Hindawi
Scientific Programming
Volume 2022, Article ID 5203022, 10 pages
https://doi.org/10.1155/2022/5203022

Research Article

Expression Recognition of Classroom Children's Game Video Based on Improved Convolutional Neural Network

Xiaohong Li

Henan Institute of Economics and Trade, Zhengzhou, Henan 450000, China

Correspondence should be addressed to Xiaohong Li; xiaoxiao518@henetc.edu.cn

Received 12 February 2022; Revised 8 March 2022; Accepted 11 March 2022; Published 8 April 2022

Academic Editor: Baiyuan Ding

Copyright © 2022 Xiaohong Li. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Humans express emotions in many ways, such as gestures, limbs, and expressions. Among them, facial expressions are the most intuitive way to express human inner emotional activities in human-to-human communication. With the rapid development of computer vision, facial expression recognition is an important research topic in the field of computer vision. It plays a key role in nonverbal communication and can be applied to human-computer interaction, social robotics, video games, and other fields. Traditional expression recognition algorithms require complex manual feature extraction, which takes a long time, and the accuracy of expression recognition in complex scenes is not high. However, with the development of deep learning, especially the convolutional neural network, facial expression recognition technology has also developed rapidly, and the recognition accuracy has been greatly improved. This paper studies the facial expression recognition method of classroom children's game video based on convolutional neural network and proposes a convolutional neural network with deeper layers. The full connection is modified to 4 layers of convolution, 4 layers of pooling, and 2 layers of full connection. Firstly, the facial expression image is preprocessed by, for example, key point location, face cropping, and image normalization; then, the convolutional layer is used to extract the low-dimensional and high-dimensional feature information of the face image; and the pooling layer is used to extract the face image. The feature information is dimensionally reduced. Finally, the softmax classifier is used to classify and recognize the expressions of the training sample images. In order to improve the accuracy of expression recognition, a self-made set of labeled pictures was added to the expression training set. Simulation and comparison experiments show that the improved model has higher accuracy and smoother loss curve, which verifies the effectiveness of the improved network.

1. Introduction

Expression recognition refers to the separation of a specific expression state from a given static image or dynamic video sequence, thereby determining the psychological emotion of the recognized object. With the rapid development of computer technology, artificial intelligence technology, and related disciplines, the automation degree of the whole society continues to increase, and people's demand for human-computer interaction similar to the way people communicate with each other is growing. If computers and robots can understand and express emotions like humans, this will fundamentally change the relationship between humans and computers, enabling computers to serve humans better. Expression recognition is the basis of emotion understanding, the premise for computers to understand people's emotions, and an effective way for people to explore and understand intelligence [1–5].

Facial expression recognition (as shown in Figure 1) refers to separating a specific expression state from a given static image or dynamic video sequence, so as to determine the psychological emotion of the recognized object, realize the understanding and recognition of facial expression by computer, and fundamentally change the relationship between humans and computers, so as to achieve better human-computer interaction. Therefore, facial expression recognition has great potential application value in the field of education. In particular, the evaluation of classroom teaching efficiency can have a great application, but there is currently a lack of an effective expression recognition method for classroom teaching. In order to solve the above shortcomings, it is urgent to provide a solution [6–9].

Facial expression recognition technology has a wide range of application scenarios and is mainly used in the following real-world scenarios:

(1) In the field of human-computer interaction, the traditional mouse and keyboard and human input commands are abandoned, and expressions, actions, and voices are used to control and operate the machine, so that the machine can understand human emotions and make corresponding responses, thereby saving time and improving machine performance operating efficiency.

(2) In the field of safe driving, the facial expression recognition technology can be used to monitor the driver's facial expression status at all times, so as to determine whether driver fatigue occurs.

(3) In the field of short video, with the launch and continuous development of 5G networks and smartphones, mobile phones have gradually become one of the main ways for people to understand current affairs and for leisure and entertainment. People can watch short videos anytime and anywhere through their mobile phones. The time of each video is only a few dozen seconds. In this short period of time, users can watch what the publisher wants to express or learn about current news and current affairs. However, the current short videos can only be recommended according to the type of videos that viewers usually like. If it can be assisted by facial expression recognition, the camera captures the user's facial features and then analyzes and judges the category of their emotions to recommend short videos that match their current emotions type.

(4) In terms of case detection, when examining suspects, the machine can automatically identify and learn complex psychological changes based on changes in the suspect's facial expressions, so as to figure out the other party's behavioral motives and provide certain help for the police to solve the case [10–16].

Facial expression recognition is a complex learning process, and how to improve the recognition rate is a problem that researchers need to solve. With the rapid development of artificial intelligence, many scholars have devoted themselves to the field of expression recognition, applied some related algorithms such as image processing and feature extraction, and achieved good results. However, how to further improve the recognition rate still needs to be studied today.

In 1862, French researcher Duchenne studied electrical stimulation of the individual facial muscles responsible for the production of facial expressions. Later, in 1872, in Darwin's work "The Expression of the Emotions in Man and Animals," he described the changing process of facial expressions in the process of human-to-human communication, indicating that one of the important ways for humans to express emotions is the facial expressions. A comprehensive and in-depth study of facial expression recognition began with the MIT Media Lab led by Professor Picard, which applied facial emotion recognition technology to the analysis of social behaviors of autistic teenagers. MIT has also developed a robot that can recognize the facial expressions of the other person in the communication with the interlocutor, then analyze the facial expression of the interlocutor, and make different responses according to the results of the analysis. Since then, with the rapid development of computer vision, the research on facial expression recognition by domestic and foreign researchers has progressed rapidly, and different methods for facial expression recognition research have emerged and have achieved remarkable results. For example, the optical flow method proposed by Mase et al. achieves an 80% facial expression recognition rate. With the rise of deep learning, neural networks are favored by researchers due to their high recognition rate. Therefore, the facial expression recognition algorithm based on deep learning has become a research hotspot. For example, Liu et al. proposed an AUDN (AU-inspired Deep Network), which divides facial expressions into different facial expression units, encodes them, and uses deep neural networks for deeper feature extraction, so that the network model can achieve better facial expression recognition effect. Lopes et al. used preprocessing to extract specific features of facial expression. They established a five-layer convolutional neural network to extract facial expression features and input the extracted facial expression features into a classifier for facial expression analysis. The proposed method achieved a facial expression recognition rate of 97.75% on the facial expression dataset CK+. However, the traditional facial expression recognition method requires step-by-step processing. First, the facial image features are manually extracted, and then the corresponding classifiers are selected for classification. This process is relatively complicated. The images collected in real scenes are mainly affected by the illumination angle and posture. The influence of factors such as difference and occlusion greatly reduces the robustness of traditional methods [17–23].

