Facial expression recognition technology has been more and more in demand in security, entertainment, education, medical, and other domains as artificial intelligence has advanced, and face expression recognition technology based on deep learning has become one of the research hotspots. However, there are still some issues with the existing deep learning convolutional neural network; the feature extraction technique has to be improved, and the design of the detailed network structure needs to be optimized. It is critical to do more research on the deep learning convolutional neural network model in order to increase the accuracy of face facial expression detection. In this paper, a deep learning convolutional neural network structure combining VGG16 convolutional neural network and long and short-term memory networks is designed to address the shortcomings of existing deep learning methods in face expression recognition, which are prone to overfitting and gradient disappearance, resulting in low test accuracy. This structure easily and effectively collects facial expression information and then classifies the retrieved features using a support vector machine to detect face expressions. Finally, the fer2013 dataset is used to train face expression recognition, and the results demonstrate that the built deep convolutional neural network model can effectively increase face expression identification accuracy.

1. Introduction

There are many ways for people to communicate with each other: language, text, action, etc., and expression is also one of the important ways. With the rapid development of artificial intelligence, how to achieve the detection and accurate recognition of human face expressions has become a hot research topic in artificial intelligence nowadays [1]. Facial expression recognition technology has broad application scenarios in psychological guidance, entertainment, security, online education, and intelligent medical care [2]. Facial expression recognition can be applied to fatigue driving monitoring, using the camera to capture the real-time expression of the driver, analyze the driver’s mental state, and send out alerts when the driver’s face is tired, which can prevent traffic accidents caused by fatigue driving to a certain extent; in nursing homes or empty nest elderly homes, the installation of human-computer interaction system with facial expression recognition function can help the elderly with limited mobility and mental state. In the process of online teaching, it is difficult for the teacher to observe each student’s reaction as in offline teaching, so as to adjust the progress of the course in time, but the facial expression recognition system can monitor the students’ reaction to the lecture and give feedback to the teacher in time, which can help improve the teaching quality to a certain extent [3].

The traditional face expression recognition method is to capture the image and extract the image features first and then use the machine learning method to recognize the image, but this method has certain limitations, the feature extraction process is complicated, and the recognition performance is easily disturbed by the external environment and the face action [4]. In recent years, with the development of deep learning, convolutional neural networks are gradually applied to the image processing field and have
high accuracy, but there are still some shortcomings; for example, the network size needs to be further compressed so as to save space, and the operation efficiency and recognition accuracy still need to be improved; the image features may be overfitted when extracted; the shift of input data dimensionality leads to easy gradient disappearance and gradient. The existing algorithm is still difficult to handle dynamic images; the adaptability of the algorithm application is poor; more practical applications about face expression recognition are yet to be developed, etc. [5].

To address the aforementioned issues, this paper proposes a structure that combines VGG16 convolutional neural network with long and short-term memory networks. First, VGG16 convolutional neural network is used to extract face features efficiently and accurately, and then, long and short-term memory networks are combined to predict and analyze image sequences for efficient and accurate expression recognition. To maximize the network parameters, a multilayer network TAFMN activation function is presented. The TAFMN activation function can convert difficult nonlinear conditions into linear conditions and extract deep features of pictures more easily and rapidly, allowing deep learning convolutional neural networks to recognize facial expressions with greater accuracy [6]. The model can effectively detect facial expressions on the fer2013 dataset after experimental tests, and this optimized network model has good feasibility and accuracy [7].

2. Review of the Literature

People have been researching and applying face expression recognition since the dawn of artificial intelligence. The present state of facial recognition technology may be classified into two categories: classical approaches and deep learning methods.

2.1. Application of Traditional Methods in Face Expression Recognition. In 2005, Peterson Joshua et al. proposed the method of using infrared illumination to achieve the capture and recognition of human face expressions [8]. Even if the head moves in different situations, this approach employs an active near-infrared light source imaging with a higher intensity than ambient light, and with the matching wavelength optical filter, a steady facial picture may be created. On this image, dynamic Bayesian network (DBN) modeling is used to extract the image features and further eliminate the monotonic changes of the image, forming a dynamic and highly reliable face expression recognition system [9].

