IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 3, 2019, accepted October 16, 2019, date of publication October 28, 2019, date of current version November 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2949741

A Face Emotion Recognition Method Using Convolutional Neural Network and Image Edge Computing

HONGLI ZHANG[1], ALIREZA JOLFAEI[2], AND MAMOUN ALAZAB[3]

[1]Department of Educational Technology, Inner Mongolia Normal University, Hohhot 010022, China
[2]Department of Computing, Macquarie University, Sydney, NSW 2109, Australia
[3]Charles Darwin University, Darwin, NT 0810, Australia

Corresponding author: Hongli Zhang (zhanghl_1974@163.com)

This work was supported in part by the Natural Science Foundation Project of Inner Mongolia Autonomous Region under Grant 2015MS0634, and in part by the Scientific and Technological Projects of Institutions of Higher Learning in Inner Mongolia Autonomous Region under Grant NJZY033.

ABSTRACT To avoid the complex process of explicit feature extraction in traditional facial expression recognition, a face expression recognition method based on a convolutional neural network (CNN) and an image edge detection is proposed. Firstly, the facial expression image is normalized, and the edge of each layer of the image is extracted in the convolution process. The extracted edge information is superimposed on each feature image to preserve the edge structure information of the texture image. Then, the dimensionality reduction of the extracted implicit features is processed by the maximum pooling method. Finally, the expression of the test sample image is classified and recognized by using a Softmax classifier. To verify the robustness of this method for facial expression recognition under a complex background, a simulation experiment is designed by scientifically mixing the Fer-2013 facial expression database with the LFW data set. The experimental results show that the proposed algorithm can achieve an average recognition rate of 88.56% with fewer iterations, and the training speed on the training set is about 1.5 times faster than that on the contrast algorithm.

INDEX TERMS Face expression recognition, convolutional neural network, edge computing, deep learning, image edge detection.

I. INTRODUCTION

Human-computer interaction technology refers to a kind of technology which takes computer equipment as the medium, so as to realize the interaction between human and computer. In the recent years, with the rapid development of pattern recognition and artificial intelligence, more and more research has been conducted in the field of human-computer interaction technology [1], [2]. The facial expression recognition, as an important means of intelligent human-computer interaction, has a broad application background. It has been applied in the fields of assistant medicine, distance education, interactive games and public security [3]–[5]. The facial expression recognition extracts the information representing the facial expression features from the original input facial expression images through computer image processing technology, and classifies the facial expression features according to human emotional expression, such as happiness, surprise, aversion and neutrality [6], [7]. The facial expression recognition plays an important role in the research of emotional quantification. Under the trend of artificial intelligence, the communication between human and computer becomes easier and easier. Therefore, vigorously promoting the research of facial expression recognition technology is of great value to the development of individuals and society [8], [9]. The facial expression recognition is a technology which uses computer as an assistant tool and combines it with specific algorithms to judge the inner emotion of the human face expression. The facial expression recognition is also applied to the medical field. To know the effect of new antidepressants, more accurate drug evaluation can be made according to the daily record of patients' facial expressions. In the treatment of autistic children, facial expression recognition can be used to help interpret the emotions of

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

autistic children and help doctors understand themselves. Psychological changes in autistic children, so as to develop more accurate treatment programs [10]. The application of facial expression recognition in teaching field can enable the teaching system to capture and record students' emotional changes in learning, and provide better reference for teachers to teach students in accordance with their aptitude. The application of facial expression recognition in traffic field can be used to judge the fatigue state of pilots or drivers, and to avoid the occurrence of traffic hazards by technical means. Applying facial expression recognition to daily life, life management robots can understand people's mental state and intention according to facial expression recognition, and then make appropriate responses, thus enhancing the experience of human-computer interaction.

In the recent years, the development of facial expression recognition technologies has been rapid and many scholars have contributed to the development of facial expression recognition [11], [12]. Among them, the Massachusetts Institute of Technology Media Laboratory and Japan's Art Media Information Science Laboratory are representative. The research of expression recognition in computer field mainly focuses on the feature extraction and feature classification. The so-called feature extraction refers to extracting features that can be used for classification from input pictures or video streams [13], [14]. There are many methods of feature extraction. According to the type of data input, the existing methods of feature extraction can be divided into two categories: one is based on static images and the other is based on a dynamic sequence. Feature extraction methods based on static images include Gabor wavelet transform [15], Haar wavelet transform [16], Local Binary Pattern (LBP), and Active Appearance Models (AAM) [17]. Generally speaking, the dimension of feature is large before and after the completion of feature, and thus the dimension reduction is usually carried out [18]. The facial expression classification refers to the use of specific algorithms to identify the categories of facial expressions according to the extracted features. Commonly used methods of facial expression classification are Hidden Markov Model (HMM), Support Vector Machine (SVM), AdaBoost, and Artificial Neural Networks (ANN) [6]. To avoid the complex process of explicit feature extraction and low-level data manipulation in traditional facial expression recognition, a fast R-CNN (Faster Regions with Convolutional Neural Network Features) facial expression recognition method is proposed in the literature [19]. The trainable convolution kernel is used to extract the implicit features, and the maximum pool is used to reduce the dimension of the extracted implicit features. The work in [20] presents a Feature Redundancy-Reduced Convolutional Neural Network (FRR-CNN). Unlike traditional CNN, the convolution core of FRR-CNN diverges due to the more discriminant differences between feature maps at the same level, resulting in fewer redundant features and a more compact image representation. In addition, the transformation invariant pool strategy is used to extract representative cross-transform features. The work in [21] presents a hierarchical Bayesian topic model based on pose to solve the challenging problem in multi-user facial expression recognition. The model combines local appearance features with global geometric information and learns intermediate representation before recognizing expression. By sharing a set of functions with different postures, it provides a unified solution for multi-functional facial expression recognition, bypassing the individual training and parameter adjustment of each posture, so it can be extended to a large number of postures.