Figure 1: Facial expression recognition.

With the introduction and continuous development of deep learning, deep learning methods have gradually shown good results in computer vision tasks, among which convolutional network and recurrent network algorithms have been used in feature extraction, image recognition, and classification tasks. CNN is a deep neural network composed of convolutional layers, pooling layers, nonlinear activation functions, and fully connected layers. The local features of the input data itself are used for autonomous learning, the global features in the image, and the data enhancement methods such as translation, scaling, and rotation of the image make it robust. The CNN algorithm does not need to manually extract features but performs end-to-end learning and training by directly inputting the pixel values of the image sequence, reducing the degree of dependence on facial image samples and data preprocessing methods. Therefore, CNN has made breakthroughs in tasks such as image processing, face recognition, automatic detection, and scene analysis [24–30].

2. Convolutional Neural Network

At present, the research method using deep learning is the most effective method in solving the problem of facial expression recognition, and using artificial neural network models of different depths in the convolutional neural network has different effects on the lighting environment, different angles, posture changes, whether there is occlusion, and other factors. Feature extraction comes from facial images. In most computer vision tasks, the artificial neural network of deep learning method can avoid the tedious process of manually extracting facial expression image features in traditional methods and extract facial expression features through autonomous learning, so that the obtained image features have strong discrimination. At the same time, the CNN model has high accuracy and better robustness. In addition, the convolutional network model in the deep learning method integrates each link of the traditional method into an end-to-end network model for learning and training, which effectively reduces the complexity of the target task.

Neural network is an abstract mathematical model proposed and developed on the basis of modern neuroscience, which aims to reflect the structure and function of the human brain. A neural network is composed of one or several neurons; that is to say, a neuron is the basic unit of a neural network, as shown in Figure 2, which is a neuron structure. From the neuron structure diagram, the output h can be obtained, and its formula is as follows:

$$h = f\left(\sum_{i=1}^{n} w_i x_i + b\right), \tag{1}$$

where x_i is the ith input, w_i is the weight of the ith input data, b is the bias value, f is the activation function, and h is the output.

2.1. Convolutional Layer. Convolution calculation is the core operation of convolutional neural network, and it is also a special linear operation. The convolution layer is calculated by multiple convolution kernels to form multiple feature maps. Figure 3 is an example of a convolution calculation, where the size of the input image is $3 * 3$, the red square indicates a size of the convolution kernel of $2 * 2$, the sliding step size is 1, the size of the generated feature map is $2 * 2$, and the convolution calculation process is represented in the dotted box.

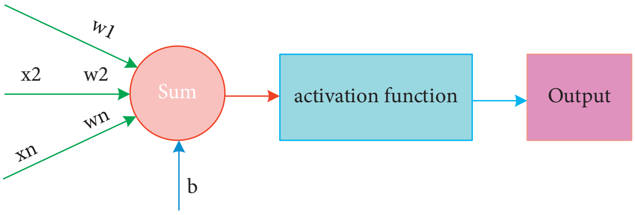


Figure 2: The basic unit of a neural network.

Suppose the size of the input image is $W_0 * W_0$, the size of the convolution kernel is $K * K$, the sliding step size is S, the padding is P, and the size of the output feature map is $W_1 * H_1$; then, the calculation formulas of W_1 and H_1 are as follows:

$$\begin{aligned} W_1 &= \frac{(W_0 - K + 2P)}{S} + 1, \\ H_1 &= \frac{(H_0 - K + 2P)}{S} + 1. \end{aligned} \tag{2}$$

The number of channels of the output feature map is equal to the number of convolution kernels. The specific operation of padding is to add 0 to the periphery of the input image and add a layer, and padding is recorded as 1. The role of padding is to prevent the loss of image edge information.

2.2. Pooling Layer. In the convolutional neural network, the function of the pooling layer is to compress the image and reduce the dimension of the feature map, so it is also called the downsampling layer. In the convolutional neural network, usually after a series of convolution operations, a pooling layer is used to halve the width and height of the feature map extracted by the convolutional neural network, and through the compression of the features, the main image of the image is achieved. The purpose of efficient extraction of feature information is to simplify the computational complexity and improve the computational speed of the network. There are many types of pooling, such as max pooling, average pooling, overlap pooling, and spatial pyramid pooling. The most commonly used pooling methods in image recognition and classification tasks are max pooling and average pooling. The maximum pooling refers to taking the largest feature point in the neighborhood, which means that the maximum value of the feature is saved; the average pooling refers to averaging the feature points in the neighborhood, which means that the average value of the feature is saved. The calculation process of pooling is similar to the calculation process of convolution. During pooling, the convolution kernel goes through the feature map from top to bottom and from left to right according to a certain step size, and the corresponding window area will be pooled. The calculation process of pooling is

