution of facial grayscale images in FER is larger than that of LBP facial images. The fusion approach achieves the highest performance when α is set to 0.7. Thus, in the present study, we manually set the weight α to 0.7.

C. QUANTITATIVE EVALUATIONS OF THE PROPOSED APPROACH

Fig. 11 illustrates the performance of our approach in recognizing six basic expressions in the different datasets. For each dataset, the recognition results are provided through confusion matrixes. For the "CK+" dataset (Fig. 11(a)), our approach recognizes different expressions with very high accuracies (higher than 0.96), except for the expression "disgust" (recognition accuracy of 0.94). For the "JAFFE" dataset (Fig. 11(b)), the recognition accuracies of expression

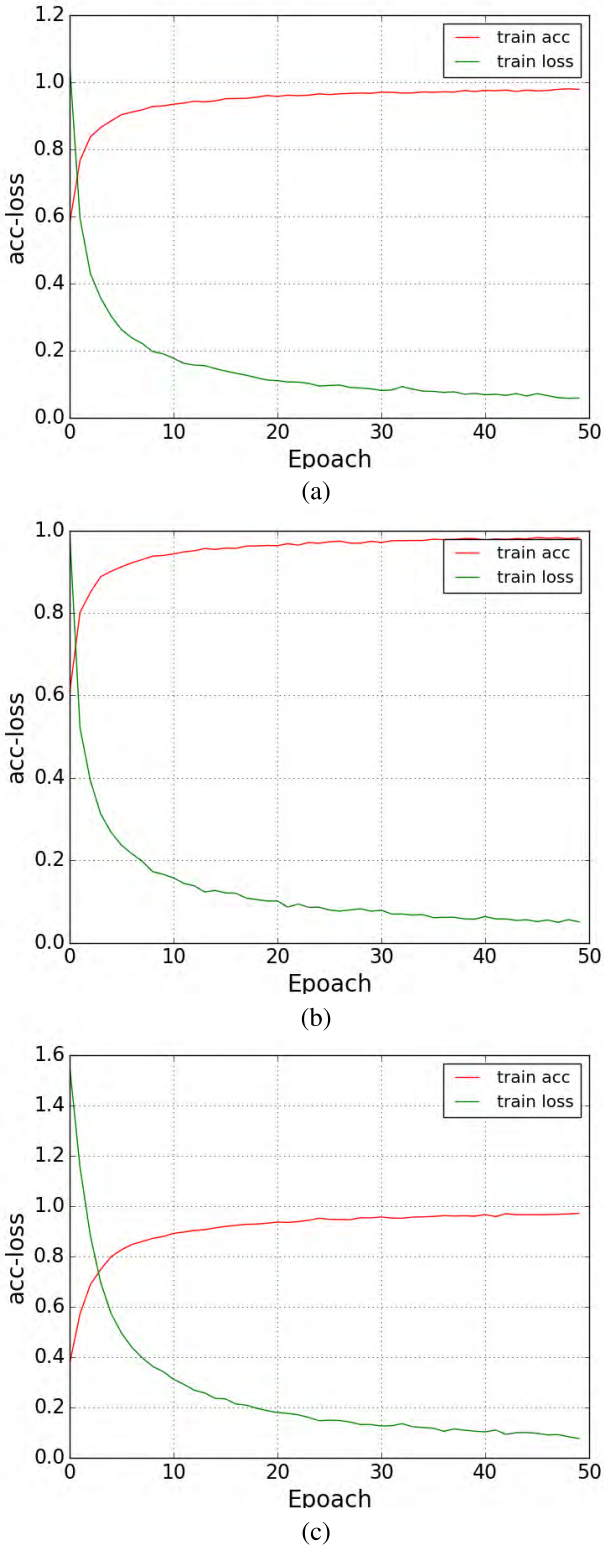


FIGURE 9. Curves of "Loss" and "Accuracy" during training in (a) "CK+", (b) "JAFFE", and (c) "Oulu-CASIA" datasets.