Although the CNN algorithm has made some progress in the field of facial expression recognition, it still has some shortcomings, such as too long training time and low recognition rate in the complex background. To avoid the complex process of explicit feature extraction in traditional facial expression recognition, a facial expression recognition method based on CNN and image edge detection is proposed in this paper. The main innovations of this method are as follows:

(1) The edge of each layer of the input image is extracted, and then the extracted edge information is superimposed on each feature image to preserve the edge structure information of the texture image.

(2) In this paper, the maximum pooling method is used to reduce the dimension of the extracted implicit features, which shortens the training time of the convolutional neural network model.

(3) The Fer-2013 facial expression database and LFW (Labeled Faces in the wild) data set are scientifically mixed to design a simulation experiment, which proves that the method proposed in this paper has a certain robustness for facial expression recognition under a complex background.

II. FACIAL EXPRESSION DATA PREPROCESSING

Because the original pictures of facial expressions have complex background, different sizes, different shades and other factors, a series of image pre-processing processes have to be completed before facial expressions are input into the network for training. Firstly, we locate the face in the image and cut out the face image. Then, we normalize the face image to a specific size. Next, we equalize the histogram of the image to reduce the influence of illumination and other factors. Finally, we extract the edge of each layer of the image in the convolution process. The extracted edge information is superimposed on each feature image to preserve the edge structure information of texture image.

A. FACE DETECTION AND LOCATION

This paper uses a Haar classifier for human detection. The Haar classifier is trained by Haar-like small features and an integral graph method combined with the AdaBoost algorithm. The Haar-like is a commonly used texture descriptor, and its main features are linear, edge, center and diagonal. Adaboost is an improvement of Boosting algorithm and its core idea is to form a strong classifier by iterating not only

weak classifiers but also weak classifiers. The Viola-Jones detector is a milestone in the history of face detection. It has been widely used because of its high efficiency and fast detection. This method uses the Haar-like to extract facial features, and uses an integral graph to realize fast calculation of Haar-like features, and screens out important features from a large number of Haar-like features. Then, we use the Adaboost algorithm to train and integrate the weak classifier into a strong classifier. Finally, several strong classifiers are cascaded in series to improve the accuracy of the face detection.

The Haar-like feature can reflect the gray level change of image, so it is very effective to describe human face, because many features of human face have obvious contrast change characteristics. However, the calculation of eigenvalues is very time-consuming. In order to improve the calculation speed, this paper uses the integral graph method to calculate the Haar like eigenvalues. The concept of an integral graph is expressed in Figure 1 (a). The integral graph of the coordinate $A\,(x, y)$ in a graph is defined as the sum of all the pixels in its upper left corner.

$$A\,(x, y)\, ii\,(x, y) = \sum_{x' \le x, y'} i\left(x', y'\right) \tag{1}$$

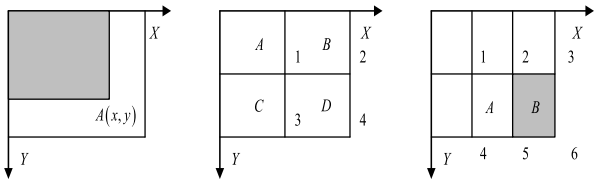


FIGURE 1. Integral graph method to calculate eigenvalues.

Here, $ii\,(x, y)$ represents the integral image. $i\left(x', y'\right)$ represents the original image; for gray image, here represents the gray value and for color image, here represents the color value.

The pixel value of an area can be calculated by using the integral graph of the end points of the area, as shown in Figure 1 (b). The pixel value of region D can be calculated by

$$S\,(D) = ii\,(4) + ii\,(1) - ii\,(2) - ii\,(3) \tag{2}$$

where $ii\,(1)$ represents the pixel value of region A, $ii\,(2)$ represents the pixel value of region $A + B$, $ii\,(3)$ represents the pixel value of region $A + C$, $ii\,(4)$ represents the pixel value of regions $A + B + C + D$. The eigenvalues of rectangular features can be calculated by integral graphs of feature endpoints. Taking the edge feature a as an example, the eigenvalue calculation can be expressed by Fig. 1 (c). The pixel values of point A and point B are:

$$S\,(A) = ii\,(5) + ii\,(1) - ii\,(2) - ii\,(4) \tag{3}$$

$$S\,(B) = ii\,(6) + ii\,(2) - ii\,(5) - ii\,(3) \tag{4}$$

According to the definition, the eigenvalue of rectangular feature is the pixel value of region A minus the pixel value of region B. According to formula (3) and formula (4), the formula for calculating eigenvalue is as follows.

$$T = ii(5) - ii(4) + ii(3) - ii(2) - (ii(2) - ii(1)) - (ii(6) - ii(5)) \tag{5}$$

It can be seen that the eigenvalues of rectangular features are only related to the integral graph of rectangular endpoints. Through simple integral graph addition and subtraction operation, the eigenvalues can be calculated, which greatly improves the speed of target detection. Next, the extracted Haar-like features are used to train the classifier, and the AdaBoost algorithm is used to train the classifier. Finally, the trained classifier is used to extract the face from the image.