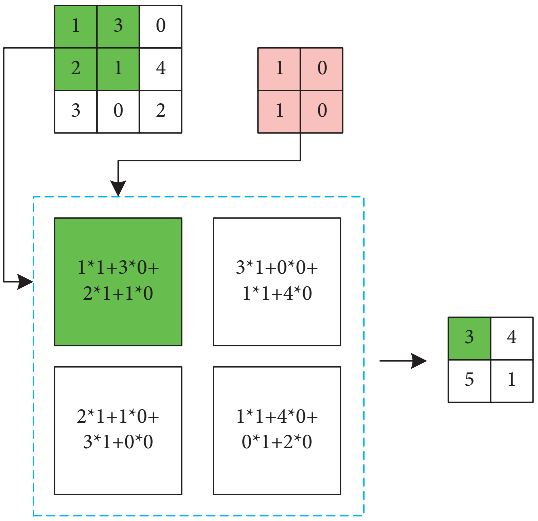


Figure 3: An example of a convolution calculation.

$$y = \text{pooling}(a), \tag{3}$$

where pooling ($*$) represents the pooling function and a is the result after convolution.

2.3. Activation Function. In order to ensure that the convolutional network can better fit the data model, an activation function is usually added after the feature information is extracted by the convolution module for nonlinear calculation. The four activation functions often used in current convolutional network models are the Sigmoid function, the tanh function, the ReLU function, and the Leaky ReLU function.

(a) Sigmoid function: it is also called S-model function. By mapping variables to (0, 1), it has the characteristics of monotonic continuity, limited output range, and easy derivation. The function formula is expressed as follows:

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \tag{4}$$

(b) Tanh function: it is the hyperbolic tangent function. Its shape is similar to the Sigmoid function, but the whole function takes the zero point as the center of symmetry, and the transformation range is (−1, 1), which can solve the mean shift phenomenon of the Sigmoid function. The function expression is shown as follows:

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}. \tag{5}$$

(c) ReLU function: ReLU function is the most commonly used nonlinear activation function in CNN model, which has the characteristics of simple expression and fast operation speed. The most important thing is to solve the problem of gradient dispersion during model backpropagation. In addition, the ReLU function's feature of setting the negative semiaxis input to zero can make the connections between the network's convolutional layers more sparse during training, but this operation can also lead to neuron death. The function expression is shown as follows:

$$\text{relu}(x) = \begin{cases} x, & \text{if } x > 0, \\ 0, & \text{if } x \leq 0. \end{cases} \tag{6}$$

(d) Leaky ReLU function: this function is the same as the ReLU function on the positive half axis and performs a simple weighting operation on the negative half axis. It resolved the issue of neuron death caused by the ReLU function in training. Its function expression is shown in (7), where a is a very small fixed value. In this case, the function becomes PReLU.

$$\text{leaky relu}(x) = \begin{cases} x_i, & \text{if } x > 0, \\ ax_i, & \text{if } x \leq 0. \end{cases} \tag{7}$$

2.4. Dropout Layer. Convolutional neural networks are prone to overfitting problems. The so-called overfitting means that the training set shows a high accuracy rate, but the accuracy rate is poor in the test set. The proposal of dropout technology has effectively improved this problem. The idea is to randomly discard some neurons with a certain probability during the training process to reduce the network's dependence on certain neurons. The network structure after each dropout is different. The results can be regarded as the average of multiple models, which can improve the generalization ability of the network.

Suppose the output of the ith hidden unit in the $l+1$th layer of the standard neural network is y_i, y represents the output of the lth layer, w represents the weight, b represents the bias, and $\sigma()$ represents the activation function; then, the calculation formula of y_i is as follows:

$$y_i = \sigma\left(w_i{}^{*} y + b_i\right). \tag{8}$$

If dropout is added, the calculation formula of y_i is as follows:

$$\begin{aligned} &r \sim \text{Bernoulli}(p), \\ &\tilde{y} = r^{*} y, \\ &y_i = \sigma\left(w_i{}^{*} \tilde{y} + b_i\right). \end{aligned} \tag{9}$$

In the formula, r is the generated probability vector, which is multiplied by y to get a reduced version of the output vector y, and y is applied to each hidden unit of the next layer to obtain y_i, which is equivalent to sampling a large network.

3. Improving Convolutional Neural Networks

The solution of expression recognition mainly consists of 4 steps: face detection, data preprocessing, feature extraction, and expression classification. The most used method at present is facial expression recognition through deep learning, which combines the two steps of feature extraction and classification to achieve an end-to-end training mode. However, the traditional method is to extract various features in the image and then select the corresponding classifier for identification.

For expression recognition tasks, the selection of expression data samples is equally important. Therefore, this paper will describe in detail different expression datasets and related theoretical and practical methods used in expression recognition tasks. The facial expression recognition process based on deep learning is shown in Figure 4.

In the study of the FER problem, many databases have been subjected to comparative experiments by many researchers. Traditional methods use two-dimensional still images or two-dimensional dynamic image sequences for the research of expression recognition. In recent years, the spontaneous facial expression task has formed a new research hotspot in the process of research development. The application of 3D face image and expression analysis is of great help to the understanding of the internal subtle structural changes of spontaneous expressions. This section will briefly introduce relevant datasets related to the research of expression recognition problem, among which there are various common 2D and 3D dynamic image sequences as well as still images.

3.1. CK+ Facial Expression Database. The CK+ (Extended Cohn–Kanade) facial expression database contains 593 image sequences. The last frame of each image sequence contains action unit (AU) markers. Among all image sequences, 327 have expression labels, including spontaneous expressions and posed expressions. The 123 participants were between the ages of 18 and 30, and most were women. The resolutions of the images are the precision of the grayscale values, which is 8 bits.

3.2. JAFFE Expression Database. There are 213 gray images in the facial expression database, including seven facial expressions, six of which are basic facial expressions and one neutral facial expression. The original unit of each image is described as pixel.

