IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 29, 2019, accepted February 15, 2019, date of publication February 19, 2019, date of current version March 7, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2900231

HERO: Human Emotions Recognition for Realizing Intelligent Internet of Things

WENTAO HUA, (Student Member, IEEE), FEI DAI, (Student Member, IEEE), LIYA HUANG, JIAN XIONG, (Member, IEEE), AND GUAN GUI, (Senior Member, IEEE)

Key Laboratory of Broadband Wireless Communication and Sensor Network Technology, Ministry of Education, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

Corresponding authors: Liya Huang (huangly@njupt.edu.cn) and Guan Gui (guiguan@njupt.edu.cn)

This work was supported in part by the Priority Academic Program Development of Jiangsu Higher Education Institutions, in part by the National Natural Science Foundation of China under Grant 61671253, in part by the Jiangsu Specially Appointed Professor under Grant RK002STP16001, in part by the Innovation and Entrepreneurship of Jiangsu High-Level Talent under Grant CZ0010617002, and in part by the 1311 Talent Plan of Nanjing University of Posts and Telecommunications.

ABSTRACT Human emotions recognition (HERO) is considered as one of the important techniques for realizing the intelligent Internet of Things. The demand for a robust and precise facial expression recognition algorithm is urgent for the HERO. In this paper, we propose a deep recognition algorithm based on the ensemble deep learning model. The proposed algorithm consists of three sub-networks with different depths. Each sub-network is comprised of convolutional neutral networks and trained independently. The sub-network with more convolutional layers recognizes emotions by extracting local details such as the features of eyes and mouth, while the sub-network with less convolutional layers focuses on the macrostructure of the input image. The three sub-networks are assembled together to constitute the whole model. The experiment is based on the Kaggle facial expression recognition challenge database (FER2013), the Japanese female facial expression database, and the AffectNet database. The experimental results show that the proposed algorithm achieves a test accuracy of 71.91%, 96.44%, and 62.11% better than other competitors, and increases the test accuracy by approximately 2–3% than unique sub-networks.

INDEX TERMS Facial expression recognition, ensemble learning, deep learning, convolution neutral networks.

I. INTRODUCTION

Human emotions recognition (HERO) is considered one of important techniques for realizing intelligent Internet of Things (IIoT) which plays the interactive role among human-to-human, human to things [1]– [3] and even security considerations [4]– [7]. Facial expression is one of the most direct human expressions, and it plays an important role in interpersonal communication [8]. Hence, it is necessary to develop robust and precise facial expression recognition algorithm effective for HERO. Through it, people can not only express their intentions correctly but also easily explore other's ideas. However, it is not an easy job for emotionless machines. Therefore, it is important to design a robust and precise facial expression recognition algorithm for human-machine interaction. The process of traditional methods usually includes data preprocessing, feature extraction, and expression recognition. Traditional methods and procedures are not only complex but also easily make error recognition. What's more, it is difficult to guarantee that the features that are extracted from the pre-processed images are effective and positive. In recent years, with the development of deep learning, we witness a new research direction to design the facial expression recognition algorithm and many applications. The algorithm based on deep learning [9] can extract features of pictures, from low to high levels, and gain more abstract features through the connection of several non-linear layers. Deep-learning algorithms can automatically extract useful features and avoid the invalid features extraction problem caused by traditional methods.

Paper [10] designed a discriminative learning convolutional neutral network (CNN) for facial expression recognition. The proposed algorithm combined the central loss function and the verification recognition model to provide

The associate editor coordinating the review of this manuscript and approving it for publication was Jun Wu.

the algorithm with better discrimination ability. Paper [11] applied the conditional generative adversarial network [12] to increase the sample size and used a CNN-based network for facial expression classification. A recognition algorithm based on the visual geometry group model was proposed by Fathallah *et al.* [13]. The model was tested on the CK+ and MUG database and achieved perfect performance. However, the above-mentioned networks are based on the single structure model, and still have much room for improvement.

In this paper, a deep facial expression recognition algorithm for HERO is proposed based on CNNs and the ensemble deep learning algorithm to predict facial expressions. The proposed algorithm combines the advantages of neutral networks and ensemble learning by connecting three subnetworks with different structures. By ensuring the stability of the algorithm, the test accuracy and generalization ability are improved.

The rest of this paper is organized as follows. In section II, the latest research results on facial expression recognition and the application of CNNs in computer vision are introduced. In Section III, the structure and principle of our proposed algorithm are introduced. Section 4 provides the experimental results on the FER2013 and JAFFE datasets. Finally, the conclusion and future work will be introduced in section V.

II. RELATED WORKS

In 1971, Ekman and Friesen [14] defined six basic facial expressions: surprise, fear, disgust, angry, happy, and neutral. Following this, most of the FER research was based on this theory. Pantic and Rothkrantz [15] considered that FER mainly consists of three steps: face detection, feature extraction, and facial expression recognition. Each of these steps affects the accuracy of FER. The research on FER worldwide are conducted based on these three aspects.

At present, the methods of extracting facial features are mainly divided into two categories. The first category is based on deformation. Zhou and Zhang [16] combined the Gobar transform and local binary patterns (LBP) for facial feature extraction and achieved good performance. However, the algorithm has the problem of high computational cost in the processing of the Gobar transform. Kotsia *et al.* [17] extracted texture and shape features of facial expression images and combined the two through neutral networks. The fusion features were then applied to the FER. The second method of feature extraction is based on motion. The experiment of Bassili [18] showed that the visual features of human faces can be identified by describing the movement of facial feature points and analyzing the relationship between them, which laid a theoretical foundation for the movement-based facial features extraction.

In recent years, with the development of neural networks, a large number of new neural networks are proposed, such as forward neural networks [19] (FNNs), CNNs, and recurrent neural networks [20] (RNNs). These neural networks are composed of several neurons and can extract features from low to high level through inner products [21]. Paper [22] proposed a CNN-based recognition algorithm, which used support vector machines (SVMs) to replace the softmax layer [23] and achieved significant gains on the ICML 2013 facial expression database. Nwosu *et al.* [24] designed a model using a two-channel CNNs. The facial images were divided into two parts: the eyes and mouth. The first channel used the extracted eyes as the input, while the second channel used the mouth as the input. The performance of the proposed algorithm was verified on the Jaffe and CK+ dataset and achieved a significant increase in accuracy. Jinwoo *et al.* realized a real-time FER algorithm with high accuracy and low computational cost using CNNs. This proposed algorithm could be applied to real-time human-computer interaction. Sanger *et al.* [19] proposed an ensemble model based on CNNs, which consisted of three subnets with different structure. The proposed model achieved 65.03% accuracy on the FER dataset. Diederich and Wasserschaff [20] designed a facial expression analysis model and improved the assembling of the individual nets through supervised learning. Ma and Leijon [21] designed a specific image pre-processing algorithm and achieved outstanding performance in face expression datasets. Tang *et al.* [22] proposed a CNN based model which is consisted of three sub-CNNs. Compared with the above outstanding researches, our work mainly made the following improvements. First, we proposed a CNN-based ensemble model with three sub-networks and each sub-network was assigned with different weighting factor according to their recognition accuracy. Second, the data augmentation technique was applied to the training and testing phase. Third, a new testing method was proposed to improve the recognition accuracy.