B. SCALE NORMALIZATION

Because the input of the network is a fixed sized picture, before the picture is input into the network, the original picture should be normalized to generate a specific size picture. Let point (x, y) in the original picture be normalized and mapped to point $\left(x', y'\right)$. The mapping is as follows:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{6}$$

where s_x represents the scaling ratio of the image in the direction of x axis and s_y represents the scaling ratio of the image in the direction of y axis. In the process of image scaling, bilinear interpolation algorithm is also needed to fill the image. A, B, C and D are the four points around the pixel (x, y). The corresponding gray values are $g\,(A)$, $g\,(B)$, $g\,(C)$, $g\,(D)$. To get the gray value of point (x, y) and calculate the gray value of points E and F, the formula is as follows:

$$g\,(E) = (x - x_D)\,(g\,(C) - g\,(D)) + g\,(D) \tag{7}$$

$$g\,(F) = (x - x_A)\,(g\,(B) - g\,(A)) + g\,(A) \tag{8}$$

x_A and x_D are the abscissa of point A and point D, respectively. The gray scale formula of (x, y) is as follows:

$$g\,(x, y) = (y - y_D)\,(g\,(F) - g\,(E)) + g\,(E) \tag{9}$$

where y_D represents the ordinates of CD points. Through normalization, the input image is scaled to 128×128 size. As shown in Figure 2, it is a normalized contrast map.

(a)Before normalization (b)After normalization

FIGURE 2. Contrast before and after normalization.

C. GRAY LEVEL EQUALIZATION

In the actual image acquisition process, it is easy to be affected by illumination, shadows and other factors, which makes the collected image show a state of uneven distribution of light and shade, which will increase the difficulty of feature extraction. Therefore, it is necessary to average the gray level of the image to enhance the contrast of the image. In this paper, the Histogram Equalization (HE) method is used to process images. The basic idea is to transform the histogram of the original graph into a uniform distribution form [22]. If the gray level of the gray image is L, the size is $M \times N$, and the number of pixels in the r_i gray level is E, the corresponding probability of gray level occurrence is as follows:

$$P_r(r_i) = \frac{n_i}{M \times N}, \quad i = 0, 1, \cdots, L-1 \tag{10}$$

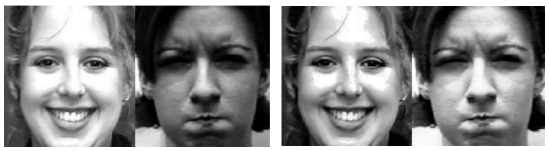
Subsequently, the cumulative distribution function is calculated using the following equation.

$$T(r_i) = \sum_{j=0}^{i} P_r(r_j), \quad i = 0, 1, \cdots, L-1 \tag{11}$$

Finally, the image histogram is averaged using the following mapping relations:

$$e_j = INT\left[(e_{\max} - e_{\min}) \underset{j=0,1,\cdots,L-1}{T(r)} + e_{\min} + 0.5\right] \tag{12}$$

The processing results are shown in Figure 3. When the histogram of the image is completely uniform, the entropy of the image is the largest and the contrast of the image is the largest. In fact, gray level equalization realizes the uniform distribution of image histogram, which enhances the contrast of the image and makes the details clearer, and is conducive to the extraction of facial features.



(a)Before gray level equalization (b)After gray level equalization

FIGURE 3. Grayscale equalization before and after contrast.

D. IMAGE EDGE DETECTION

The edge information of an image is often reflected in the area where the gradient information of the image changes dramatically. The edge of the image gives people a stronger visual sense. Therefore, the edge information of the image cannot be ignored in the process of texture synthesis. Some edge information of the image is lost, which results in the blurred edge information in the final synthesis result and affects the table. In this paper, we extract the edge of each layer of the image in the convolution process, and then superimpose the extracted edge information on each feature map, which preserves the edge structure information of texture image. Kirsch edge operator is used to extract image edge information. The template of eight directions of Kirsch operator is respectively.

$$a_0 = \begin{bmatrix} 5 & 5 & 5 \\ -3 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix}, \quad a_1 = \begin{bmatrix} -3 & 5 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & -3 \end{bmatrix},$$
$$a_2 = \begin{bmatrix} -3 & -3 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & 5 \end{bmatrix}, \quad a_3 = \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & 5 \\ -3 & 5 & 5 \end{bmatrix}$$
$$a_4 = \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & -3 \\ 5 & 5 & 5 \end{bmatrix}, \quad a_5 = \begin{bmatrix} -3 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & 5 & -3 \end{bmatrix},$$
$$a_6 = \begin{bmatrix} 5 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & -3 & -3 \end{bmatrix}, \quad a_7 = \begin{bmatrix} 5 & 5 & -3 \\ 5 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix} \tag{13}$$

Assuming that any pixel P_A in the image is surrounded by the gray level of 3×3 area, and that g_i $(i = 0, 1, \cdots, 7)$ is the gray level of point A obtained by convolution of the $i+1$ template of the Kirsch edge operator of the image, the gray level of point A can be obtained by the convolution of the D template of the Kirsch edge operator

$$g_0 = 5 \times (a_3 + a_4 + a_5) - 3(a_2 + a_6) - 3(a_1 + a_0 + a_7) \tag{14}$$

In Equation (14), a_i $(i = 0, 1, \cdots, 7)$ is the neighborhood pixel of the arbitrary point A. The gray value of point A in other directions can be calculated by the same method of Equation (14). After processing, the gray value of point A is calculated by

$$g_A = \max(g_i) \quad i = 0, 1, \cdots, 7 \tag{15}$$

III. FACE EXPRESSION RECOGNITION NETWORK MODEL BASED ON CNN

The essence of deep learning method is to construct a deep neural network similar to human brain structure, which learns more advanced feature expression of data layer by layer through multi-hidden non-linear structure. This mechanism of automatically learning the internal rules of large data makes the extracted features have more essential characterization of the data, and thus the classification results can be greatly enhanced. For a two-dimensional image input, the neural network model can interpret it layer-by-layer from the pixels initially understood by the computer to edges, parts, contours of objects, objects understood by the human brain, and then can classify it directly within the model to obtain recognition results.