With the development of 3D technology, in 2006, Oruganti and Namratha proposed to apply 3D technology to facial expression recognition. First, the facial expression images were captured by a camera, and a preliminary 3D model was obtained by stereo reconstruction of these images. Different expression images will construct different 3D models, extract the deformation parameters of the 3D models, and recognize and classify different expression images by analyzing the deformation parameters [10].

In 2009, Hongxue and Kongtao used local binary pattern (LBP) to do face expression recognition. First, the theoretical positions of the five facial features are calculated using the projection method. Then, the actual positions of the five senses captured by the camera are compared with the previously calculated theoretical positions to perform face expression recognition [11].

AG et al. used support vector machine (SVM) in the face expression identification process in 2015. With the use of the SVM classifier, AG et al. categorised facial emotions into angry, pleased, astonished, sad, disgusted, natural, and scared [12]. Then, the input images were compared with each category.

In 2017, Weilong et al. used a combination of methods for expression recognition, which include local binary patterns, facial markers, histogram of oriented gradient (HOG), and SVM. Firstly, facial flags are set on the input expression image and local binary model is built, then local binary features are extracted using histogram of oriented gradient, and finally, SVM is used to classify the features to complete the final expression recognition [13].

2.2. Application of Deep Learning on Face Expression Recognition. In recent years, the mainstream way of face expression recognition is to use convolutional neural network (CNN) for deep learning. The multiple convolution and acquisition layers in convolutional neural networks can extract higher-level features of face expression images and accurately classify these deep image features, which have achieved good results in FER. After training, convolutional neural networks are currently the best among many kinds of neural networks for image recognition [14].

In 2014, Ijina and Mohan first proposed the use of deep convolutional neural networks for face expression recognition. The previous convolutional neural networks do not analyze face expressions deeply enough and are easily disturbed by external conditions. In contrast, deep convolutional neural networks can use multilayer structure and activation function for local discrimination to analyze deep features of face expression images and reduce the interference of external environment and face action. It was trained on EURECOM kinect facial dataset and achieved good results in face expression recognition.

In 2015, Lei et al. proposed that multichannel convolutional neural networks can be used for face expression recognition [15]. They designed a two-channel convolutional neural network with two channels using different feature extractors: an unsupervised trained convolutional autoencoding feature extractor and a standard convolutional neural network feature extractor. After extracting the features, a fully connected layer is used to synthesize the image features for analysis and classification. Using JAFFE dataset for training, this two-channel convolutional neural network has higher recognition rate [16].

In 2016, Shijia et al. focused on the application of convolutional neural networks for the recognition of real-world faces. The dataset used for the training of the convolutional neural network is mostly from the face expression images on the Internet, which is too small and lacks universality. Since the face expressions to be recognized in real-world scenarios are more complex and variable and have more
environmental interference, the convolutional neural network trained with these expression images cannot solve the real-world face expression recognition problem well. So Shi-jia et al. created their own new face expression dataset and used several different camera models for different lighting conditions during the filming process in order to make the data more general and trained the convolutional neural network with the new dataset created [17].

K. Sha et al. employed deep convolutional neural networks to create a face expression detection system with four components: an input module, a preprocessing module, a recognition module, and an output module in 2017. The system is able to extract deeper image features due to the preprocessing of the input facial expression images, the learning rate of the convolutional neural network is significantly increased, and the K-nearest neighbor algorithm is added to the recognition module for data fitting. In order to have higher generality and make the results more convincing, the system is trained simultaneously with the public dataset Cohn-Kanade and the expression dataset JAFFE, which focuses on Japanese female facial expressions. The experimental results show that the facial expression recognition system can accurately recognize human facial expressions.

In 2018, Zheng et al. improved the algorithm of deep learning and proposed a simpler and faster centralized coordinate learning (CCL) method [18]. This method forces the feature vectors to be decentralized and arranged on a sphere spanning the whole coordinate space in order to enhance the discriminative ability, merges the multiplicative angle residual and additive cosine residual into the soft-max loss function, respectively, and further proposes the adaptive angle residual to enhance the discriminative ability of facial features. To improve the training efficiency, the centralized coordinate learning model is trained with a small CASIAWebface dataset, which has about 10,000 460K facial images, and the training results show that the centralized coordinate learning model can accurately recognize facial expressions of human faces [19].