In recent years, deep learning has made brilliant achievements in the field of wireless communications and computer vision, such as object detection and image classification. Paper [25] developed deep learning based non-orthogonal multiple access scheme for obtaining the high sum rate and reducing the computational complexity significantly. H. Huang *et al.* [26] proposed deep learning based super-resolution channel estimation method for millimeter wave (mmWave) massive multiple-input multiple-output (MIMO) systems. Paper [27] proposed an unsupervised learning based fast beamforming design method for downlink MIMO systems. Liu *et al.* [28] proposed a deep learning based message passing algorithm for achieving efficient resource allocation in cognitive radio networks. Paper [29] designed a deep learning based unmanned surveillance systems for observing water levels in the applications of internet of things. Jeon *et al.* [30] designed a real-time object detection method called Redmon [31], which solved the object-detection task as a regression problem. Based on a separate end-to-end network, the transformation from the original image to the object location and class was realized. Girshick [32] proposed a fast region-based CNN (Fast R-CNN) method for object detection to improve the shortcomings of R-CNN [33]. Therefore, it is an inevitable trend to apply deep learning to emotion recognition.

TABLE 1. The structures of the three sub-networks.

Sub-network 1					Sub-network 2					Sub-network 3				
type	kernel	stride	output	dropout	type	kernel	stride	output	dropout	type	kernel	stride	output	dropout
input			48×48×1		input			48×48×1		input			48×48×1	
convolutional	5×5	1	48×48×32		convolutional	5×5	1	48×48×32		convolutional	5×5	1	48×48×64	
Max-pooling	3×3	2	24×24×32		Max-pooling	5×5	2	24×24×32		Max-pooling	5×5	2	24×24×64	
convolutional	4×4	1	12×12×32		convolutional	4×4	1	24×24×32		convolutional	3×3	1	24×24×64	
Max-pooling	3×3	2	24×24×32		Max-pooling	3×3	2	12×12×32		Max-pooling	3×3	2	12×12×64	
convolutional	5×5	1	12×12×32		convolutional	5×5	1	12×12×64		convolutional	3×3	1	12×12×64	
Av-pooling	3×3	2	12×12×64		Max-pooling	3×3	2	8×8×64		Max-pooling	3×3	2	6×6×64	
FC2			1024×1		convolutional	5×5	1	8×8×64		convolutional	3×3	1	6×6×128	
FC1			1024×1	30%	Av-pooling	3×3	2	4×4×64		Max-pooling	3×3	2	3×3×128	
Output			7×1	30%	FC			2048×1		convolutional		1	3×3×128	
					FC1			1024×1	30%	FC1			1024×1	20%
					FC2			1024×1	30%	FC2			1024×1	20%
					Output			7×1	30%	Output			7×1	

III. THE PROPOSED ALGORITHM

Our proposed algorithm is based on deep learning. Combining the advantages of CNNs and ensemble learning, the algorithm can not only automatically extract expression features, but also integrate several weak classifiers together to form a strong classifier. Meanwhile, in the testing phase, we preprocess the test images and create eight images without changing the image information through translation and flip operation. Through the above operations, the accuracy of emotion recognition will be significantly improved.

A. TRAINING DATA PREPROCESS

Our experiments were conducted on the Fer2013 dataset, the JAFFE dataset and the AffectNet dataset. The Fer2013 dataset and the JAFFE dataset are the traditional facial expression dataset which are widely recognized by related researchers. And the AffectNet dataset is the biggest facial expression dataset at present. The training datasets of the three datasets consists of seven different facial expressions categories (angry, disgust, fear, happy, sad, surprise, and neutral). However, the distribution of the training data is usual uneven. For example, to the Fer2013 dataset, the number of training images of disgust is 434, while the other categories of training pictures range from 4,000-7,000. The imbalance of training data distribution will have a negative impact on network performance. To eliminate this negative effect, we performed data augmentation (DG-1) (shown in Figure 1) on the disgust training data. The number of each image in the disgust category can expand to 10 through DG-1. During the training phase, to train our model with more data, we perform another data augmentation (DG-2) (shown in Figure 1) on all of the training data. Through this operation, our total training amount has increased to eight times as much as it was before. In summary, we used the DG-1 operation to balance the data distribution and the DG-2 operation to generate more training data.

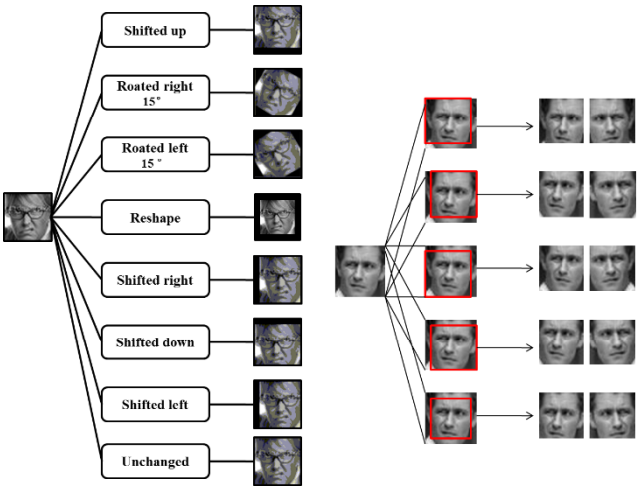

FIGURE 1. The structure of data argument-1 & data argument-2.