The CNN is a feedforward neural network, which can extract features from a two-dimensional image and optimize network parameters by using back propagation algorithm. Common CNNs usually consist of three basic layers: a convolution layer, a pooling layer and a connective layer. Each layer is composed of several two-dimensional planes, that is, feature maps, and each feature map has many neurons. In convolution neural network, the input layer is a two-dimensional

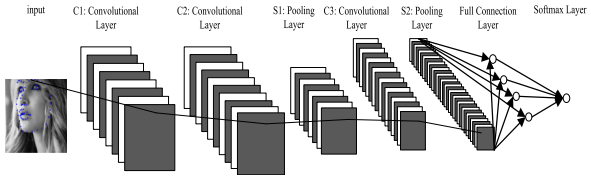


FIGURE 4. CNN structure for facial expression recognition.

matrix composed of image pixels. The alternation of convolution layer C and pooling layer S is the core module to realize feature extraction of convolution neural network. This paper designs a CNN structure for facial expression recognition, as shown in Figure 4. Excluding input layer, the network consists of seven layers, including three convolution layers ($C1$, $C2$ and $C3$), two pooling layers ($S1$ and $S2$), one full connection layer and one Softmax layer. The input layer is a 96×96 face pixel matrix. Each feature map is connected locally with its previous feature map. There are several feature maps in the convolution layer and the pooling layer.

The CNN has three important characteristics: local perception, weight sharing and down sampling. These characteristics overcome the problem of too many parameters and difficult calculation of the traditional feedforward neural network in high-dimensional input, and make the model obtain certain translation, rotation and distortion invariance. Common sense holds that people's perception of the outside world is generally from the local to the whole. The image has a certain spatial connection. The adjacent pixels are closely related, while the distant pixels have little correlation. Therefore, neurons only need to perceive the local pixels, and then integrate the local information at the bottom to get the global information at the high level. This is the concept of local perception, which greatly reduces the number of parameters needed to learn.

Convolutional layers $C1$, $C2$ and $C3$ use 32, 64 and 128 convolutional nuclei for convolution operation respectively. The size of convolutional nuclei used in each convolution layer is 5×5; the size of sampling window used in pooling layer $S1$ and $S2$ is 2×2; the full connection layer contains 300 neurons, which are fully connected with pooling layer S2; and the Softmax layer contains 7 neurons, which are fully connected. The features of the output layer were classified, and the facial expressions were divided into seven categories: fearful, angry, sad, happy, surprised, disgusted and neutral.

A. CONVOLUTION LAYER

The convolution layer is the core of CNN, which has the characteristics of local connection and value sharing. The specific method is to use the convolution core to calculate on the upper input layer by sliding windows one by one [23], [24]. The input image and several trainable convolution filters are convoluted to produce the $C1$ layer of the feature mapping layer. Then the feature mapping map is processed, including summation, weighting and bias operations. A new feature mapping layer $S2$ is obtained through an activation function. Then the convolution operation is carried out on the $S2$ layer, and the $C3$ layer is obtained. According to the same principle and operation, the $S4$ layer is obtained. Finally, these feature maps are rasterized and connected into a set of feature vectors, which are then transferred to the neural network classifier for training. Generally, the computational expression of convolution layer is

$$y_j^l = \theta\left(\sum_{i=1}^{N_j^{l-1}} w_{i,j} \otimes x_i^{l-1} + b_j^l\right), \quad j = 1, 2, \cdots, M \tag{16}$$

The layer l is the current layer, and the layer $l-1$ is the previous layer. $w_{i,j}$ represents the convolution kernel of the jth feature graph of the current layer and the i feature graph of the previous layer; y_j^l represents the j-th feature graph of the current layer; b_j^l represents the bias of the jth feature graph of the current layer; x_i^{l-1} represents the ith feature graph of the previous layer. In the experiment, $b_j^l = 0$ is adopted to enable the network to train rapidly and reduce learning parameters. M represents the number of feature maps of the current layer; $\theta(\cdot)$ stands for activation function; N_j^{l-1} means connected to the current layer 1 j a characteristic figure of the previous layer all the characteristics of the figure, the number of experiments, the modified Linear unit (Rectified Linear Units, ReLU) functions, rather than the commonly used sigmoid or hyperbolic tangent function $\tanh(\cdot)$, because ReLU can produce a more sparse ReLU function.

$$\theta(x) = \max(0, x) \tag{17}$$

Practice has proved that the network trained with ReLU activation function has moderate sparsity. At the same time, it can solve the problem of gradient disappearance which may occur in the process of adjusting back propagation parameters and accelerate the convergence of the network. The feature extracted by convolution operation can be directly used to train classifiers, but it still faces huge computational challenges. In order to further reduce the parameters, a down sampling operation is proposed after the convolution operation. The basis of down sampling is that the pixels in the continuous range of the image have the same characteristics (local correlation), so the features of different locations can be aggregated and counted. For example, we can calculate the average or maximum value of a specific feature in an image region. This statistical dimensionality reduction method not only reduces the number of parameters, prevents fitting, but also makes the model obtain the scaling invariance of the image.

The convolution layer $C1$ convolutes 96×96 pixels of input image with a 5×5 convolution core, i.e. each neuron specifies a 5×5 local receptive field, so the size of the feature map obtained by convolution operation is $(96 - 5 + 1) \times (96 - 5 + 1) = 92 \times 92$. Through convolution operations of 32 different convolution kernels, 32 feature maps are obtained, that is, 32 different local expression features are extracted. Convolution layer $C2$ uses 64 5×5 convolution kernels and

then convolutes the characteristic graphs of convolution layer $C1$ output. 64 feature graphs are obtained. The size of each feature graph is $(92-5+1)\times(92-5+1)=88\times 88$. In convolution layer $C3$, 128 5×5 convolution kernels are used to convolute the characteristic maps of pool layer $S1$ output, and 128 feature maps are obtained. The size of each feature map is $(44-5+1)\times(44-5+1)=40\times 40$.