"angry" is as high as 0.95, whereas the expression "disgust" is lower than 0.9. The four remaining expressions, "fear," "happiness," "sadness," and "surprise" have recognition

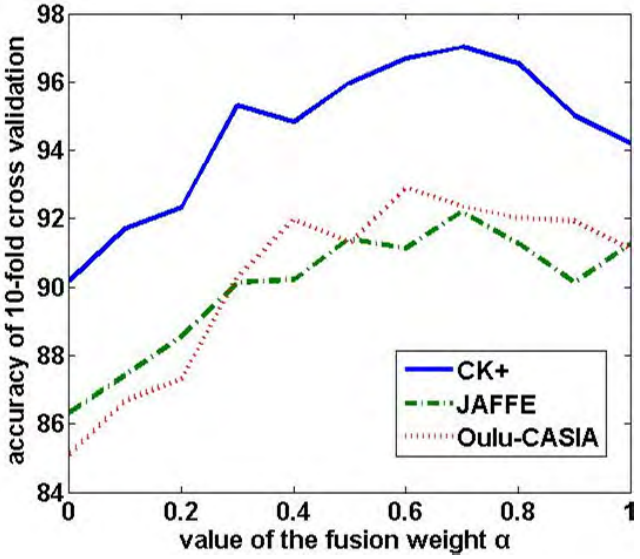


FIGURE 10. Evaluations of fusion weight α on different datasets.

accuracies of around 0.92. The unstable recognition performance is due to the fact that expressions from the "JAFFE" dataset are difficult to distinguish even by manual manipulation. For the "Oulu-CASIA" dataset (Fig. 11(c)), the performance is similar to that in the "JAFFE" dataset. As indicated by the above results, the proposed method can accurately recognize the expressions "anger," "fear," and "happiness" due to their drastic changes in appearance. For the expression "disgust," the proposed approach frequently misclassifies it as "fear" or "sadness". We check this problem and find that the reason is that several subjects in the chosen dataset appear similar when they are showing different expressions. Meanwhile, generalization ability is promised due to the effective combination of LBP and grayscale facial images. A fine-tuning strategy can be used to further improve the generalization ability.

The proposed method and several state-of-the-art methods [17], [34]–[36] are also compared in the benchmarking datasets. We also evaluate the effectiveness of our approach by calculating recognition accuracies based on single channel facial images. We term those approaches as partial ones which include only partial VGG16 for facial grayscale images, shallow CNN for facial grayscale images and LBP facial images, respectively.

For each dataset, the recognition results are listed in Table 4. The parameters of the employed methods are set according to the original work that proposed them. Our method outperforms the methods that use hand-crafted features [34], [35] in both datasets. This result verifies the superiority of deep learning-based FER approaches in automatically extracting expression related features. Aly *et al.* and Rivera *et al.* manually extracted expression-related features, such as HOG and local directional number pattern. Compared with other CNN-based FER methods, our method also achieves better performance than the two