B. THE TRAINING MODEL

The proposed algorithm is based on ensemble learning, which combines multiple weak supervision models to obtain a better and more comprehensive strong supervision model. The potential idea of ensemble learning is that even if a weak classifier obtains the wrong prediction, the other weak classifier can also correct the error. Based on this idea, we design three weakly supervised models (sub-net1, sub-net2, and sub-net3) with CNNs. The structures of the three sub-nets are shown in Table 1. To achieve an ensemble model with strong generalization performance, the individual weak classifier in integration should be as independent as possible. Therefore, the structure of each sub-network

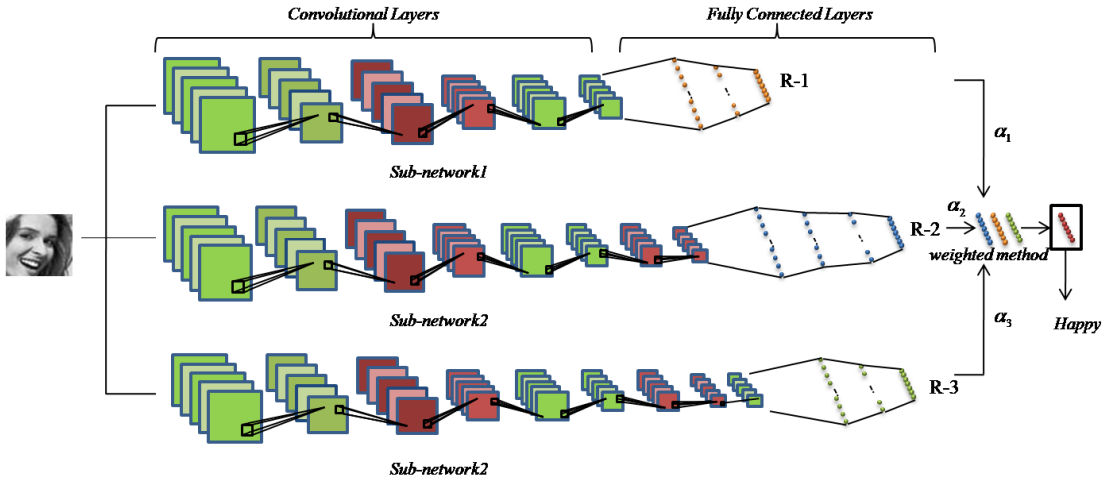

FIGURE 2. The structure of the final test model. R-1, R-2, and R-3 are the probability vectors of sub-network1, sub-network2, and sub-network3 respectively, and α_1, α_2, α_3 are the weights of the sub-networks.

is different, and each sub-network is trained independently. The back propagation algorithm (BP) [34] and stochastic gradient descent algorithm (SGD) [35] were used to minimize the loss function. To overcome the overfitting, the dropout operation [36] was used on the full connection layer. The training dataset for each sub-network is the same. Before training, the dataset was processed by DG-1 and DG-2 to balance the data distribution and generate more training data.

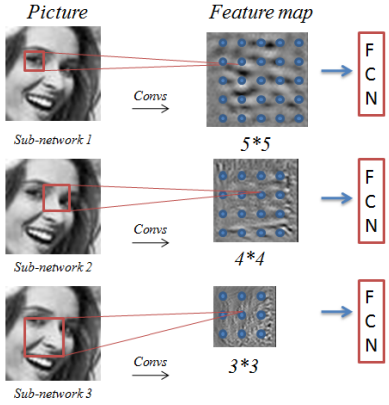

FIGURE 3. The feature maps of three sub-networks.

The three sub networks recognize facial expression from different perspectives. The sub-network1 focuses on extracting local features of images such as the corners of the eye and mouth. The size of the feature map [37] of the last convolutional layers is 5×5 (shown in Figure 3) so the each pixel point contain subtle feature information. The purpose of the sub-network 1 is to recognize facial expression from these local details. At the same time, the feature map of the last convolutional layers of sub-network3 is 3×3 (shown in Figure 3), which means the pixel point contains more feature information. In other words, the sub-network3 focuses on recognizing the facial expression through the relevance between different fine features such as the simultaneous appearance of slightly upturned lips and squinting eyes represents happiness. The sub-network 2 is between the two. The three sub-networks recognize the same expression from different perspectives. Therefore, the recognition accuracy can be greatly improved by combining the three sub-networks.

C. TEST DATA PREPROCESSING

All of the test images will be preprocessed to generate eight images before the testing phase through DG-1. The eight images are generated by translation and flip without changing any feature information of the original image. We combined the recognition results of these generated images using the average weighted method as the final recognition results of each sub-network to the original images. Using this strategy, the influence of the outliers will be eliminated and the recognition accuracy will be improved. We combine the results of the eight generated images as the final result of the sub-network to the test image, referred to as the step average strategy.

The application of ensemble learning can effectively improve the recognition accuracy, but it will also lead to a large increase in computing time. In the testing phase, to improve the efficiency of computing, the parallel computing method on GPU were introduced. As shown in the Figure 4, during the testing phase, the three sub-networks computed on three independent GPUs at the same time. After that, the outputs of the three sub-networks will merge to get the final results through Eq. (2). The application of parallel computing method can vastly reduce testing time, and make the time consumption of ensemble model close to that of single sub-network.

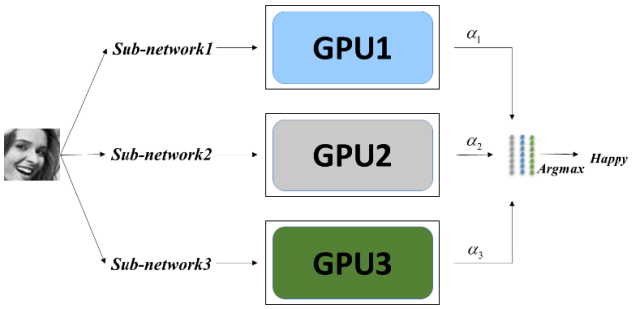

FIGURE 4. The parallel computing strategy on the testing phase.

D. THE TEST MODEL

The final test model, which is shown in Figure2, is obtained by merging the three trained sub-networks. To obtain better combination performance, each sub-network will be assigned a weight, which is based on the performance of each sub-network on the validation dataset. The sub-network with high recognition accuracy will be assigned a large weight and vice versa. The weight formula is defined as follows:

$$\alpha_i = \frac{1}{2} \log \frac{1 - e_i}{e_i} \tag{1}$$

where e_i is the recognition error rate of the sub-network i on the validation dataset. Through these operations, a complete test model is obtained. The specific testing process of an image is shown in the Figure 5.