3.3. FER2013 Expression Database. The FER2013 expression database is a face recognition contest database provided by the Kaggle website in 2013. There are 35,887 grayscale images in the database, including a total of seven facial expressions. Each expression corresponds to a numerical label, where 0 = anger, 1 = disgust, 2 = fear, 3 = happy, 4 = sad, 5 = surprised, and 6 = neutral. The original size of each image is that all images were downloaded from the Internet.

3.4. MMI Expression Database. The MMI Facial Expression Database consists of over 2900 video sequences and high-resolution still images of 75 participants. It fully annotates the presence of AU in video sequences and partially encodes it at the frame level, indicating whether each frame is in the neutral, onset, vertex, or offset phase. There were 75 participants, both male and female, ranging in age from 19 to 62. The original size of each face image is pixels.

3.5. BP4D Spontaneous Expression Database. BP4D Spontaneous (Binghamton-Pittsburgh 4D Dynamic Spontaneous) Expression Database is a 3D video database including spontaneous expressions of 41 young adults (23 females, 18 males). Participants were 18–29 years old, 11 were Asian, 6 were African American, 4 were Hispanic, and 20 were European American. This database facilitates the exploration of 3D spatiotemporal features in fine facial expressions, leading to a better understanding of the relationship between posture and motion dynamics in facial AU, and a deeper understanding of naturally occurring facial movements. The original size of each face image is pixels.

3.6. B+ Expression Database. The B+ (Extended Yale B face expression) database consists of 16,128 facial images of 28 subjects. The subjects are photographed from 9 different poses as needed, and 64 shooting parameters are used to shoot under a single set of light sources. The original size of each face image is pixels.

3.7. KDEF Expression Database. KDEF (The Karolinska Directed Emotional Face) expression database consists of 4,900 facial expression images of 70 subjects, each photographed from five different angles, and each angle takes multiple facial expressions, including seven different expressions. The original size of each face image is pixels.

Convolutional layer, downsampling layer, fully connected layer, and output layer are the general structure of CNN. It mainly uses the two basic ideas of local perception and weight sharing. This paper uses the CNN after transforming the number of layers to do the task of expression classification. We improve the convolutional neural network LeNet, from the original 2-layer convolution, 2-layer pooling, and 1-layer full connection to 4-layer convolution (C1, C2, C3, C4), 4-layer pooling (S1, S2), S3, S4), and 2-layer full connection. After the convolutional layer, a Rectified Linear Unit (ReLU) activation function is added, and Batch Normalization (BN) is added before the activation function for normalization to prevent the disappearance of gradients. Finally, dropout technology is used to prevent overfitting.

In the convolution layer, features are extracted by using convolution kernels, and the number of convolution kernels is the same as the feature map. Generally, the number of convolution kernels increases with the depth of the convolutional neural network, in order to better extract the high-level features of the input image. The corresponding operation is shown in Figure 5(a). The dark area in the left image of Figure 5(a) indicates that the convolution kernel acts on the image pixel. The convolution kernel is multiplied by the corresponding image pixel and then added to obtain the value of the dark area in the right image; the rest of the

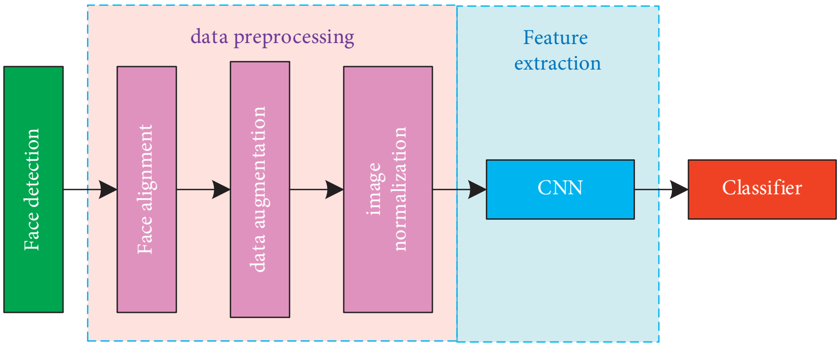
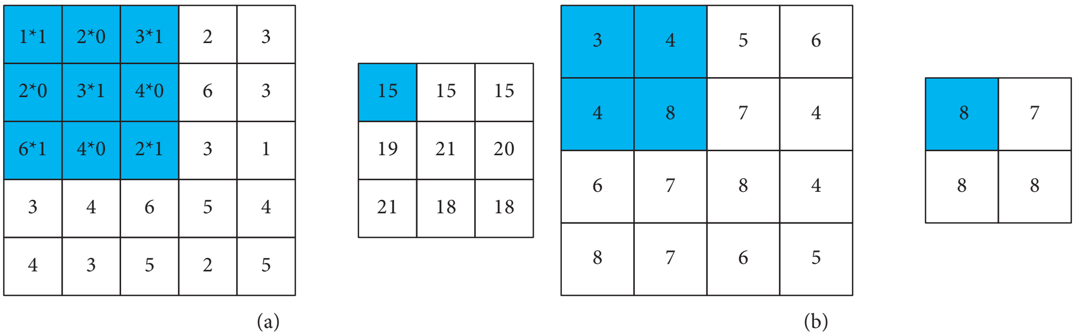


Figure 4: The facial expression recognition process based on deep learning.



1*1	2*0	3*1	2	3
2*0	3*1	4*0	6	3
6*1	4*0	2*1	3	1
3	4	6	5	4
4	3	5	2	5

15	15	15
19	21	20
21	18	18

(a)

3	4	5	6
4	8	7	4
6	7	8	4
8	7	6	5

8	7
8	8

(b)

Figure 5: Operation. (a) Convolution operation. (b) Max pooling operation.

values in the right image of Figure 5(a) are calculated by moving the position of the convolution kernel in the image.