In 2019, Ananth and Rajendrane proposed a new network architecture for end-to-end facial expression recognition combined with an attention model, which constructs an attention model on the human face and uses a Gaussian function to process the input data and analyze the stereoscopic structure to recognize facial expressions. The core of this network architecture has two parts: the first part is responsible for face expression image capture and correction, first extracting features with encoder, decoder network, and convolutional feature extractor, then arranging pixels into matrices, and performing multiplication of matrices to obtain feature attention maps; the second part is responsible for obtaining the representation and classification of embedded facial expressions. To demonstrate that the designed network architecture has higher recognition rate, Ananth and Rajendrane combine the traditional BU3DFE and CK facial dataset to form a larger and more comprehensive synthetic dataset, use this synthetic dataset for training, and compare the results with other previous network architectures to verify that the network architecture of end-to-end facial expression recognition combined with attention model has recognition with high accuracy [20].

### 3. Emotion Recognition Network Construction

Since the captured facial expression information usually exists in image format, it is often trained using image sequences in deep learning to recognize facial expressions more accurately. Convolutional neural networks have good performance in picture recognition and can extract the depth features of images, while long short-term memory networks are obtained by improving on the recurrent neural network model and have good ability to handle sequence data. As a result, the two may be utilized together for facial expression sequence identification, with the convolutional neural network's feature recognition and the long and short-term memory network's data processing. To begin, convolutional neural networks are utilized to extract deep visual characteristics from sequential facial expression image sequences, removing the impact of face motions and lighting conditions to increase face expression recognition accuracy. The optimal network is then built by combining the long and short-term memory network with a convolutional neural network that extracts deep visual features, and the long and short-term memory network synthesizes the sequence information into the network model using deep learning.

#### 3.1. Convolutional Neural Network Selection

Convolutional neural network is a feed-forward artificial neural network divided into three processes: input, processing, and feedback. It is inspired by the neural tissue cell structure of animals, where individual nerve cells in the brain have only the edge part to respond to external stimuli. Based on this special structure, convolutional neural network performs well in many deep learning tasks such as speech-to-text conversion, image recognition, and language translation. At the receiver side of the convolutional neural network, various types of data such as audio, video, and image can be input, which form one or more vector arrays. The vector arrays have their own unique features, and these features are extracted and compared with the predefined features for classification, and the similarity reaches a certain degree to be classified into one category. In the process of supervised learning, the convolutional neural network is trained with training data of known mapping relationships, and the convolutional neural network can automatically adjust the parameters to form the mapping relationships that meet the actual needs after learning. In this process, a large amount of training data is usually required, and both the amount of data and the depth of the network have a great impact on the model accuracy.

Convolutional neural networks have several layers: a convolutional layer, a pooling layer, a rectified linear unit layer, a fully connected layer, and a loss layer. The convolutional layer is the core component of the convolutional neural network algorithm. The number of layers in the convolutional layer may be adjusted according to the technique and the real requirements, and each layer can extract features from the input picture using different methods. Under nonlinear situations, the pooling layer samples the input data. Maximum pooling layer and average pooling layer are the two most popular forms of pooling layers.
The maximum pooling layer can be used in the final display of the image. Rectified linear layers are usually applied after each convolutional layer. When dealing with nonlinear features such as hyperbolic tangent and S-shaped functions, rectified linear layers use a variety of basic operations to solve the problem of gradient disappearance and gradient explosion, transforming nonlinear problems into linear ones and improving the accuracy rate. The last part of the convolutional neural network is the fully connected layer, which is responsible for processing the final output of the whole neural network. The output is usually an N-dimensional vector, and N is the number that must be chosen for program classification. Suppose a numerical classifier is needed, and if there are 8 numbers, N equals 8. Assuming that the numerical classification process results in an N-dimensional vector of \([0.18 \ 0.02 \ 0.64 \ 0.06 \ 0 \ 0 \ 0 \ 0]\), the test image has a probability of 1 for 18%, 2 for 2%, 3 for 64%, 5 for 6%, and 0 for all other probabilities.