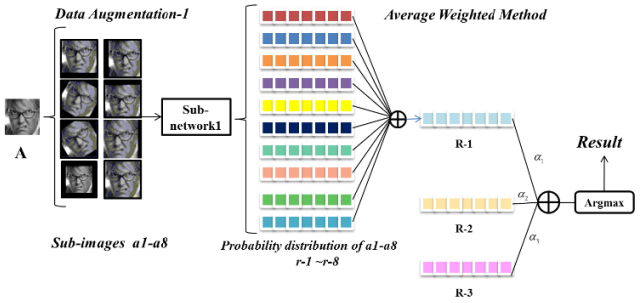

FIGURE 5. The specific process of testing to image A.

Firstly, the tested image A is preprocessed to generate eight sub-images (a1-a8), and the generated images will be sent to sub-network1 in turn. Through the soft-max activation function of the final fully connected layer, the probability vector (each value of this vector represents the probability that the input image belongs to each class) of each sub-image (a1-a8) will be obtained. The weighted average method is applied to combine the eight probability vectors and obtain the final probability vector R-1 of sub-network1 to image A. At the same time, the same operation is performed on the other two sub-networks to obtain the probability vectors R-2 and R-3. Finally, through the weight voting strategy, the probability vector α_i of each sub-network will be weighted according to their respective weights, and the function *argmax* will be used to determine the final category of the image A, as follows:

$$Result = \arg\max \begin{bmatrix} \alpha_1 \times vector(R-1) \\ +\alpha_2 \times vector(R-2) \\ +\alpha_3 \times vector(R-3) \end{bmatrix} \tag{2}$$

IV. EXPERIMENTS AND PEROFORMANCE ANALYSIS

In this session, the performance of our proposed recognition algorithm is shown. We chose Keras to build our model and the training process was implemented on a NVidia 1080Ti GPU to improve the training speed.

A. FER DATASET

The Fer2013 dataset is retrieved from a Kaggle recognition competition (Facial Expression Recognition Challenge). The dataset consists of grayscale images of faces from females of all ages. The size of each image is 48x48 pixels. The dataset is divided into seven categories (angry, disgust, fear, happy, sad, surprise, and neutral). The Fer2013 dataset contains 28,709 training images, 3,589 validation images, and 3,589 test images. The dataset contains facial images of all ages and various directions, including cartoon faces. Fer2013 is an ideal dataset for facial expression recognition. DG-1 and DG-2 (which are proposed in section 2) are implemented on the Fer2013 dataset to balance the data distribution and generate more training data. FER2013 is a challenging and difficult dataset, and the average accuracy for humans on this dataset is approximately 65% (±5).

First, we trained three sub-networks, which are shown in Figure 2, independently on the Fer2013 dataset to ensure that each sub-network has no interference with each other in the testing phase. At the same time, the number of convolutional and fully-connect layers in each sub-network were set to be different. The training epoch was set to 200 and the Adam-optimizer is used to reduce the cross-entropy loss function. The training accuracy (the on-accuracy of the training set) and val accuracy (the accuracy of the test set) on the Fer2013 dataset in shown in Figure 6.

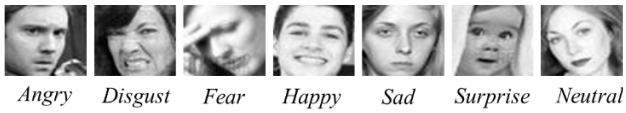

FIGURE 6. A sample of images from Fer2013 dataset.

TABLE 2. The accuracy and val-accuracy of three sub-networks.

	Accuracy	Val-accuracy
Sub-network 1	**0.9611**	**0.6785**
Sub-network 2	**0.9255**	**0.6818**
Sub-network 3	**0.9320**	**0.6815**

The highest val accuracy of each sub-network is shown in Table 2. We note that the recognition accuracy of each sub-network is close to 68%. That is a good result for CNN-based networks, but there is still room for improvement.

TABLE 3. The comparison on the test dataset Fer2013 database.

Method	Accuracy
Eric Cartman	0.64474
Ryank	0.65087
Lor.Voldy	0.65254
Radu+Marius+Cristi	0.67483
Sub-network 1	**0.6785**
Sub-network 2	**0.6818**
Sub-network 3	**0.6815**
Maxim Milakov	0.68821
Unsupervised	0.69267
RBM	0.71161
Proposed method	**0.71911**

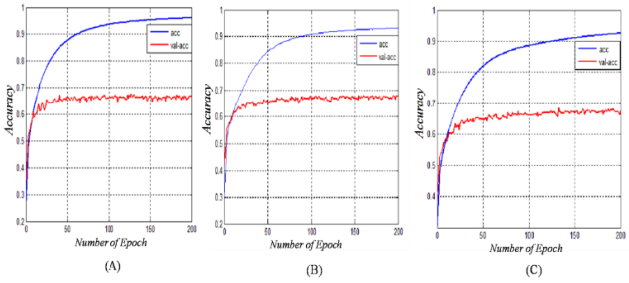

FIGURE 7. The training-accuracy and val-accuracy curves of the three sub-networks on the FER2013 dataset. (A)Sub-network 1. (B) Sub-network 2. (C) Sub-network 3.

Simple voting and weight-voting strategies based on ensemble learning were used to combine the three sub-networks into a complete recognition network. The exact value of α_i was calculated using Eq. (1), and the proposed algorithm performance was tested using the test dataset. We eventually achieved an accuracy of 71.191% on the test set. Compared with the sub-networks, the recognition accuracy increased approximately 3%. A comparison of the performance between our proposed algorithm and other recognition algorithms is shown in Table 3.

The confusion matrix of the recognition accuracy for seven different facial expressions is shown in Table 4. The final recognition accuracy on the Fer2013 dataset using our method is approximately 71.91%. Table 4 shows that the recognition accuracy of angry, disgust, fear, and sad are lower than the final recognition accuracy. This phenomenon occurs for two reasons. First, the mentioned facial expressions are easy to confuse with other facial expressions. For example, fear is usually confused with sad and angry. On the other hand, the quality of the training data is not high. The training data for fear and sad are hard to differentiate on the training set. The method achieves high recognition accuracy on happy, surprise, and neutral, because the three facial expressions are characteristic and easy to recognize.

Lastly, in terms of computational efficiency we calculated the recognition time of each picture during the testing phase on an NVidia 1080Ti GPU. The calculation was performed on both the ensemble model and the three sub-networks. On average, one picture in the test dataset needs 0.823 seconds to be tested in a sub-network. At the same time, the same picture needs 2.518 seconds to be tested in the proposed ensemble model. The confusion matrix of the recognition accuracy for seven different facial expressions is shown in Table 4.