In the convolutional neural network, sampling operations are divided into two types, namely, upsampling and downsampling. There are two commonly used downsampling methods: maximum downsampling and average downsampling. Maximum downsampling is also called max pooling, and average downsampling is also called average pooling. In order to avoid the problem that the feature dimension of the feature map extracted by the convolutional layer is too high, a pooling layer is often connected after the convolutional layer for dimensionality reduction. In this paper, the maximum pooling layer is used for feature dimensionality reduction. If the input image is large, it is also possible to connect continuous 2-layer pooling and perform 2 dimensionality reduction operations. The feature training classifier learned by this method will not have the problem of excessive dimensionality. At the same time, the downsampling operation can reduce the sensitivity of the feature map output to rotation, scaling, translation, etc. The size of the feature map after downsampling becomes the original $2n/s$, where n is the size of the sampling window. This paper uses max pooling, and its operation is shown in Figure 5(b). The dark area in the left picture shows that the sampling window acts on the image pixels, and the maximum value of the image pixel in the sampling window is taken out as the final sampling result, which corresponds to the value of the dark area in the right picture of Figure 5(b). The rest of the values on the right are calculated by shifting the position of the sampling window in the image.

The fully connected layer is connected before the output layer of the CNN, and there is generally 1 or 2 layers at the back of the CNN. Its connection method is special; each neuron in the fully connected layer must be connected with all the neurons in the previous layer to integrate the local information in the convolutional layer and the pooling layer. A ReLU activation function is added after each neuron in the fully connected layer. The input of the fully connected layer must be an array, and it must be one-dimensional. Therefore, the two-dimensional array output by the pooling layer S4 is converted into a one-dimensional array, and then all the converted one-dimensional arrays are connected, and finally becomes 1 A 4 608-dimensional ($3 \times 3 \times 512 = 4\,608$) feature vector, which is used as the input of the fully connected layer. The fitting ability and training speed of the network are closely related to the number of neurons in the fully connected layer, so it is necessary to select a suitable number of neurons. Through the experimental test, the network learning effect is better when the number is 800.

4. Simulation Experiments

The experimental environment of the algorithm in this paper is as follows: Ubuntu 16.04 system, Intel Core i5-7200U CPU, and 8 GB of memory. Because most of the images in the public dataset have complex backgrounds and different light intensities, the static images in the public dataset CK+ are selected for the experiments in this paper, and some self-collected images are added to expand the dataset. We select 7 kinds of expressions, with a total of 648 face image samples,

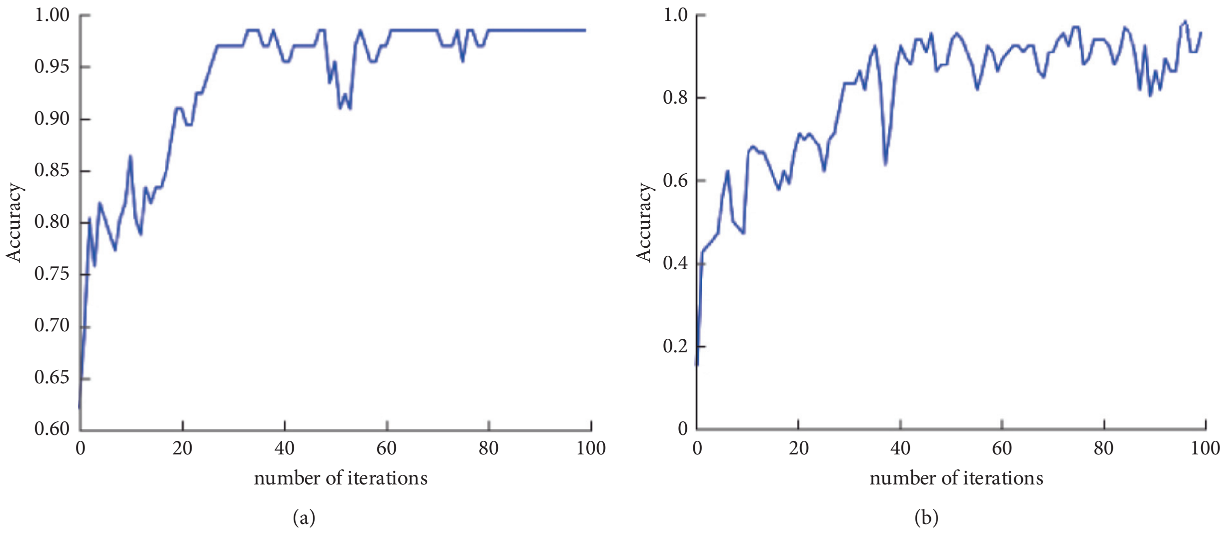


Figure 6: Accuracy. (a) Proposed method. (b) LeNet.