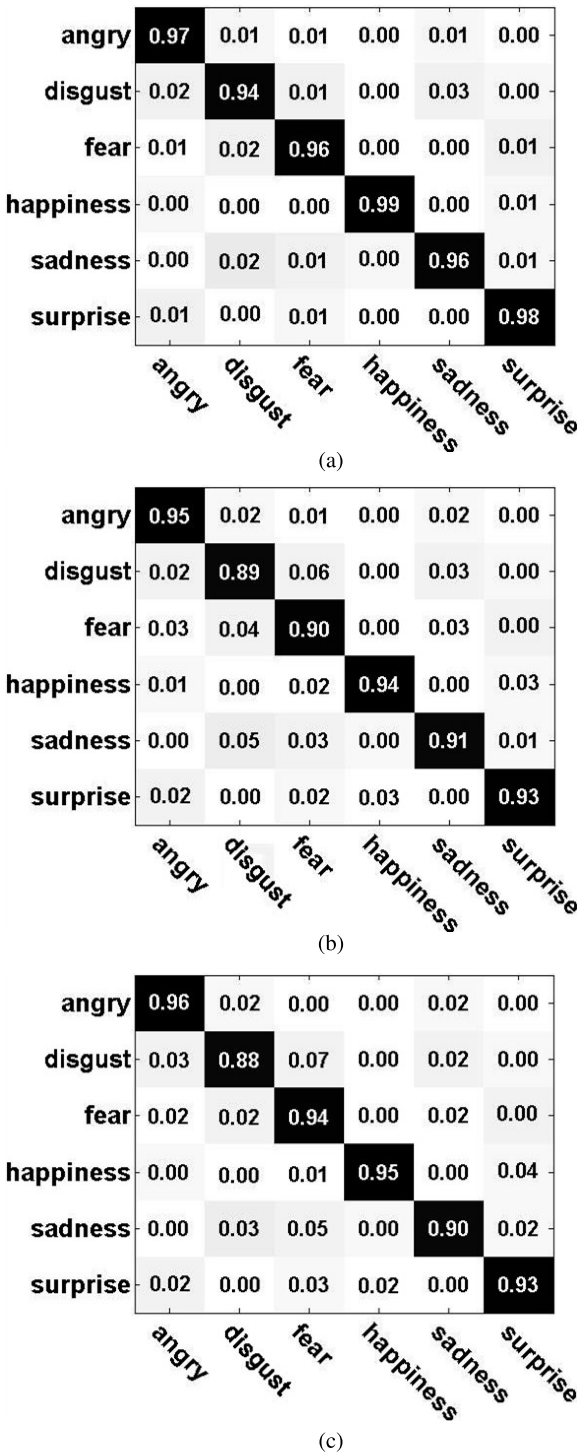


FIGURE 11. Recognition results of six basic facial expressions on three benchmarking datasets. (a) "CK+" dataset. (b) "JAFFE" dataset. (c) "Oulu-CASIA" dataset.

employed methods. For example, our method outperforms the FER method based on a single-modal CNN proposed by Lopes *et al.* [36]]. The advantage of our method is achieved by fully utilizing the complementarity of different channels of facial images, while the other method only uses

TABLE 4. Comparisons between our approach and the state-of-the-art FER approaches.

Employed approach	Dataset		
	"CK+"	"JAFFE"	"Oulu-CASIA"
Aly *et al.*[34]	88.14%	87.32%	84.21%
Rivera *et al.*[35]	91.51%	88.75%	85.18%
Lopes *et al.*[36]	93.68%	88.73%	86.42%
Zhang *et al.*[17]	95.12%	91.48%	87.88%
Partial VGG16 for grayscale images	94.98%	90.86%	87.30%
Only Shallow CNN for grayscale images	93.12%	88.92%	86.73%
Only Shallow CNN for LBP images	92.08%	88.33%	85.52%
Our approach	97.02%	92.21%	92.89%

facial grayscale images. Our method also outperforms the method proposed by Zhang *et al.*, who proposed a similar FER framework based on multi-channel CNN, in both global and local facial regions (regions around the eyes, nose, and mouth) [17]. However, extra effort is necessary to detect facial landmark points, which are useful in finding local facial regions. Wrong detections of local facial regions may decrease the recognition accuracies of Zhang *et al.* Moreover, the employed LBP facial information and the weighted fusion way make our method more accurate in recognizing different facial expressions than Zhang's method. Finally, it is obvious that our whole approach outperforms our partial approaches which only handle single channel facial images. Furthermore, fine-tuning using partial VGG16 achieves the best recognition performance among the three partial approaches due to its ability in extracting effective features from a given image. Only shallow CNN for LBP facial images plays the worst in recognizing different facial expressions in all the three datasets. This result is in line with the conclusion indicated by the evaluations of fusion weight α. Meanwhile, we evaluate our model (trained in "CK+") on the middle frames of each sequence of CK+ dataset and the recognition accuracy is 96.68%. It reveals the effectiveness of our approach in processing dynamic sequences of a facial expression.

D. QUALITATIVE EVALUATIONS OF THE PROPOSED APPROACH

To evaluate the qualitative performance of the proposed approach, practical facial images are collected for evaluation. Partial occlusions are considered to test the robustness of our approach. In each sub-figure, the detected facial region is represented by a red rectangle. The recognition result is shown on the top left corner of the image. The red characters indicate the ground truth of the given facial expression, whereas the blue characters indicate the recognized facial expression with certain recognition accuracy.

Fig. 12 illustrates some cases of successful recognition of facial expressions with high accuracies. All recognition accuracies of different facial expressions are above 0.9, even when the subjects are partially occluded by a notebook. Obviously, drastic changes in appearance exist in these

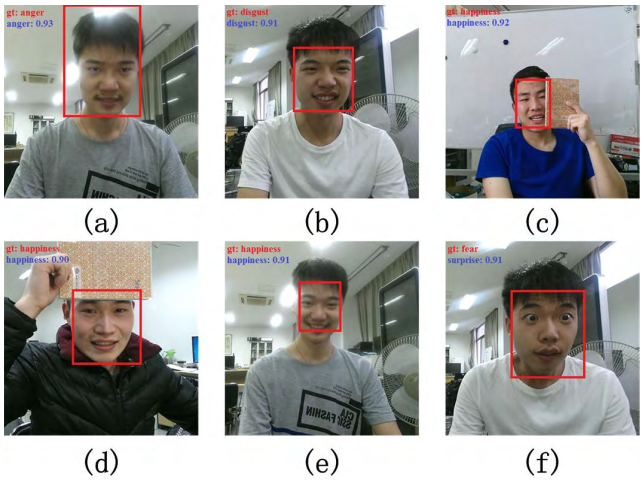


FIGURE 12. Successful recognition of facial expressions with high accuracies.

facial expressions. We can conclude that facial expressions, such as "anger," "happiness," and "surprise" are easy to recognize. This conclusion is in line with the results that are indicated by the three confusion matrixes, as illustrated in Fig. 11.