The final recognition accuracy on the Fer2013 dataset using our method is approximately 71.91%. Table 4 shows that the recognition accuracy of angry, disgust, fear, and sad are lower than the final recognition accuracy. This phenomenon occurs for two reasons. First, the mentioned facial expressions are easy to confuse with other facial expressions. For example, fear is usually confused with sad and angry. On the other hand, the quality of the training data is not high. The training data for fear and sad are hard to differentiate on the training set. The method achieves high recognition accuracy on happy, surprise, and neutral, because the three facial expressions are characteristic and easy to recognize. Lastly, in terms of computational efficiency we calculated the recognition time of each picture during the testing phase on NVidia 1080Ti GPUs. The calculation was performed on both the ensemble model and the three sub-networks. A comparison was made between the test method with the parallel computing (PC) strategy and the method without the PC, and the results are shown in Table 5. On average, one picture in the test dataset needs 0.823 seconds to be tested in a single sub-network. At the same time, the same picture needs 2.518 seconds to be tested in the proposed ensemble model without the PC strategy. However, when the PC strategy is applied, the time-consuming will be shortened to 0.865s which is close to the cost of single sub-network.

B. JAFFE DATASET

The JAFFE dataset consists of 213 grayscale facial images with seven facial expressions (angry, disgust, neutral, sad, surprise, happy, and fear) posed by seven Japanese female models. The size of each image is 256 × 256 pixels. The JAFFE dataset is divided into three parts: the training set, the validation set, and the test set. The training set contains 129 images (60%), the test set contains 42 images (20%), and the validation set contains 42 images (20%). The dataset was created by Michael Lyons, Miyuki Kamachi, and Jiro Gyoba of Kyushu University. It is a traditional facial expression dataset and most facial expression recognition tasks are based on this dataset.

Because the image size of the input layer of the proposed algorithm is 48 × 48 × 1, we first standardized the size of all pictures to 48 × 48 × 1. Then, we used DG-1 (which is proposed in section 2) to balance the data distribution. The JAFFE dataset does not need this operation, because the data distribution is balanced. Finally, DG-2 (which is proposed in section 2) is used to generate more training data. We trained

TABLE 4. The confusion matrix on the test dataset from FER2013 database.

	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Angry	**66.7%**	0.41%	5.7%	3.5%	14.8%	1.0%	7.9%
Disgust	25.5%	**63.4%**	1.8%	3.6%	3.6%	0%	1.8%
Fear	10.6%	0%	**48.4%**	4.3%	22.9%	5.6%	7.9%
Happy	1.8%	0%	1.3%	**88.1%**	4.3%	1.3%	3.1%
Sad	6.4%	0%	6.1%	6.1%	**65.5%**	0.5%	15.5%
Surprise	2.1%	0%	6.7%	5.3%	3.1%	**79.7%**	2.8%
Neutral	1.7%	0.16%	2.4%	3.5%	16.3%	0.9%	**74.8%**

TABLE 5. The time consumption of FER dataset in the testing phase.

	Consumption of time (second)
Single sub-network	**0.823s**
Ensemble model (without PC)	**2.518s**
Ensemble model (with PC)	**0.865s**

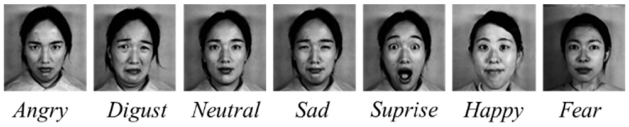

FIGURE 8. A sample of images from the JAFFE dataset.

the three sub-networks using the generated train dataset. The weights α_1 of each sub-network are calculated using Eq. (1). The three sub-networks are combined according to the principle of ensemble learning. Finally, we tested the performance of our proposed algorithm on the test dataset of JAFFE using the test preprocessing and weight voting strategy. The recognition accuracy of each sub-network is approximately 94%-95%. After combining the three sub-networks as into complete recognition network using the ensemble learning algorithm, the recognition accuracy is greatly improved and is approximately 96.44%. The comparison between our algorithm and other recognition algorithms is shown in Table 6, which shows that the recognition accuracy for each facial expression is approximately 96%.

The method achieves high accuracy on happy, disgust, neutral, and surprise, which are easier to express. It achieves relatively low recognition accuracy on fear, sad, and angry, which are easy to confuse with other facial expressions. In terms of computational cost, one picture needs 0.816 seconds to be tested in a sub-networks and 2.562 seconds in the ensemble model. The recognition accuracy of each facial expression using our proposed model is shown in Table 7, which shows that the recognition accuracy for each facial expression is approximately 96%. The method achieves high accuracy on happy, disgust, neutral, and surprise, which are easier to express. It achieves relatively low recognition accuracy on fear, sad, and angry, which are easy to confuse with other facial expressions. In terms of computational cost, one picture needs 0.816 seconds to be tested in a sub-networks on average and 2.562 seconds in the ensemble model without PC strategy. When the PC strategy is applied, the time-consuming will be shortened to 0.859 seconds.

C. AFFECTNET DATASET

The AffectNet [43] dataset is one of the biggest facial dataset for facial expression research. It contains more than 1,000,000 facial images collected from three search engines: Google, Bing and Yahoo. All in total 1250 different related keywords were applied to query the three search engines. Figure 9 shows some facial images from the AffectNet dataset.

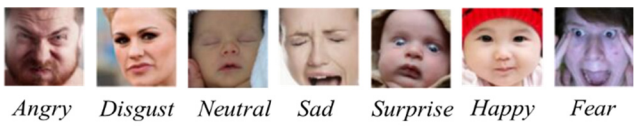

FIGURE 9. A sample of images from the AffectNet dataset.

About half of the facial images (420,299) in AffectNet dataset were tagged by twelve professional annotators. These facial images were annotated using 11 different categories: Neutral, Happiness, Sadness, Surprise, Fear, Anger, Disgust, Contempt, None, Uncertain and Non-face. The dataset is divided into three parts: the training dataset, the validation dataset and the testing dataset. However, the testing dataset is not released at present. According to Mollahosseini *et al.* [43] advice, in this paper, our experiment was carried on a total of 287,400 images with 7 basic facial expressions: Neutral, Happiness, Sadness, Surprise, Fear, Anger and Disgust. Table 9 shows the number of facial images in the training dataset, the validation dataset and the testing dataset.