The convolutional neural network model used in this paper is the VGG16 convolutional neural network model. The convolutional layers use \(3 \times 3\) kernels, and the step size and padding are set to 1 to ensure that two adjacent convolutional layers have the same spatial dimension. Rectified linear unit activation is performed immediately after each convolution, and the maximum intersection operation is used at the end of each block to reduce the overall size of the network. The maximum pooling layer uses a filter of size \(2 \times 2\). To ensure that the two adjacent pooling layers have the same spatial dimension, the padding and step size are set to the same, and to ensure that each spatial dimension of the activation mapping from the previous layer is halved, the padding and step size are both set to 2. Then, it goes through 3 fully connected layers, the first two fully connected layers consist of Relu activation function and the last fully connected layer consists of soft-max loss function.

The VGG16 convolutional neural network has two main drawbacks: first, the training speed is very slow because of the complex and large structure with many layers; second, the network architecture has too many parameters, with more than 100 million parameters in total, 90% of which are located in the fully connected layers, which will occupy a large amount of memory space and seriously affect the operation speed.

Based on these two shortcomings, this paper improves the VGG16 convolutional neural network in two aspects, namely, improving the training speed and reducing the number of parameters.

### 3.2. Long Short-Term Memory Network Selection

Long short-term memory network (LSTM) is a special form of recurrent neural network (RNN). Recurrent neural network was born in the 1980s, but it was not until the beginning of this century that it was really used as an algorithm for more effective convolutional neural networks. Recurrent neural networks have different structures, one of which is the long and short-term memory network. The input data of a recurrent neural network is usually in the form of a sequence. After the input sequence, the recurrent neural network processes the sequence recursively in order and outputs it after reaching a certain number of cycles. The structure of a recurrent neural network is shown in Figure 1.

From Figure 1, we can see that in recurrent neural network, the hidden layer, delayers, and outputs share the weights among them, so the recurrent neural network has some memory function. Based on the memorability of recurrent neural network, this network has important applications in video and audio processing recognition, contextual text detection and prediction, and contextual semantic recognition.

The long and short-term memory network introduces unit gates and functions with special functions based on the recurrent neural network, which can not only remember something but also selectively forget and input.

#### 3.3. Optimized Neural Network Structure Construction

In this study, we combine a convolutional neural network with a long and short-term memory network to capture the characteristics of face expression sequences. The long and short-term memory network learns the laws of face expression sequences in detail to perform face expression identification. The neural network contains the following three parts: CNN feature sample layer, LSTM feature learning layer, and SVM feature classification layer. First, the input original image sequence \(X = \{x_1, x_2, \ldots, x\}\), the image of one expression in the database is extracted into \(N\) frames, and the features of the expression image corresponding to \(N\) CNNs are extracted to form the CNN feature sample layer. The CNN feature sample layer extracts the features of the input face image, which takes two-dimensional image data for feature extraction, forms a one-dimensional array, and retains most of the information by performing convolution on the original data, generating image abstract features by cross-convolution and merging, pooling, and other layer-by-layer operations. The CNN feature sample layer uses the VGG16 convolutional neural network, but since the parameters of VGG16 are concentrated in the last three fully connected layers, in order to reduce the size of the network, the last three fully connected layers are removed and only the previous convolutional and pooling layers are taken and input to the average sampling layer, which averages the \(X = +1, x_{-2}\), and \(X\) of the \(K\) consecutive features of the images. Each layer of the CNN is connected to a single LSTM, and the upper \(N\) and lower \(N\) LSTM layers are connected to each other to form an LSTM feature learning layer. The function of the LSTM feature learning layer is to perform feature learning of the feature vectors generated from the images, which includes image data preprocessing, face detection, convolutional feature learning, and feature sampling. Finally, an optimized support vector machine- (SVM-) based feature recognition.
classification layer is formed to perform feature selection using the SVM classifier. The overall structure of the designed DCNNBEA network is shown in Figure 2.