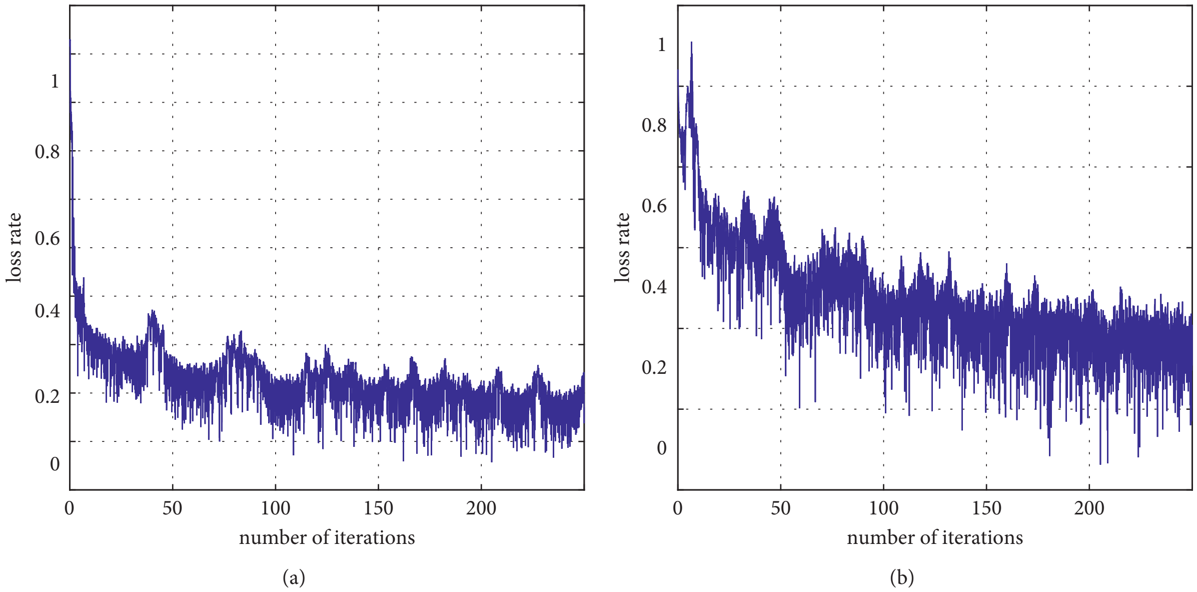


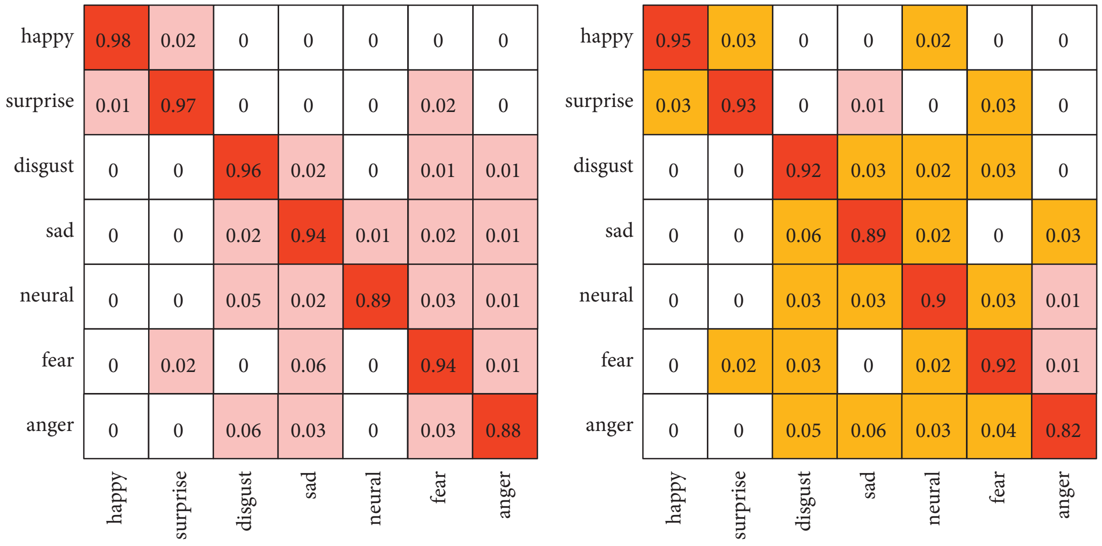
Figure 7: Loss rate. (a) Proposed method. (b) LeNet.

of which there are 90, 50, 138, 36, 50, 166, and 118 samples of angry, scared, happy, neutral, sad, surprised, and disgusted expressions, respectively. These 648 samples are all used to train the training set of the model. The test set used in the experiment includes some remaining static images and self-collected images of the CK+ dataset, including 128 samples of 7 kinds of expression labels.

We convert the image format of the training set to CSV format, then input it into the improved convolutional neural network in this paper for training, and also input it into the LeNet network before the improvement for training. The network accuracy image in this paper is shown in Figure 6(a), the LeNet accuracy image is shown in Figure 6(b), the image of the network loss function in this paper is shown in Figure 7(a), and the image of the LeNet loss function is shown in Figure 7(b).

From the accuracy curve, it can be seen that the accuracy curve of the improved network in this paper is faster and smoother than the accuracy curve of the LeNet network. According to the loss function curve, it can also be seen that the loss function of the improved network and the loss of the LeNet network are faster and smoother. By evaluating the performance of the model based on these two indicators, it can be seen that the improved network is more robust than the LeNet network.

In order to further illustrate the performance advantages of the improved network in this paper compared with the original network LeNet, the test set is input into the improved network and the LeNet network, and the confusion matrix is used as the performance evaluation index. The confusion matrices of the two models are shown in Figure 8. It can be seen that the average recognition rate of the seven



	happy	surprise	disgust	sad	neural	fear	anger
happy	0.98	0.02	0	0	0	0	0
surprise	0.01	0.97	0	0	0	0.02	0
disgust	0	0	0.96	0.02	0	0.01	0.01
sad	0	0	0.02	0.94	0.01	0.02	0.01
neural	0	0	0.05	0.02	0.89	0.03	0.01
fear	0	0.02	0	0.06	0	0.94	0.01
anger	0	0	0.06	0.03	0	0.03	0.88

(a)

	happy	surprise	disgust	sad	neural	fear	anger
happy	0.95	0.03	0	0	0.02	0	0
surprise	0.03	0.93	0	0.01	0	0.03	0
disgust	0	0	0.92	0.03	0.02	0.03	0
sad	0	0	0.06	0.89	0.02	0	0.03
neural	0	0	0.03	0.03	0.9	0.03	0.01
fear	0	0.02	0.03	0	0.02	0.92	0.01
anger	0	0	0.05	0.06	0.03	0.04	0.82

(b)

Figure 8: Confusion matrix on test set. (a) Proposed method. (b) LeNet.