Different from the structure shown in Figure 2, we chose VGG19 [45] as the sub-network 1, Resnet50 [46] as the sub-network2 and Resnet101 as the sub-network3. Figure 10 shows the structure of the ensemble model which we applied on the AffectNet dataset. An extra fully-connected layer (FC) was added after each sub-network. Softmax function was

TABLE 6. The comparison on the test dataset from JAFFE database.

Method	Accuracy
KCCA [38]	0.7705
Information Projection [5]	0.8318
Classifier Selection [39]	0.8592
Gabor [40]	0.9330
Sub-network 2	**0.9375**
Two-channel CNNs [41]	0.9440
Sub-network 3	**0.9465**
Sub-network 1	**0.9531**
SVM [42]	0.9560
Proposed method	**0.9644**

TABLE 7. The confusion matrix on the test dataset from JAFFE database.

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	**96.2%**	1.4%	0.4%	0.2%	0.2%	1.6%	0%
Disgust	1.4 %	**96.8%**	0.1%	0%	0.6%	0.7%	0.4%
Fear	1.0%	1.3%	**95.8%**	0.4%	0 %	1.3%	0.2%
Happy	0%	0%	0.3%	**97.8%**	0.7%	0%	1.2 %
Neutral	0%	1.2%	0%	0.3%	**96.5%**	1.1%	0.9%
Sad	1.2%	1.4%	0.8%	0.1%	0.3%	**96.2%**	0%
Surprise	0.1%	0.2%	0.2%	1.1%	0.5%	0%	**97.9%**

TABLE 8. The time consumption of JAFFE dataset in the testing phase.

	Consumption of time (Second)
Single sub-network	0.816s
Ensemble model (without PC)	2.562s
Ensemble model (with PC)	0.859s

TABLE 9. Train, validation, and test split of the AffectNet database.

Expression	Train	Validation	Test
Neutral	72357	3016	500
Happy	129515	5400	500
Sad	24919	1040	500
Surprise	14006	584	500
Fear	6602	276	500
Disgust	4131	172	500
Anger	24366	1016	500
Total	275896	11504	3500

selected as the activation function. The number of the neutrons in the last FC layer was set to 7 to make sure that the dimension of the probability vector of each sub-network was 7.

The average size of the facial images in the AffectNet dataset is 425×425 pixels. All the original facial images were pretreated by graying to reduce the scale of data processing and rescaled to 224×224 pixels. DG-1 (which is proposed in

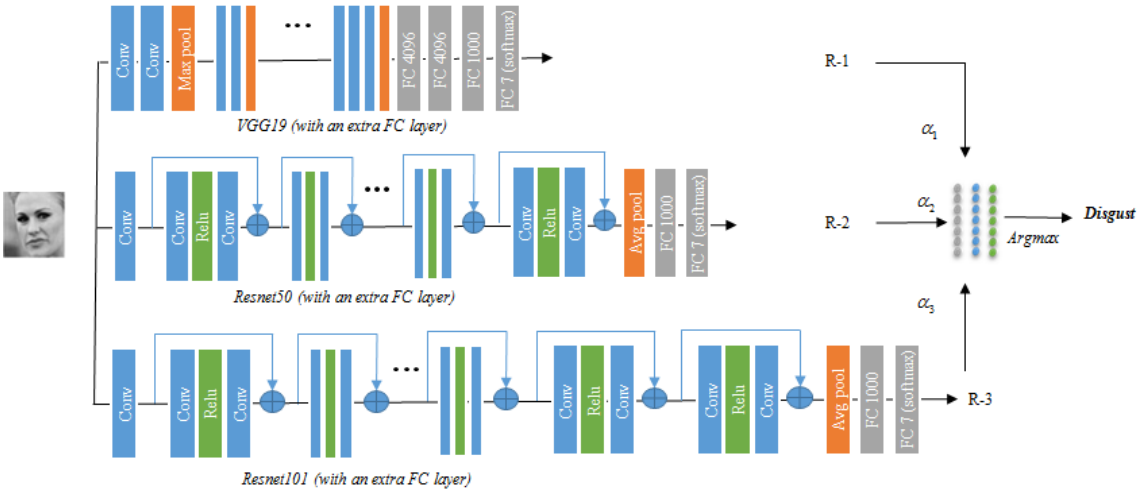

FIGURE 10. The structure of the ensemble model on the AffectNet dataset.

TABLE 10. The comparison on the test dataset from the AffectNet dataset.

METH

Method	Accuracy
Sub-network 1	0.5740
Baseline [43]	0.5800
Sub-network 2	0.5891
Sub-network 3	0.5937
VGG-FACE [44]	0.6000
Proposed method	**0.6211**

TABLE 11. The confusion matrix on the test dataset from the AffectNet database.

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	**57.4%**	15.6%	8.0%	2.0%	6.6%	5.8%	4.6%
Disgust	6.8%	**58.6%**	16.0%	3.4%	4.6%	5.8%	4.8%
Fear	6.8%	5.2%	**61.4%**	4.6%	3.4 %	6.4%	12.2%
Happy	4.4%	4.2%	2.6%	**70.2%**	6.2%	4.0%	8.4 %
Neutral	5.2%	3.6%	8.4%	6.8%	**58.8%**	9.2%	8.0%
Sad	7.2%	6.4%	4.8%	4.2%	12.8%	**62.6%**	2.0%
Surprise	4.6%	5.2%	6.0%	6.2%	8.8%	3.4%	**65.8%**

TABLE 12. The time consumption of the AffectNet dataset in the testing phase.

	Consumption of time (Second)
Single sub-network	**1.232s**
Ensemble model (without PC)	**3.749s**
Ensemble model (with PC)	**1.656s**

section 2) was used to balance the training data distribution and generate more training data. The base learning rate of each sub-network was set to 0.001 and decreased step-wise by a factor of 0.1 every 30 epochs. The weight α_i of each sub-network were calculated by Equation (1). During the testing phase, DG-2 (which is proposed in section 2) was used to further improve the recognizing accuracy of the ensemble model. In terms of evaluation criteria, we adopted the

TABLE 13. Training and testing accuracy for each architecture on the FER dataset.

Model structure	Accuracy(training)	Accuracy(testing)
Resnet101	**0.9912**	**0.5823**
VGG19	**0.9845**	**0.6231**
CNN(sub-network)	**0.9611**	**0.6685**

weighted-loss method for evaluation as Mollahosseini *et al.* mentioned in their paper [43]. Table 10 shows the recognition accuracy of our proposed model and other relative algorithms.