However, there are some problems with this model: first, the initial trial values of the model parameters need to be set reasonably, and the selection of the initial values is very important, and inappropriate values may lead to overfitting of the data; second, although the model has good feature extraction capability, the number of calculations is too many, which leads to too slow operation speed.

To address these two problems, the CNN feature sample layer is improved by using an average sampling layer to remove random interference and increase stability, and its basic principle diagram is shown in Figure 3.

4. Experimental Results and Analysis

4.1. Optimization of Expression Recognition Algorithm Network Parameters. Although the DCNNBEA network model can significantly improve the accuracy of convolutional neural network, it may be overfitted in the process of picture feature extraction, the problem of gradient disappearance in the process of converting two-dimensional data to one-dimensional data, the difficulty of taking reasonable initial values of network parameters, and the poor adaptability of algorithm application. These problems affect the accuracy of the facial expression feature recognition model to a certain extent. Therefore, this paper investigates the problems related to the gradient disappearance during the conversion of two-dimensional data into one-dimensional data and the low recognition accuracy due to the difficulty of taking reasonable initial values of the network parameters and proposes a multilayer network TAFMN activation function based on the Relu function.

4.1.1. Single-Layer Activation Function Research Selection. Since the activation function can substantially improve the working accuracy and efficiency of neural networks, it is widely used in neural network models.

The sigmoid function and tanh function are the two most common activation functions, where the tanh function is a kind of deformation of the sigmoid function.

The formula of the sigmoid function is shown in

\[ s(y) = \frac{1}{1 + e^{-y}}. \]  

(1)

The tanh function can be expressed as

\[ \tanh y = \frac{2}{1 + e^{2y}} - 1. \]  

(2)

Rectified linear unit (Relu) is an integral part of the neural convolutional network, and its main role is to carry out the transformation between nonlinear features and linear features. Since the calculation of nonlinearity is too complicated and difficult, it is necessary to transform the input data into linear features, and the rectified linear unit can effectively improve the accuracy and efficiency of convolutional neural network model training.

Rectified linear unit is a kind of nonlinear activation function, and according to the characteristics of the Relu function itself, its computational expression is as in

\[ \text{Relu}(y) = \begin{cases} 0 & \text{if } y < 0, \\ y & \text{if } y \geq 0. \end{cases} \]  

(3)

From equation (3), it can be seen that the Relu function is much simpler to calculate and only needs to be taken according to two ranges of the input value. It does not need exponential and fractional calculation like the sigmoid function or tanh function, so most activation functions of convolutional neural networks are designed based on the Relu function and improved according to the desired function. The TAFMN activation function proposed in this paper is an improvement of the Relu activation function. To simplify the calculation, the linear calculation method is still used as the main calculation method, but because the gradient of the convolutional neural network model is unstable and will have problems such as gradient disappearance and gradient explosion as the number of layers increases, the initial value selection may also lead to problems such as overfitting or poor fitting performance, so it needs to be combined with other computational methods so that it can be applied to more complex convolutional neural network models.

4.1.2. Activation Function Design for Multilayer Networks. Although the Relu activation function is simple and fast to compute, commonly used, and can effectively increase the accuracy and efficiency of convolutional neural network model training, it still cannot meet the need for higher accuracy and efficiency of convolutional neural network models. In order to improve the computational rate, the parameters...
of convolutional neural network should not be too many; attention should be paid to prevent the problems of gradient disappearance and gradient explosion as well as overfitting. To address these problems, a trainable multilayer maximum network TAFMN activation function is proposed in this paper.

The TAFMN activation function can be used to represent various linear and nonlinear relationships, and can simultaneously have functions that different single-layer activation functions have, such as the ReLu activation function commonly used in convolutional neural networks. Because the maximum output function can provide similar functions as the ReLu activation function, the TAFMN activation function proposed in this paper is a multilayer network activation function that aggregates multiple maximum output activation functions and thus has the activation function functions of each single layer. Even for the hidden and difficult to extract potential features that need to be extracted, it can be achieved by applying the multilayer activation function to the convolutional neural network. There are many complex problems to be solved in the process of face expression recognition, because faces are differentiated, the position and shape of the five senses are different for different expressions, and some subtle expressions are not easy to be detected, so deep features need to be extracted to improve the accuracy of face expression recognition.