The principle of weight sharing is that the statistical characteristics of one part of an image are similar to those of other parts, so the same convolution kernel can be used to extract features for all positions on the image. However, it is not enough to use only one convolution kernel to learn the features. Therefore, in the actual training of convolution neural network, many convolution kernels are used to increase the diversity of feature mapping. Each kind of convolution can get the mapping plane of different features of the image. By using weight sharing, not only abundant image information can be obtained, but also the number of parameters needed for network training can be greatly reduced. Under the condition of reasonable control of network structure, the generalization ability of convolutional neural network can be enhanced. The feature extracted by convolution operation can be directly used to train classifiers, but it still faces huge computational challenges. In order to further reduce the parameters, a down sampling operation is proposed after the convolution operation. The basis of down sampling is that the pixels in the continuous range of the image have the same characteristics (local correlation), so the features of different locations can be aggregated and counted. For example, we can calculate the average or maximum value of a specific feature in an image region. This statistical dimensionality reduction method not only reduces the number of parameters, prevents fitting, but also makes the model obtain the scaling invariance of the image.

B. POOLING LAYER

The main purpose of the pooling operation is to reduce the dimension. A pooling window of 2×2 step size can reduce the dimension of the next feature map by half. Although there is no direct reduction in the number of training parameters, halving the dimension of feature graph means that the computational complexity of convolution operation will be greatly reduced, which greatly improves the training speed.

If we train the Softmax classifier directly with all the features we have learned, it will inevitably bring about the problem of dimension disaster. To avoid this problem, a pooling layer is usually used after the convolution layer to reduce the feature dimension [25], [26]. Down sampling does not change the number of feature maps, but reduces the output of feature maps, which reduces the sensitivity to translation, scaling, rotation and other transformations. If the size of the sampling window is $n\times n$, then after one down-sampling, the size of the feature graph becomes $1/n\times 1/n$ of the original feature graph. The general expression of pooling is

$$y_j^l=\theta\left(\beta_j^l down\left(y_j^{l-1}\right)+b_j^l\right) \tag{18}$$

Among them, y_j^l and y_j^{l-1} represent the j-th feature map of the current layer and the first layer respectively; $down(\cdot)$ represents a down sampling function; β_j^l and b_j^l represent the multiplicative and additive biases of the j-th feature map of the current layer, respectively. In the experiment, $\beta_j^l=1$, $b_j^l=0$ and $\theta(\cdot)$ are used as activation functions, and identical functions are used in the experiment.

After sharing the local receptive fields and weights, the number of training parameters is greatly reduced, but the dimension of the feature map is not much reduced. There are two problems. Firstly, if the dimension of feature graph is too large, the number of training parameters generated by full connection will be very large; secondly, the computer will waste a lot of time on convolution calculation in the process of training network.

C. FULL CONNECTION LAYER

The input of the full connection layer must be a one-dimensional array, whereas the output of the previous pooling layer $S2$ is a two-dimensional array. First, the two-dimensional array corresponding to each feature graph is converted into a one-dimensional array, and then 128 one-dimensional arrays are connected in series to a feature vector of 51200 dimensions ($20\times 20\times 128=51200$) as the full connection. The output of each neuron is

$$h_{w,b}(x)=\theta\left(w^T x+b\right) \tag{19}$$

where $h_{w,b}(x)$ denotes the output value of neurons. x denotes the input eigenvector of neurons. w denotes the weight vector. b denotes bias. $b=0$ denotes the activation function in experiments. $\theta(\cdot)$ denotes the activation function, and ReLU function is used in experiments. The number of neurons will affect the training speed and fitting ability of the network. The experimental results show that when the number of neurons is 300, the effect is better.

D. SOFTMAX LAYER

The last layer of the CNN uses a Softmax classifier. The Softmax classifier is a multi-output competitive classifier. When a given sample is input, each neuron outputs a value between 0 and 1, which represents the probability that the input sample belongs to that class. Therefore, the category corresponding to the neuron with the largest output value is selected as the classification result.

E. CNN PARAMETER TRAINING

The training process of CNN is essentially the process of optimizing and updating network weights. Appropriate initialization of weights has a great impact on the updating of weights. The commonly used initialization methods include constant initialization, uniform distribution initialization and Gauss distribution initialization. The CNN essentially implements a mapping relationship between input and output. The CNN carries out supervised training. Before starting training, it initializes the ownership value of the network with some

different small random numbers. The training of convolution neural network is divided into two stages:

1) Forward propagation stage. Sample x is extracted from the training sample set. Its corresponding category label is y, $\tilde{y}$ is a 7-dimensional vector whose elements represent the probability that x is divided into different categories. x is input to the CNN network. The output of the upper layer is the input of the current layer. Then, the output of the current layer is calculated by activation function, which is passed down layer by layer. Finally, the output $\tilde{y}$ of the Softmax layer is obtained.

(2) Back propagation stage, also known as error propagation stage. Calculate the error between the output $\tilde{y}$ of Softmax layer and the class label vector y of a given sample (y is a 7-dimensional vector, only the element corresponding to the class label y is 1, the other elements are 0), and adjust the weight parameters by minimizing the mean square error cost function.

IV. EXPERIMENT

In this section, two sets of experiments are designed to verify the performance of the proposed method. The first group of experiments analyze the performance of the algorithm and verify that the training time of the algorithm is lower than that of the traditional CNN algorithm model. The experimental data comes from the Fer-2013 expression database. The second group of experiments is used to verify that the recognition rate of the algorithm has increased under complex background. The experimental data comes from the Fer-2013 facial expression database and LF mixed sets of W data sets [27], [28]. The Fer-2013 facial expression database contains 28,709 training pictures and 7,178 test pictures, each of which is a 48×48 gray scale image. Each face is more or less in the middle of the picture. Therefore, in the experiment, the image data can be directly input into the network for training without any other pre-processing.