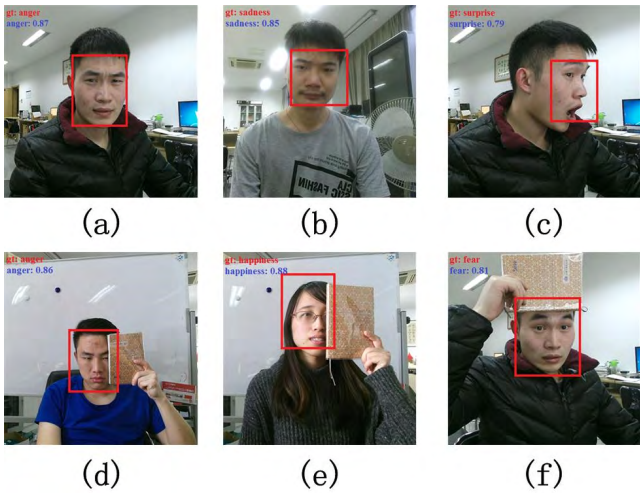


FIGURE 13. Successful recognition of facial expressions with low accuracies.

Fig. 13 illustrates some cases of successful recognition of facial expressions with low accuracies listed from 0.7 to 0.9. Poor performance is caused by several causes, including gentle changes in appearance, partial occlusions, and head deflection. For example, the "anger" expression in Fig. 13(a) is not as obvious as the expressions that appeared in Figs. 12(b) and 12(e). Thus, the recognition accuracy is less than 0.9. As shown in Fig. 13(b), the recognition accuracy of the expression "sadness" is 0.85, which is lower than the recognition accuracies of the expressions "anger" and "happiness". This result is in agreement with the conclusions of three confusion matrixes. Sometimes, an obvious expression may be difficult to recognize because of other factors, except for gentle changes in appearance. For instance, the "surprise" expression is sufficiently obvious, but its recognition accuracy is 0.79 only. This low recognition accuracy is attributed to head deflection, which may lead to loss in face information. Moreover, partial occlusions may influence the recognition of given expressions to a certain extent (Figs. 13(e) and 13(f)), especially when changes in these expressions are insufficiently drastic (Fig. 13(d)).

FIGURE 14. Failed recognition of facial expressions.

Fig. 14 illustrates some cases of failed recognition of facial expressions, which are represented as "Unknown" or a wrong label. Only expressions with recognition accuracies larger than a given threshold (we manually set this threshold as 0.7 and the value can be changed based on specific tasks) are denoted with recognized expressions and corresponding accuracies. Otherwise, "Unknown" is used to indicate the failure of FER. Furthermore, "Unknown" is also used when no faces are detected in the given facial image. The detected facial region in Fig. 14(a) is too small to detect the facial expression, especially when the subject shows gentle changes in appearance. This case is hard to recognize accurately even when occlusion does not occur. For example, the subject in Fig. 14(b) insists that she shows an expression of anger, but our approach cannot recognize the facial expression. We cannot easily recognize her expression through a manual approach. Sometimes, our approach cannot easily detect precise facial regions because of different factors, such as large occlusions (Fig. 14(c)) and poor lighting conditions. Moreover, inaccurate recognition of facial expression is inevitable, especially when the changes of the detected expression in appearance are not so drastic. For example, the subject in Fig. 14(d) revealed the expression "sadness", but our approach wrongly recognized the expression "disgust" with accuracy 0.73.

V. CONCLUSION

This study proposes a FER method based on WMDNN that can process facial grayscale and LBP facial images simultaneously. We argue that both used image channels are complementary, can capture abundant (both local and global) information from the facial images, and can improve the recognition ability. A weighted fusion strategy is proposed to fully use the features that have been extracted from different image channels. A partial VGG16 network is constructed to automatically extract features of facial expressions from facial grayscale images. Fine-tuning is used to train the network with initial parameters obtained from ImageNet.