The confusion matrix of the recognition accuracy for seven different facial expressions is shown in Table 11. Our model achieved an acceptable accuracy on Happy (70.2%) and Surprise (65.8%) categories but did not perform well in Neutral (58.8%), Angry (57.4%) and Disgust (58.6%) categories. This phenomenon may be caused by two reasons. On the one hand, the size of the training data may affect the recognition accuracy. On the other hand, some emotions, such as Neutral and Angry, are easily confused with other emotions. Generally, our proposed model achieved 4.11% increase compared to the baseline reported by He *et al.* [45]. In terms of computational cost, one picture needs 1.232 seconds to be tested in a sub-networks on average and 3.749 seconds in the ensemble model without PC strategy. When the PC strategy is applied, the time-consuming will be shortened to 1.656 seconds.

V. CONCLUSION AND FUTURE WORK

In this paper, a deep recognition algorithm based on deep learning and ensemble learning was proposed for HERO in IIoT. We adopted three CNNs-based models with different structures as the sub-networks and trained these sub-networks independently. The structures of three sub-networks which we adopted to carry out experiment on the FER and JAFFE dataset were simple and had less layers. Because of the huge data scale of AffectNet dataset, deeper neutral sub-networks were used to make up the ensemble model of the AffectNet dataset. According to the val-accuracy, each sub-network was assigned a weight. Simple voting and weighted voting strategies were applied to combine the three sub-networks into the final ensemble model. During the test phase, the average strategy was used to further improve the recognition accuracy and the parallel computing method was used to improve the computational efficiency. The results of the experiments on the Fer2013, JAFFE and AffectNet datasets are ideal, indicating that our proposed facial expression recognition algorithm has a relatively good performance compared with other algorithms.

However, there is room for improvement. First, during the training phase, the training set of each sub-network is the same, which may have a negative effect on the final recognition results and lead to overfitting. We can apply the bagging method to randomly extract images to build the training set to solve the problem. Second, in the future, we can consider using an unsupervised pre-training strategy from transfer learning, which may further decrease the recognition error rate.

REFERENCES

[1] M. Mohammadi, A. Al-Fuqaha, S. Sorour, and M. Guizani, "Deep learning for IoT big data and streaming analytics: A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 2923–2960, 4th Quart., 2018.
[2] J. Wu, M. Dong, K. Ota, J. Li, and Z. Guan, "Big data analysis-based secure cluster management for optimized control plane in software-defined networks," *IEEE Trans. Netw. Service Manag.*, vol. 15, no. 1, pp. 27–38, Mar. 2018.
[3] H. Li, K. Ota, and M. Dong, "Learning IoT in edge: Deep learning for the Internet of Things with edge computing," *IEEE Netw.*, vol. 32, no. 1, pp. 96–101, Jan./Feb. 2018.
[4] W. Zheng, X. Zhou, C. Zou, and L. Zhao, "Facial expression recognition using kernel canonical correlation analysis (KCCA)," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 233–238, Jan. 2006.
[5] H. Wang, S. Chen, Z. Hu, and W. Zheng, "Locality-preserved maximum information projection," *IEEE Trans. Neural Netw.*, vol. 19, no. 4, pp. 571–585, Apr. 2008.
[6] N. Zhang, N. Lu, N. Cheng, J. W. Mark, and X. S. Shen, "Cooperative spectrum access towards secure information transfer for CRNs," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 11, pp. 2453–2464, Nov. 2013.
[7] J. Wu, M. Dong, K. Ota, J. Li, and Z. Guan, "FCSS: Fog computing based content-aware filtering for security services in information centric social networks," *IEEE Trans. Emerg. Topics Comput.*, to be published. doi: 10.1109/TETC.2017.2747158.
[8] K. Zhao, H. Zhang, Z. Ma, Y.-Z. Song, and J. Guo, "Multi-label learning with prior knowledge for facial expression analysis," *Neurocomputing*, vol. 157, pp. 280–289, Jun. 2015.
[9] Y. Sun, X. Wang, and X. Tang, "Deep deep learning face representation from predicting 10,000 classes," in *Proc. CVPR*, 2014, pp. 1891–1898.
[10] Z. Li, "A discriminative learning convolutional neural network for facial expression recognition," in *Proc. IEEE Int. Conf. Comput. Commun.*, Dec. 2017, pp. 1641–1646.
[11] H. Yang, Z. Zhang, and L. Yin, "Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2018, pp. 294–301.
[12] M. Mirza and S. Osindero. (2014). "Conditional generative adversarial nets," [Online]. Available: https://arxiv.org/abs/1411.1784
[13] A. Fathallah, L. Abdi, and A. Douik, "Facial expression recognition via deep learning," in *Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl.*, Oct./Nov. 2017, pp. 745–750.
[14] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *J. Pers. Soc. Psychol.*, vol. 17, no. 2, pp. 124–129, 1971.
[15] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1424–1445, Dec. 2000.
[16] Q. Zhou and S. Zhang, "A comparative study of geometry, Gabor wavelets representation and local binary patterns for facial expression recognition," *Adv. Biomed. Eng.*, vol. 11, p. 200, Jan. 2012.
[17] I. Kotsia, S. Zafeiriou, N. Nikolaidis, and I. Pitas, "Texture and shape information fusion for facial action unit recognition," in *Proc. Int. Conf. Adv. Comput.-Hum. Interact.*, 2008, pp. 77–82.
[18] J. N. Bassili, "Facial motion in the perception of faces and of emotional expression," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 4, no. 3, pp. 373–379, 1978.
[19] T. D. Sanger, "Optimal unsupervised learning in feedforward neural networks," *Neutral Netw.*, vol. 2, no. 4, pp. 459–473, 1989.