We use a Keras framework to build the network. Keras is a python-based neural network framework, which supports seamless switching between theano and tensorflow [29]. The hardware platform of the experiment is Intel (R) Core (TM) i5-6500 CPU main frequency 3.2GHz, 16GB memory and 6GB NVIDIA GeForce GTX 1060 GPU display memory.

A. PERFORMANCE ANALYSIS EXPERIMENTS

The purpose of this experiment is to test the performance of the proposed algorithm, and verify that the proposed algorithm has a lower training time than the original algorithm. The experimental data come from the images of Fer-2013 expression database. The Fer-2013 expression database has been introduced in Section 3.D. In order to improve the reliability of the experimental results, three cross-validation experiments were carried out, which divided 35,886 facial expression images into three parts on average. Two of them were used as training samples in each experiment, and the remaining one was used as test samples. The experiments were repeated three times, and the average recognition results of three times were taken as the final recognition performance. The training set and the test set contained seven expressions of happiness, fear and surprise, respectively.

In order to verify the performance of the proposed algorithm, the proposed method and R-CNN model [19] are compared experimentally. In the experiment, the same experimental environment and experimental data were used. Through simulation experiments, the relationship between iterations and Accuracy of training sets of the two models is obtained, as shown in Figure 5.

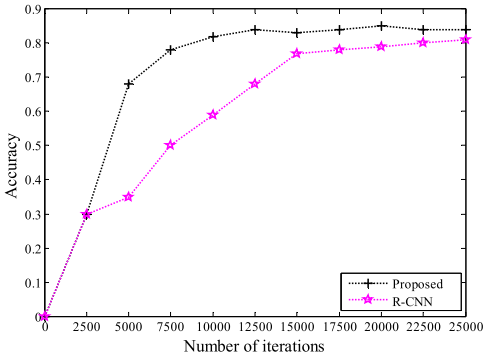


FIGURE 5. Expression recognition rate under different iterations.

From Figure 5, we can see that both the proposed method and R-CNN algorithm converge. It needs to be explained that this experiment takes the training set as an example. There are 23,924 images in the training set, 48 facial expressions (batch_size $=$ 48) are processed at a time, and 499 samples can be processed at a time. This experiment has trained 50 generations in the training set, that is, 25,000 iterations. From the above figure, the following conclusions can be drawn:

(1) As can be seen from Figure 5, both the proposed algorithm and the R-CNN algorithm converge after a certain number of iterations. When the model converges, the recognition rates are 85.54% and 77.78%, respectively. It can be seen that the proposed algorithm improves by nearly 8 percentage points compared to the R-CNN algorithm. Thus, the proposed algorithm has certain advantages in facial expression recognition rate.

(2) As seen from Figure 5, both models converge after a certain number of iterations. When the proposed model is iterated to 10,000 times, the model begins to converge. The R-CNN algorithm converges after 15,000 iterations. This shows that the proposed algorithm can achieve satisfactory results after fewer iterations, that is to say, the training speed of the proposed method on the training set is 1.5 times faster than that of R-CNN algorithm.

The proposed method is compared with R-CNN and FRR-CNN algorithms [20]. The experimental data come from Fer-2013 facial expression data set. Table 1 lists the recognition rate comparison of the three algorithms, and the time comparison on the test set and the training set.

The training time and test time of training set indicated in Table 1 refer to the time used to process a batch of images. In this paper, 48 images are processed in batches.

TABLE 1. Performance comparison of three algorithms.

METHOD	TRAINING TIME(S)	TEST TIME(S)	RECOGNITION RATE(%)
THE PROPOSED ALGORITHM	178	24.89	88.56
R-CNN ALGORITHM	256	33.97	79.34
FRR-CNN ALGORITHM	148	17.92	70.63

Table 1 shows that the proposed algorithm is smaller than R-CNN algorithm and FRR-CNN algorithm in both training time and testing time. It can be concluded that the maximum pooling method is used to reduce the dimension of the extracted implicit features, which can shorten the training time of the convolutional neural network model. Moreover, the proposed algorithm achieves the highest average recognition rate of 88.56%.

B. COMPARATIVE ANALYSIS EXPERIMENT

The purpose of this experiment is to verify the robustness of the proposed algorithm for facial expression recognition under complex background. The experimental data in the previous experiment are from the Fer-2013 facial expression database. Because the facial expression database is a standard facial expression image set, there is no complex recognition background. Therefore, in order to verify the recognition effect of this experiment under complex background, the experimental data is a mixture of LFW data set and Fer-2013 facial expression database. The LFW data set contains 13,000 face images, each of which is named after the person being photographed. The expression images of the database are not collected in the laboratory, but in the complex external environment collected by the network. The purpose of this experiment is to verify the recognition performance of the proposed algorithm under complex background.

In order to verify the robustness of this method for facial expression recognition under complex background, the proposed algorithm is compared with the classical convolutional neural network FRR-CNN model and R-CNN algorithm. This experiment can be divided into two steps. The first step is to scientifically mix the Fer-2013 facial expression database and LFW data set, and take 3,269 images of each of the seven kinds of expressions in the Fer-2013 facial expression database, totaling 22,883 images. It combines with 13,000 images of LFW data set, and forms a data set containing 35,883 images for training. The second step is to use 28,341 pictures in the mixed data set as training set and 7,542 pictures as test set to get the expression recognition rate of the above algorithm under different iteration times in different test sets. The experimental results are shown in Figure 6.