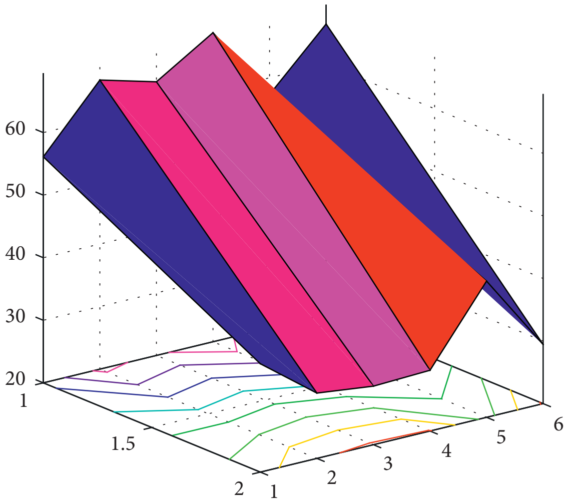


Figure 9: Prediction.

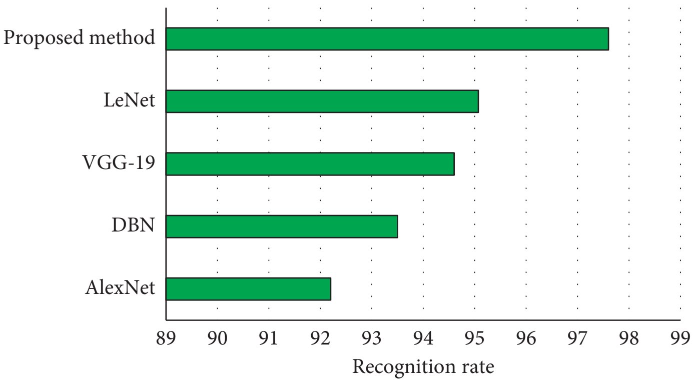


Figure 10: Recognition rate.

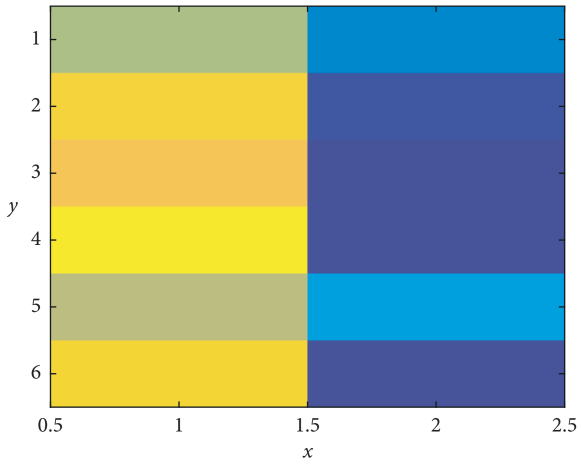


Figure 11: x and y variation.

expressions on the test set of the improved network is 93.7%, and the average recognition rate of the seven expressions of LeNet on the test set is 90.5%, indicating that the recognition accuracy of the improved network has been improved. It is also proved that the robustness of the improved network is better compared with the original network. The prediction is shown in Figure 9.

In order to further verify the recognition rate of the CNN model proposed in this paper, under the same hardware environment, the same dataset was input into the following four networks: DBN, AlexNet, VGG-19, and LeNet for experiments, and they were obtained in training. It can be seen from the results in Figure 10 that the average recognition rate of the improved CNN in this paper is increased by 5.4% compared with the classic AlexNet algorithm. Compared with the LeNet algorithm before the improvement, the recognition rate is also increased by 2.53%. The x and y variation are shown in Figure 11.

5. Conclusion

Humans express emotions in many ways, such as gestures, limbs, and expressions. Among them, facial expressions are the most intuitive way to express human inner emotional activities in human-to-human communication. With the rapid development of computer vision, facial expression recognition is an important research topic in the field of computer vision. It plays a key role in nonverbal communication and can be applied to human-computer interaction, social robotics, video games, and other fields. Traditional expression recognition algorithms require complex manual feature extraction, which takes a long time, and the accuracy of expression recognition in complex scenes is not high. However, with the development of deep learning, especially the convolutional neural network, facial expression recognition technology has also developed rapidly, and the recognition accuracy has been greatly improved. This paper studies the facial expression recognition method of classroom children's game video based on convolutional neural network and proposes a convolutional neural network with deeper layers. The full connection is modified to 4 layers of convolution, 4 layers of pooling, and 2 layers of full connection. Firstly, the facial expression image is preprocessed by, for example, key point location, face cropping, and image normalization; then, the convolutional layer is used to extract the low-dimensional and high-dimensional feature information of the face image; and the pooling layer is used to extract the face image. The feature information is dimensionally reduced. Finally, the softmax classifier is used to classify and recognize the expressions of the training sample images. In order to improve the accuracy of expression recognition, a self-made set of labeled pictures was added to the expression training set. Simulation and comparison experiments show that the improved model has higher accuracy and smoother loss curve, which verifies the effectiveness of the improved network.

Although the recognition effect has been greatly improved, the training speed of the improved network is lower than the training speed of the original classic network. Therefore, how to optimize the network to improve the recognition rate and also ensure the speed of network training is also the future problem to be improved.

Data Availability

The dataset can be accessed upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest.

Acknowledgments

The author thanks 2022 Henan Province Key Research and Development and Promotion Special Project (Science and Technology Research, Soft Science Research Support Project) and Research on the Cultivation Path of "Artisan Spirit" for Students in Henan Vocational Colleges under the Background of Quality Improvement and Excellence (no. 149).