The TAFMN activation function has two main features: first, although the TAFMN activation function is a nonlinear function, it is a linear function on the local segments, so the TAFMN activation function can solve the problem of gradient disappearance and gradient explosion during the training of neural network like the ReLu function, which effectively improves the effectiveness and accuracy of the training work of neural network structure; second, the TAFMN can perform deep learning, and since it is a deformation of the maximum output function, the parameters of the activation function which can be changed according to the different neural network models to achieve the purpose of assisting the neural network to extract the effective features. Based on these two features of the TAFMN activation function, it is clear that the TAFMN activation function is capable of solving more complex problems and can be applied to complex network models. However, care should be taken not to have too many parameters; otherwise, it will lead to longer training processing time.

The technique of face facial expression recognition is used in the application, and the process is shown in Figure 4.

4.2. Experimental Dataset. During the experiments, the designed deep learning convolutional neural network is trained using the expression images in the fer2013 dataset. The fer2013 dataset is derived from the 30th Machine Learning Conference held in Atlanta, Georgia, USA. The fer2013 database has nearly 30,000 face expression training images with 48 × 48 pixels, and there are seven expressions in this database: neutral, happy, sad, surprised, angry, disgusted, and fearful. The distribution of fer2013 expressions is shown in Figure 5.

4.3. Experimental Training Design and Result Analysis of VGG16 Model. In this paper, we first combine the improved VGG16 pretraining model and the long and short-term networks to form the DCNNBEA network and then train the network model with fer2013 dataset. The framework of TensorFlow is used to write the training network in python, and a computer with Windows 10 is used for training and accelerated with a GTX1060 GPU. The learning rate is initialized to 1, and batch normalization is added after each convolutional layer and pooling layer. The batch size was set to 256. The time used for training was 78 hours. The best classification accuracy was obtained after 270 time periods for the selected models. 128000 randomly selected data in fer2013 dataset were used as training data.

Figure 6 shows the training accuracy and testing accuracy of the VGG16 pretrained model.

From Figure 6, we can see that as the number of training steps increases, the accuracy rate will be higher. When the number of training steps is 50,000, the training accuracy is 0.65 and the test accuracy is 0.42; when the number of training steps is 50,000, the training accuracy is 0.65 and the test accuracy is 0.42; when the number of training steps is 50,000, the training accuracy is 0.65 and the test accuracy is 0.42;
steps is 70,000, the training accuracy is 0.74 and the test accuracy is 0.46; finally, when the number of training steps is 100,000, the training accuracy is 0.967 and the test accuracy is 0.498. After convergence using batch normalization of about 100000 steps, the accuracy basically no longer varies with the number of training steps. Meanwhile, the learning rate has a great influence on the convergence speed of the curve, and the curve converges fastest when the learning rate is 1.

When the number of training steps is less than 100,000, both the training accuracy and test accuracy grow rapidly, but when the number of training steps is greater than 100,000, the training test rate and test accuracy gradually level off, and finally, the training test rate grows to 96.7% and the test accuracy grows to 49.8%. It indicates that the designed DCNNBEA network structure has a high accuracy rate and can effectively recognize human face expressions.

From Figure 7, to prove that the neural network structure designed in this paper has better results compared with other networks, the experiment was repeated on AlexNet, VGG16, VGG19, and Resnet152 network structures. The experimental results are shown in Figure 7. The final test accuracies obtained from this experiment on these networks were 15.2%, 37.4%, 39.7%, and 48.7%, respectively. On these
networks, Resnet152 had the highest test accuracy: 48.7%, but the identification method proposed in this paper obtained an even higher test accuracy: 49.8%.

4.4. Experimental Results and Analysis after Activation Function Improvement. The TAFMN activation function was added to the DCNNBEA network for training. The TensorFlow framework was used to write the training network in python, and the training was performed on a Windows 10 computer with a GTX1060 GPU for acceleration to improve the running efficiency. The total training time was 68 hours. To ensure maximum convergence speed, the learning rate was set to 1, and batch normalization was added after each convolutional layer and pooling layer, with batch size set to 256.