As can be seen from Figure 6, the three models converge after iteration to a certain algebra. Taking the test set as an example, the following conclusions can be drawn from the analysis of Figure 6 in our laboratory:

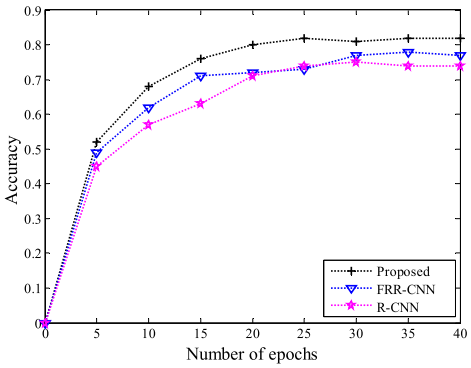


FIGURE 6. Expression recognition rate of different methods in complex background.

(1) From the figure it can be seen that the proposed algorithm began to converge after 20 generations of training, the R-CNN algorithm began to converge after about 26 generations, and the convergence speed of the FRR-CNN model was relatively slow. For R-CNN model and FRR-CNN model, the number of layers is similar, but R-CNN model has stronger feature extraction ability than FRR-CNN model, which makes the convergence speed of R-CNN model in complex background faster to some extent.

(2) It can be seen from the figure that in the experimental environment of complex background, although the overall recognition rate is not high in the Fer-2013 data set, the recognition rate of the proposed algorithm is still higher than that of the other two methods after iteration to a certain extent. Therefore, it can be explained that the proposed algorithm can improve the recognition rate of facial expressions in complex background to a certain extent.

V. CONCLUSION

In this paper, we propose a facial expression recognition method using a CNN model which extracts facial features effectively. Compared to traditional methods, the proposed method can automatically learn pattern features and reduce the incompleteness caused by artificial design features. The proposed method directly inputs the image pixel value through training sample image data. Autonomous learning can implicitly acquire more abstract feature expression of the image. The training process of the proposed method uses appropriate initialization of weights which has a great impact on the updating of weights. Our extensive experimental analysis shows that compared to the past literatures, the proposed algorithm can improve the recognition rate of facial expressions in complex background to a certain extent. Compared to FRR-CNN and R-CNN models, the convergence speed of proposed model is much faster in complex background environments. Also, the proposed method achieves a higher recognition rate.

Facial expressions captured in reality may have various noises, such as face posture, occlusion, and blurring. To address this concern, as a future work, we will investigate

more robust models which satisfy real conditions. We will also focus on how to reduce the complexity of network structure, and will try to recognize dynamic expressions with 3D convolution technology.

REFERENCES

[1] R. M. Mehmood, R. Du, and H. J. Lee, "Optimal feature selection and deep learning ensembles method for emotion recognition from human brain EEG sensors," *IEEE Access*, vol. 5, pp. 14797–14806, 2017.

[2] T. Song, W. Zheng, C. Lu, Y. Zong, X. Zhang, and Z. Cui, "MPED: A multi-modal physiological emotion database for discrete emotion recognition," *IEEE Access*, vol. 7, pp. 12177–12191, 2019.

[3] E. Batbaatar, M. Li, and K. H. Ryu, "Semantic-emotion neural network for emotion recognition from text," *IEEE Access*, vol. 7, pp. 111866–111878, 2019.

[4] Y. Zhang, L. Yan, B. Xie, X. Li, and J. Zhu, "Pupil localization algorithm combining convex area voting and model constraint," *Pattern Recognit. Image Anal.*, vol. 27, no. 4, pp. 846–854, 2017.

[5] H. Meng, N. Bianchi-Berthouze, Y. Deng, J. Cheng, and J. P. Cosmas, "Time-delay neural network for continuous emotional dimension prediction from facial expression sequences," *IEEE Trans. Cybern.*, vol. 46, no. 4, pp. 916–929, Apr. 2016.

[6] X. U. Feng and J.-P. Zhang, "Facial microexpression recognition: A survey," *Acta Automatica Sinica*, vol. 43, no. 3, pp. 333–348, 2017.

[7] M. S. Özerdem and H. Polat, "Emotion recognition based on EEG features in movie clips with channel selection," *Brain Inf.*, vol. 4, no. 4, pp. 241–252, 2017.

[8] S. Escalera, X. Baró, I. Guyon, H. J. Escalante, G. Tzimiropoulos, M. Valstar, M. Pantic, J. Cohn, and T. Kanade, "Guest editorial: The computational face," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2541–2545, Nov. 2018.

[9] X. Yu, S. Zhang, Z. Yan, F. Yang, J. Huang, N. E. Dunbar, M. L. Jensen, J. K. Burgoon, and D. N. Metaxas, "Is interactional dissynchrony a clue to deception? Insights from automated analysis of nonverbal visual cues," *IEEE Trans. Cybern.*, vol. 45, no. 3, pp. 492–506, Mar. 2015.

[10] F. Vella, I. Infantino, and G. Scardino, "Person identification through entropy oriented mean shift clustering of human gaze patterns," *Multimedia Tools Appl.*, vol. 76, no. 2, pp. 2289–2313, Jan. 2017.

[11] S. H. Lee, K. N. K. Plataniotis, and Y. M. Ro, "Intra-class variation reduction using training expression images for sparse representation based facial expression recognition," *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 340–351, Jul./Sep. 2014.

[12] D. Ghimire, S. Jeong, J. Lee, and S. H. Park, "Facial expression recognition based on local region specific features and support vector machines," *Multimed. Tools Appl.*, vol. 76, no. 6, pp. 7803–7821, Mar. 2017.

[13] S. K. A. Kamarol, M. H. Jaward, H. Kälviäinen, J. Parkkinen, and R. Parthiban, "Joint facial expression recognition and intensity estimation based on weighted votes of image sequences," *Pattern Recognit. Lett.*, vol. 92, pp. 25–32, Jun. 2017.