References

[1] H. Hamada, S. Miki, and R. Nakatsu, "Automatic evaluation of English pronunciation based on speech recognition techniques," *IEICE—Transactions on Info and Systems*, vol. E76-D, no. 3, pp. 352–359, 1993.

[2] K. Truong, *Automatic Pronunciation Error Detection in Dutch as a Second Language: An Acoustic-Phonetic Approach*, Utrecht University, Utrecht, Netherlands, 2014.

[3] B. Dong, Q. Zhao, and Y. Yan, "Automatic scoring of flat tongue and raised tongue in computer-assisted Mandarin learning," in *Proceedings of the Proceedings of ISCSLP*, Tianjin, China, October 2016.

[4] S. M. Witt and J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 4, pp. 5–108, 2000.

[5] H. Chao, Z. Feng, F. K. Soong, M. Chu, and R. Wang, "Automatic mispronunciation detection for Mandarin," in *Proceedings of the IEEE International Conference on Acoustics*, Las Vegas, NV, USA, April 2008.

[6] Y. B. Wang and L. S. Lee, “Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training,” in *Proceedings of the Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, Kyoto, Japan, March 2012.

[7] A. Neri, C. Cucchiarini, and H. Strik, “ASR-based corrective feedback on pronunciation: does it really work?” in *Proceedings of the International Conference on Interspeechicslp*, San Francisco, CA, USA, September 2016.

[8] Y. Ishida and S. Hashimoto, “Asymmetric characterization of diversity in symmetric stable marriage problems: an example of agent evacuation,” *Procedia Computer Science*, vol. 60, no. 1, pp. 1472–1481, 2015.

[9] P. Zoha and R. Kaushik, “Image edge detection based on swarm intelligence using memristive networks,” *IEEE Trans. on CAD of Integrated Circuits and Systems*, vol. 37, no. 9, pp. 1774–1787, 2018.

[10] W. Li, S. M. Siniscalchi, N. F. Chen, and C.-H. Lee, “Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling,” in *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, March 2016.

[11] X. Qian, H. Meng, and F. Soong, “The use of DBNHMMs for mispronunciation detection and diagnosis in L2 English to support computer-aided pronunciation training,” in *Proceedings of the proc interspeech*, Brno, Czechia, September 2021.

[12] K. Li, X. Qian, and H. Meng, “Mispronunciation detection and diagnosis in L2 English speech using multidistribution deep neural networks,” *IEEE ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, pp. 193–207, 2016.

[13] A. Lee, Y. Zhang, and J. Glass, “Mispronunciation detection via dynamic time warping on deep belief network-based posteriorgrams,” in *Proceedings of the IEEE International Conference on Acoustics*, Vancouver, BC, Canada, May 2013.

[14] Y. Hua, J. Zhao, and L. Jia, “Improve mispronunciation detection with Tandem feature,” in *Proceedings of the International Symposium on Chinese Spoken Language Processing*, Hongkong, China, January 2020.

[15] J. Pais, “Random matching in the college admissions problem,” *Economic Theory*, vol. 35, no. 1, pp. 99–116, 2018.

[16] J. J. Jung and G. S. Jo, “Brokerage between buyer and seller agents using constraint satisfaction problem models,” *Decision Support Systems*, vol. 28, no. 4, pp. 291–384, 2020.

[17] Y. Liu and K. W. Li, “A two-sided matching decision method for supply and demand of technological knowledge,” *Journal of Knowledge Management*, vol. 21, no. 3, p. 0183, 2017.

[18] J. Byun and S. Jang, “Effective destination advertising: matching effect between advertising language and destination type,” *Tourism Management*, vol. 50, no. 10, pp. 31–40, 2015.

[19] A. N. Nagamani, S. N. Anuktha, N. Nanditha, and V. K. Agrawal, “A genetic algorithm-based heuristic method for test set generation in reversible circuits,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 2, pp. 324–336, 2018.

[20] C. Koch and S. P. Penczynski, “The winner’s curse: conditional reasoning and belief formation,” *Journal of Economic Theory*, vol. 174, pp. 57–102, 2018.

[21] C. K. Karl, “Investigating the winner’s curse based on decision making in an auction environment,” *Simulation & Gaming*, vol. 47, no. 3, pp. 324–345, 2016.

[22] D. Ettinger and F. Michelucci, “Creating a winner’s curse via jump bids,” *Review of Economic Design*, vol. 20, no. 3, pp. 173–186, 2016.

[23] J. A. Brander and E. J. Egan, “The winner’s curse in acquisitions of privately-held firms,” *The Quarterly Review of Economics and Finance*, vol. 65, pp. 249–262, 2017.

[24] Z. Palmowski, “A note on var for the winner’s curse,” *Economics/Ekonomia*, vol. 15, no. 3, pp. 124–134, 2017.

[25] B. R. Routledge and S. E. Zin, “Model uncertainty and liquidity,” *Review of Economic Dynamics*, vol. 12, no. 4, pp. 543–566, 2009.

[26] D. Easley and M. O’Hara, “Ambiguity and nonparticipation: the role of regulation,” *Review of Financial Studies*, vol. 22, no. 5, pp. 1817–1843, 2019.

[27] P. Klibano, M. Marinacci, and S. Mukerji, “A smooth model of decision making under ambiguity,” *Econometrica*, vol. 73, no. 6, pp. 1849–1892, 2005.

[28] Y. Halevy, “Ellsberg revisited: an experimental study,” *Econometrica*, vol. 75, no. 2, pp. 503–536, 2017.

[29] D. Ahn, S. Choi, and D. Gale, “Estimating ambiguity aversion in a portfolio choice experiment,” *Working paper*, vol. 5, no. 2, pp. 195–223, 2019.

[30] T. Hayashi and R. Wada, “Choice with imprecise information: an experimental approach,” *Theory and Decision*, vol. 69, no. 3, pp. 355–373, 2010.