A shallow CNN is constructed to automatically extract features of facial expressions from LBP facial images because of lack of effective pre-trained model based on LBP images. Subsequently, a weighted fusion strategy is proposed to fuse both features to fully use complementary facial information. The recognition results are obtained based on the fused features via a ''softmax'' operation. Furthermore, it takes about 1.3s to process a facial image, including 0.5s for pre-processing and 0.8s for recognizing different expressions. Evaluations in three benchmarking datasets verify the effectiveness of our approach in recognizing six basic expressions. On the one hand, our method outperforms FER approaches based on hand-crafted features. The ability to automatically extract features enables our method to implement more easily than approaches based on hand-crafted features, which frequently require initially detection of facial landmark points. On the other hand, by utilizing complementary facial information in a weighted fusion manner, our approach outperforms several FER approaches based on deep learning. Our future work will focus on simplifying the network used to speed up the algorithm. Furthermore, we plan to focus on other channels of facial images that can be used to further improve the fusion network.

REFERENCES

[1] C.-R. Chen, W.-S. Wong, and C.-T. Chiu, ''A 0.64 mm^2 real-time cascade face detection design based on reduced two-field extraction,'' *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 19, no. 11, pp. 1937–1948, Nov. 2011.

[2] Y. Q. Wang, ''An analysis of the Viola-Jones face detection algorithm,'' *Image Process. Line*, vol. 4, pp. 128–148, Jun. 2014.

[3] M. Demirkus, D. Precup, J. J. Clark, and T. Arbel, ''Multi-layer temporal graphical model for head pose estimation in real-world videos,'' in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2015, pp. 3392–3396.

[4] S. Jain, C. Hu, and J. K. Aggarwal, ''Facial expression recognition with temporal modeling of shapes,'' in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Nov. 2011, pp. 1642–1649.

[5] M. H. Siddiqi, R. Ali, A. Sattar, A. M. Khan, and S. Lee, ''Depth camera-based facial expression recognition system using multilayer scheme,'' *IETE Tech. Rev.*, vol. 31, no. 4, pp. 277–286, 2014.

[6] M. Valstar, M. Pantic, and I. Patras, ''Motion history for facial action detection in video,'' in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, vol. 1. Oct. 2004, pp. 635–640.

[7] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, ''A convolutional neural network cascade for face detection,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5325–5334.

[8] J. Zhang *et al.*, ''Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition,'' *Pattern Recognit.*, vol. 71, pp. 196–206, Nov. 2017.

[9] Y. LeCun, Y. Bengio, and G. Hinton, ''Deep learning,'' *Nature*, vol. 521, pp. 436–444, May 2015.

[10] Y. Sun, D. Liang, X. Wang, and X. Tang. (2015). ''DeepID3: Face recognition with very deep neural networks.'' [Online]. Available: https://arxiv.org/abs/1502.00873

[11] M. R. Mohammadi, E. Fatemizadeh, and M. H. Mahoor, ''PCA-based dictionary building for accurate facial expression recognition via sparse representation,'' *J. Vis. Commun. Image Represent.*, vol. 25, no. 5, pp. 1082–1092, 2014.

[12] C. Liu and H. Wechsler, ''Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition,'' *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.

[13] C. Shan, S. Gong, and P. W. McOwan, ''Facial expression recognition based on local binary patterns: A comprehensive study,'' *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, 2009.

[14] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, ''DISFA: A spontaneous facial action intensity database,'' *IEEE Trans. Affect. Comput.*, vol. 4, no. 2, pp. 151–160, Apr. 2013.

[15] H. Kobayashi and F. Hara, ''Facial interaction between animated 3D face robot and human beings,'' in *Proc. IEEE Int. Conf. Syst., Man, Cybern., Comput. Cybern. Simulation*, vol. 4. Oct. 1997, pp. 3732–3737.

[16] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. N. Metaxas, ''Learning multiscale active facial patches for expression analysis,'' *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1499–1510, Aug. 2014.

[17] W. Zhang, Y. Zhang, L. Ma, J. Guan, and S. Gong, ''Multimodal learning for facial expression recognition,'' *Pattern Recognit.*, vol. 48, no. 10, pp. 3191–3202, 2015.

[18] K. Mase, ''Recognition of facial expression from optical flow,'' *IEICE Trans. Inf. Syst.*, vol. E74-D, no. 10, pp. 3474–3483, 1991.