The running results of the training and testing process are shown in Figure 8.

As can be seen in Figure 8, the training accuracy and test accuracy are much closer than before using the activation function, and both values have increased. When the number of iterations was 40, the training accuracy was 0.6 and the test accuracy was 0.58; when the number of iterations was 80, the training accuracy was 0.67 and the test accuracy was 0.63; when the number of iterations was 120, the training accuracy was 0.68 and the test accuracy was 0.64. When the number of iterations exceeded 130, the test accuracy leveled off and finally stabilized at 68.4%. This indicates that the overall recognition accuracy of the convolutional neural network has been improved after adding the activation function, and the overfitting prevention performance of the network training has also been improved.
5. Conclusion

5.1. Summary of Full Paper Work. This paper focuses on deep learning-based facial expression feature recognition technology. In view of the problems of the current deep learning-based face expression recognition technology, the text makes reasonable improvements to the existing network structure and activation function to improve the accuracy of face expression recognition, and the experimental results verify the effectiveness of the network structure and activation function designed and proposed in the paper.

The main research of this paper has two main points. First, a network framework for deep learning algorithm for face expression recognition is designed. When processing dynamic face expression sequences, the size of the convolutional neural network framework is too large due to too much data and parameters, which seriously affects the training speed and operation efficiency of the system. To address these problems, this paper designs a deep learning convolutional network structure that can recognize facial expression features, namely the DCNNBEA network structure. This structure is based on the VGG16 network structure by replacing the fully connected layer with the average sampling layer, which has more concentrated parameters, and combining it with the long and short-term memory network, which can achieve both simplifying the structure of the convolutional layer of the network, reducing the parameters, and improving the processing ability of the input face expression images. Finally, the network is trained with fer2013 standard face expression dataset, and the accuracy of face expression recognition test is 49.8%.

Second, the TAFMN activation function is proposed and the parameters of the face expression recognition algorithm are optimized. The selection of parameters affects the training effect of the whole convolutional neural network. Too many parameters that are too complex will produce gradient explosion, while too few parameters or too much similarity will lead to overfitting, all of which will reduce the accuracy of face expression recognition. In this paper, we design a multilayer network TAFMN activation function, which can modify its own parameters by learning to perform linear computation, nonlinear computation, and the transformation of both. By applying the TAFMN activation function to the convolutional neural network, the parameters can be continuously optimized to reduce the possibility of gradient elimination or gradient explosion. The neural network is trained with fer2013 standard face expression dataset, and the convolutional neural network face expression recognition algorithm with optimized parameters obtains a test accuracy of 68.4%, which is much higher than that of the neural network without TAFMN activation function.

5.2. Outlook of Next Work. Although some work has been done in this paper on the research of facial expression feature recognition based on deep learning, these works cannot meet the current demand of human-computer interaction system for facial expression recognition yet, and the convolutional neural network model still has some defects and problems, so two future research directions are proposed here.

First, there is still room for improvement in data processing. When capturing face expression images, the lighting of the external environment and the degree of face tilt can interfere with the input data; in addition, the original size and pixel characteristics of the photo can also affect the accuracy of face expression recognition. Therefore, it is necessary to further improve the algorithm to extract deeper features, reduce the sensitivity of the convolutional neural network model to these influencing factors, and improve the accuracy of face expression recognition.

Second, although improvements have been made in network optimization, there is still room for further improvement. The current training method has a large randomness, so we should study the principle of convolutional neural network in depth and explore a better training method, so that the neural network can analyze the reasons for matching the results with the dataset by itself and improve the efficiency of deep learning.

Data Availability

The labeled datasets used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This study reports the results of the Basic and Applied Basic Research Projects of Guangzhou Basic Research Program in 2022 (No. 202201010106), the Special Project of Guangdong Provincial Education Department in Key Areas (No. 2021ZDZX1104), the Project of Guangzhou Philosophy and Social Science (No. 2022GZJG241), and the Key Research Project of Guangzhou Nanyang Polytechnic College (No. NY2021KYXD01).
References