[14] J. Cai, Q. Chang, X.-L. Tang, C. Xue, and C. Wei, "Facial expression recognition method based on sparse batch normalization CNN," in *Proc. 37th Chin. Control Conf. (CCC)*, Jul. 2018, pp. 9608–9613.

[15] B. Yang, X. Xiang, D. Xu, X. Wang, and X. Yang, "3D palmprint recognition using shape index representation and fragile bits," *Multimedia Tools Appl.*, vol. 76, no. 14, pp. 15357–15375, 2017.

[16] N. Kumar and D. Bhargava, "A scheme of features fusion for facial expression analysis: A facial action recognition," *J. Statist. Manage. Syst.*, vol. 20, no. 4, pp. 693–701, 2017.

[17] G. Tzimiropoulos and M. Pantic, "Fast algorithms for fitting active appearance models to unconstrained images," *Int. J. Comput. Vis.*, vol. 122, no. 1, pp. 17–33, 2017.

[18] M. Takalkar, M. Xu, Q. Wu, and Z. Chaczko, "A survey: Facial micro-expression recognition," *Multimedia Tools Appl.*, vol. 77, no. 15, pp. 19301–19325, 2018.

[19] J. Li, D. Zhang, J. Zhang, J. Zhang, T. Li, Y. Xia, Q. Yan, and L. Xun, "Facial expression recognition with faster R-CNN," *Procedia Comput. Sci.*, vol. 107, pp. 135–140, Jan. 2017.

[20] S. Xie and H. Hu, "Facial expression recognition with FRR-CNN," *Electron. Lett.*, vol. 53, no. 4, pp. 235–237, Feb. 2017.

[21] Q. Mao, Q. Rao, Y. Yu, and M. Dong, "Hierarchical Bayesian theme models for multipose facial expression recognition," *IEEE Trans. Multimedia*, vol. 19, no. 4, pp. 861–873, Apr. 2017.

[22] V. Magudeeswaran and J. F. Singh, "Contrast limited fuzzy adaptive histogram equalization for enhancement of brain images," *Int. J. Imag. Syst. Technol.*, vol. 27, no. 1, pp. 98–103, 2017.

[23] F. Zhang, Q. Mao, X. Shen, Y. Zhan, and M. Dong, "Spatially coherent feature learning for pose-invariant facial expression recognition," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 1s, Apr. 2018, Art. no. 27.

[24] H. Ma and T. Celik, "FER-Net: Facial expression recognition using densely connected convolutional network," *Electron. Lett.*, vol. 55, no. 4, pp. 184–186, Feb. 2019.

[25] L. Wei, C. Tsangouri, F. Abtahi, and Z. Zhu, "A recursive framework for expression recognition: From Web images to deep models to game dataset," *Mach. Vis. Appl.*, vol. 29, no. 3, pp. 489–502, 2018.

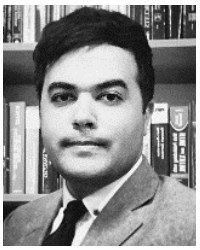
[26] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 356–370, Jan. 2019.

[27] A. V. Savchenko, "Deep neural networks and maximum likelihood search for approximate nearest neighbor in video-based image recognition," *Opt. Memory Neural Netw.*, vol. 26, no. 2, pp. 129–136, Apr. 2017.

[28] R. Massey *et al.*, "The behaviour of dark matter associated with four bright cluster galaxies in the 10 kpc core of Abell 3827," *Monthly Notices Roy. Astronomical Soc.*, vol. 449, no. 4, pp. 3393–3406, 2017.

[29] A. Moeini, K. Faez, H. Moeini, and A. M. Safai, "Facial expression recognition using dual dictionary learning," *J. Vis. Commun. Image Represent.*, vol. 45, pp. 20–33, May 2017.

HONGLI ZHANG graduated from the Beijing Institute of Technology, in 2014, and the Ph.D. degree in computer science. She is currently an Associate Professor with Inner Mongolia Normal University. She has authored more than 15 peer-reviewed articles on computer networks and intelligent algorithms. Her current research interests include artificial intelligent, data dining, and cognitive computing.



ALIREZA JOLFAEI received the Ph.D. degree in applied cryptography from Griffith University, Gold Coast, Australia. He is currently an Assistant Professor in cyber security with Macquarie University, Sydney, Australia. He has authored more than 50 peer-reviewed articles on topics related to cyber security. His current research interests include cyber security, the IoT security, human-in-the-loop CPS security, cryptography, AI, and machine learning for cyber security. He received the prestigious IEEE Australian Council Award for his research article published in the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY. He has served more than ten conferences in leadership capacities, including the Program Co-Chair, Track Chair, Session Chair, and Technical Program Committee Member, including IEEE TrustCom. He has served as a Guest Associate Editor for IEEE journals and transactions, including the IEEE INTERNET OF THINGS JOURNAL, the IEEE TRANSACTIONS ON INDUSTRIAL APPLICATIONS, and the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS.

MAMOUN ALAZAB received the Ph.D. degree in computer science from the School of Science, Information Technology and Engineering, Federation University Australia. He is currently an Associate Professor with the College of Engineering, IT and Environment, Charles Darwin University, Australia. He is also a Cyber-Security Researcher and Practitioner with industry and academic experience. His current research interests include cyber security and the digital forensics of computer systems, including current and emerging issues in the cyber environment like cyber-physical systems, and the Internet of Things, by taking into consideration the unique challenges present in these environments, with a focus on cybercrime detection and prevention. He has authored or coauthored more than 100 research articles, two of his papers were selected as the featured articles, and two other articles received the Best Paper Award. He was a recipient of the Short Fellowship from the Japan Society for the Promotion of Science (JSPS) based on his nomination from the Australian Academy of Science.