[19] G. Zhao and M. Pietikäinen, ''Dynamic texture recognition using local binary patterns with an application to facial expressions,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[20] R. Walecki, O. Rudovic, V. Pavlovic, and M. Pantic, ''Variable-state latent conditional random fields for facial expression recognition and action unit detection,'' in *Proc. 11th IEEE Int. Conf. Workshops Automat. Face Gesture Recognit. (FG)*, May 2015, pp. 1–8.

[21] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon, ''Emotion recognition in the wild challenge 2014: Baseline, data and protocol,'' in *Proc. ACM Int. Conf. Multimodal Interact.*, 2014, pp. 461–466.

[22] J. Schmidhuber, ''Deep learning in neural networks: An overview,'' *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.

[23] Y. Lv, Z. Feng, and C. Xu, ''Facial expression recognition via deep learning,'' *IETE Tech. Rev.*, vol. 32, no. 5, pp. 347–355, 2015.

[24] H. Boughrara, M. Chtourou, C. B. Amar, and L. Chen, ''Facial expression recognition based on a mlp neural network using constructive training algorithm,'' *Multimedia Tools Appl.*, vol. 75, no. 2, pp. 709–731, 2016.

[25] Z. Yu and C. Zhang, ''Image based static facial expression recognition with multiple deep network learning,'' in *Proc. ACM Int. Conf. Multimodal Interact.*, 2015, pp. 435–442.

[26] B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee, ''Hierarchical committee of deep convolutional neural networks for robust facial expression recognition,'' *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 173–189, 2016.

[27] M. Liu, S. Li, S. Shan, and X. Chen, ''Au-inspired deep networks for facial expression feature learning,'' *Neurocomputing*, vol. 159, pp. 126–136, Jul. 2015.

[28] A. Mollahosseini, D. Chan, and M. H. Mahoor, ''Going deeper in facial expression recognition using deep neural networks,'' in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2016, pp. 1–10.

[29] S. Cheng, A. Asthana, S. Zafeiriou, J. Shen, and M. Pantic, ''Real-time generic face tracking in the wild with CUDA,'' in *Proc. 5th ACM Multimedia Syst. Conf.*, 2014, pp. 148–151.

[30] K. Simonyan and A. Zisserman. (2014). ''Very deep convolutional networks for large-scale image recognition.'' [Online]. Available: https://arxiv.org/abs/1409.1556

[31] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, ''Dropout: A simple way to prevent neural networks from overfitting,'' *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

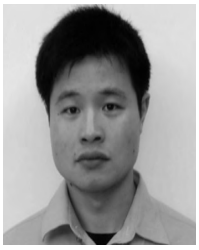
[32] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, ''The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression,'' in *Proc. IEEE Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 94–101.

[33] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, ''Coding facial expressions with Gabor wavelets,'' in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, 1998, pp. 200–205.

[34] S. Aly, A. L. Abbott, and M. Torki, ''A multi-modal feature fusion framework for kinect-based facial expression recognition using dual kernel discriminant analysis (DKDA),'' in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2016, pp. 1–10.

[35] A. R. Rivera, J. R. Castillo, and O. O. Chae, ''Local directional number pattern for face analysis: Face and expression recognition,'' *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1740–1752, May 2013.

[36] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, ''Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order,'' *Pattern Recognit.*, vol. 61, pp. 610–628, Jan. 2017.



BIAO YANG was born in Changzhou, Jiangsu, in 1987. He received the B.S. degree from the College of Automation, Nanjing University of Technology, in 2009, the M.S. and the Ph.D. degrees from the College of Instrument Science and Technology, Southeast University, Nanjing, China, in 2011 and 2014, respectively. Since 2015, he has been a Lecturer with the Department of Information Science and Engineering, Changzhou University. His research interests include pattern recognition and machine learning.

JINMENG CAO was born in Wuxi, China, in 1994. She received the B.S. degree in automation from Changzhou University in 2016. She is currently pursuing the master's degree in machine learning.

RONGRONG NI was born in Nantong, China, in 1987. She received the B.S. degree from the College of Instrument Science and Technology, Southeast University, Nanjing, China, in 2012. She is currently a Research Assistant with the College of Mechanical and Electrical, Changzhou Textile Garment Institute, Changzhou. Her research interests include computer vision and pattern recognition.

YUYU ZHANG was born in Huai'an, China, in 1992. He received the B.S. degree from the Suzhou University of Science and Technology in 2016. He is currently pursuing the master's degree in machine learning.