[20] J. Diederich and M. Wasserschaff, "Recurrent neural networks for sequence production," in *Proc. IJCAI*, 1993, pp. 1114–1119.
[21] Z. Ma and A. Leijon, "Human skin color detection in RGB space with Bayesian estimation of beta mixture models," in *Proc. Eur. Signal Process. Conf.*, 2010, pp. 1204–1208.
[22] Y. Tang. (2015). "Deep learning using linear support vector machines." [Online]. Available: https://arxiv.org/abs/1306.0239
[23] D. Opitz and R. Maclin, "Popular ensemble methods: An empirical study," *J. Artif. Intell. Res.*, vol. 11, pp. 169–198, Aug. 1999.
[24] L. Nwosu, H. Wang, J. Lu, I. Unwala, X. Yang, and T. Zhang, "Deep convolutional neural network for facial expression recognition using facial parts," in *Proc. IEEE 15th Int. Conf. Dependable, Auto. Secure Comput., 15th Int. Conf. Pervasive Intell. Comput., 3rd Int. Conf. Big Data Intell. Comput. Cyber Sci. Technol. Congr. (DASC/PiCom/DataCom/CyberSciTech)*, Nov. 2017, pp. 1318–1321.
[25] G. Gui, H. Huang, Y. Song, and H. Sari, "Deep learning for an effective nonorthogonal multiple access scheme," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8440–8450, Sep. 2018.
[26] H. Huang, J. Yang, Y. Song, H. Huang, and G. Gui, "Deep learning for super-resolution channel estimation and DOA estimation based massive MIMO system," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8549–8560, Sep. 2018.
[27] H. Huang, W. Xia, J. Xiong, J. Yang, G. Zheng, and X. Zhu, "Unsupervised learning-based fast beamforming design for downlink MIMO," *IEEE Access*, vol. 7, pp. 7599–7605, 2019.
[28] M. Liu, J. Yang, T. Song, J. Hu, and G. Gui, "Deep learning-inspired message passing algorithm for efficient resource allocation in cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 641–653, Jan. 2018.
[29] J. Pan, Y. Fan, H. Dong, S. Fan, J. Xiong, and G. Gui, "Image-based detecting the level of water using dictionary learning," in *Proc. Int. Conf. Commun., Signal Process., Syst. (CSPS)*, 2018, pp. 1–10.
[30] J. Jeon *et al.*, "A real-time facial expression recognizer using deep neural network," in *Proc. Int. Conf. Ubiquitous Inf. Manage. Commun.*, 2016, pp. 1–4.
[31] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (CVPR)*, Jun. 2016, pp. 779–788.
[32] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (CVPR)*, Dec. 2015, pp. 1440–1448.
[33] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (CVPR)*, Jun. 2014, pp. 580–587.
[34] S.-I. Horikawa, T. Furuhashi, and Y. Uchikawa, "On fuzzy modeling using fuzzy neural networks with the back-propagation algorithm," *IEEE Trans. Neural Netw.*, vol. 3, no. 5, pp. 801–806, Sep. 1992.
[35] M. Zinkevich, A. Smola, L. Li, and M. Weimer, "Parallelized stochastic gradient descent," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 23, 2010, pp. 1–9.
[36] 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.
[37] P. K. Rana, Z. Ma, J. Taghia, and M. Flierl, "Multiview depth map enhancement by variational Bayes inference estimation of Dirichlet mixture models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 1528–1532.
[38] W. Zheng, X. Zhou, C. Zou, and L. Zhao, "Facial expression recognition using kernel canonical correlation analysis (KCCA)," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 233–238, Jan. 2006.
[39] M. Kyperountas, A. Tefas, and I. Pitas, "Salient feature and reliable classifier selection for facial expression classification," *Pattern Recognit.*, vol. 43, no. 3, pp. 972–986, 2010.
[40] S. A. M. Alsumaidaee, S. S. Dlay, W. L. Woo, and J. Chambers, "Facial expression recognition using local Gabor gradient code-horizontal diagonal dedscriptor," in *Proc. IEI Int. Conf. Intell. Signal Process.*, 2016, pp. 1–5.
[41] D. Hamester, P. Barros, and S. Wermter, "Face expression recognition with a 2-channel Convolutional Neural Network," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Killarney, Ireland, Jul. 2015, pp. 1–8.
[42] H. C. Santiago, T. I. Ren, and G. D. C. Cavalcanti, "Facial expression recognition based on motion estimation," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2016, pp. 27611–27621.
[43] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affective Comput.*, to be puiblished. doi: 10.1109/TAFFC.2017.2740923.
[44] D. Kollias, S. Cheng, E. Ververas, I. Kotsia, and S. Zafeiriou. (2018). "Generating faces for affect analysis." [Online]. Available: https://arxiv.org/abs/1811.05027
[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 2058–2072.
[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778

WENTAO HUA received the B.Sc. degree in communication engineering from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2016, where he is currently pursuing the master's degree. His current research interest includes deep learning-based image processing.

FEI DAI is currently pursuing the bachelor's degree in communication engineering with the Nanjing University of Posts and Telecommunications, Nanjing, China. Her current research interest includes deep learning-based image processing.

LIYA HUANG is currently a full-time Professor with the Nanjing University of Posts and Telecommunications, Nanjing, China. She is also the Vice-Dean of the Bell Honor School. She has published over 40 peer reviewed journal papers and three books.

JIAN XIONG received the B.Sc. degree in computer science and technology from the Anhui University of Finance and Economics, Bengbu, China, in 2007, the M.Sc. degree in computer application technology from Xihua University, Chengdu, China, in 2010, and the Ph.D. degree in single and information processing from the University of Electronic Science and Technology of China, Chengdu, in 2015. In 2014, he was a Research Assistant with the Image and Video Processing Laboratory, The Chinese University of Hong Kong, Hong Kong. Since 2015, he has been a Lecturer with the College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China. His current research interests include image and video coding, pattern recognition, and machine learning.

GUAN GUI (SM'17) received the Dr.Eng. degree in information and communication engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2012.

From 2009 to 2012, he was financially supported by the China Scholarship Council and the Global Center of Education, Tohoku University, where he joined the Wireless Signal Processing and Network Laboratory (Prof. Adachi's Laboratory), Department of Communications Engineering, Graduate School of Engineering, as a Research Assistant and a Postdoctoral Research Fellow. From 2012 to 2014, he was supported by the Japan Society for the Promotion of Science Fellowship as a Postdoctoral Research Fellow of the Wireless Signal Processing and Network Laboratory. From 2014 to 2015, he was an Assistant Professor with Department of Electronics and Information System, Akita Prefectural University. Since 2015, he has been a Professor with the Nanjing University of Posts and Telecommunications, Nanjing, China. He is currently engaged in the research of compressive sensing and deep learning for physical layer signal processing. He has received four best paper awards in international conferences, including VTC 2014 (Spring), ICC 2014, ICNC 2018, and ICC 2017. He was also selected as a Jiangsu Specially Appointed Professor and Jiangsu High-level Innovation and Entrepreneurial Talent. He has received the Nanjing Youth Award and the 1311 Talent Plan of NJUPT. He has been an Editor of *Security and Communication Networks* (2012–2016), an Editor of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY (2017), and *KSII Transactions on Internet and Information Systems* (2017).