Hindawi
Wireless Communications and Mobile Computing
Volume 2023, Article ID 9760391, 1 page
https://doi.org/10.1155/2023/9760391

Retraction

Retracted: Facial Expression Recognition Based on Convolutional Neural Network Fusion SIFT Features of Mobile Virtual Reality

Wireless Communications and Mobile Computing

Received 28 November 2023; Accepted 28 November 2023; Published 29 November 2023

This article has been retracted by Hindawi, as publisher, following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of systematic manipulation of the publication and peer-review process. We cannot, therefore, vouch for the reliability or integrity of this article.

Please note that this notice is intended solely to alert readers that the peer-review process of this article has been compromised.

Wiley and Hindawi regret that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] F. Yao and L. Qiu, "Facial Expression Recognition Based on Convolutional Neural Network Fusion SIFT Features of Mobile Virtual Reality," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 5763626, 7 pages, 2021.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 5763626, 7 pages
https://doi.org/10.1155/2021/5763626

Research Article

Facial Expression Recognition Based on Convolutional Neural Network Fusion SIFT Features of Mobile Virtual Reality

Fuguang Yao[1] **and Liudong Qiu**[2]

[1]*Chongqing University of Education, Chongqing 400065, China*
[2]*Chongqing Industry Polytechnic College, Chongqing 401120, China*

Correspondence should be addressed to Fuguang Yao; yaofuguang@cque.edu.cn

Received 22 August 2021; Accepted 25 September 2021; Published 14 October 2021

Academic Editor: Balakrishnan Nagaraj

Facial expression recognition computer technology can obtain the emotional information of the person through the expression of the person to judge the state and intention of the person. The article proposes a hybrid model that combines a convolutional neural network (CNN) and dense SIFT features. This model is used for facial expression recognition. First, the article builds a CNN model and learns the local features of the eyes, eyebrows, and mouth. Then, the article features are sent to the support vector machine (SVM) multiclassifier to obtain the posterior probabilities of various features. Finally, the output result of the model is decided and fused to obtain the final recognition result. The experimental results show that the improved convolutional neural network structure ER2013 and CK+ data sets' facial expression recognition rate increases by 0.06% and 2.25%, respectively.

1. Introduction

Facial expression recognition computer technology can obtain the emotional information of the person through the expression of the person to judge the state and intention of the person. It is of great significance in human-computer interaction, safe driving, and intelligent advertising systems. The CK+ data set is a classic facial expression library, which contains expression images of anger, disgust, fear, happiness, sadness, surprise, and contempt. The expressions are video sequences [1]. It contains a series of images with the same expression ranging from calm to violent. We can extract neutral expression images from it.

Affected by factors such as distance, the image will have the problem of blurred faces and fewer face pixels. The facial expression recognition of low-pixel facial images is to recognize facial images with low quality and inconspicuous facial features [2]. The image size obtained by sampling is 32 pixel × 32 pixel, which is in line with the low-pixel characteristics. The complexity of facial expression images is high. When the facial features are not obvious, it is difficult for us to identify by extracting specific feature information. For facial expression images with a size of 32 pixel × 32 pixel, some scholars have proposed a facial expression recognition method based on the improved LeNetG5 convolutional neural network (CNN). Some scholars have proposed a CNN facial expression recognition method based on the local binary pattern (LBP). Research shows that CNN has a better effect on facial expression recognition in low-pixel facial images. This paper improves the CNN model on this basis. We propose an expression recognition method for low-pixel facial images and compare it with several other methods. The results show that this method has a better recognition effect.

2. Expression Recognition Method for Low-Pixel Facial Images

2.1. Facial Expression Image Reprocessing. General expression recognition methods include image reprocessing, facial feature extraction, and expression recognition. The reprocessing stage of image expression recognition performs face detection to obtain facial region images [3]. The recognition of expressions in low-pixel facial images also requires image

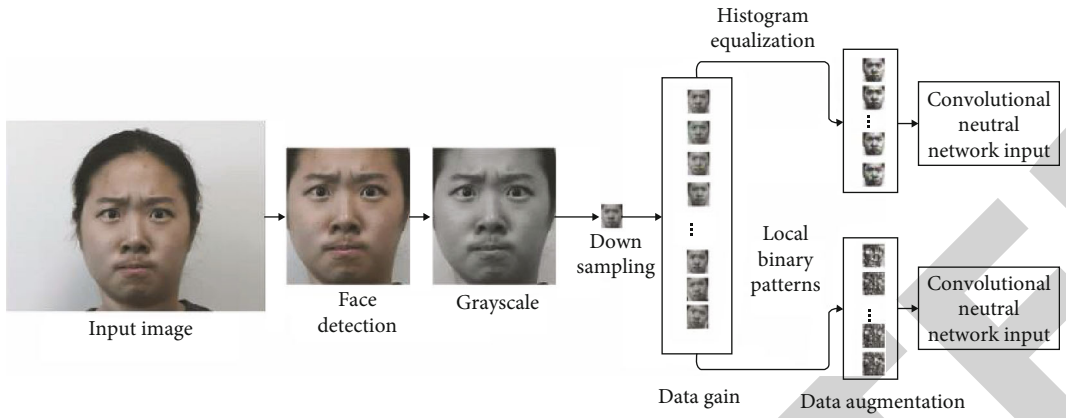

Figure 1: Facial expression reprocessing process.

enhancement or image superresolution during reprocessing. Image enhancement is to enhance the existing information of the image by changing the distribution of pixels, and image superresolution are to restore some missing pixel information by adding pixels.

The image reprocessing of this method includes face detection and cropping, gray processing, downsampling, data enhancement, and image enhancement. The purpose of face detection is to accurately calibrate the position and size of the face in the image. We use the D-lib model for face detection. The D-lib model can automatically estimate the coordinates of the facial feature points in the image and process the data in the OpenCV library. We use this to crop the image so that the image features are concentrated on the face. Grayscale processing is the process of converting a colour image into a gray image. Downsampling is to standardize the image size in the input CNN model. We use bilinear interpolation to ensure that the face position of the resampled image is the same as the original image. We use CNN for image recognition. The amount of training data directly affects the final recognition effect. The larger the amount of data, the better the effect [4]. Commonly used data enhancement methods include mirroring, rotating, and adding noise. These methods mirror the original data and rotate it in different angles and directions, enhancing the data to 13 times the original data. We then add different noise coefficients (salt and pepper noise, Gaussian noise, Poisson noise, and speckle noise) to the existing data, and the final data is enhanced to 130 times the original data. We perform histogram equalization on the image and use the local binary mode to obtain the enhanced image. Among them, histogram equalization is also called histogram flattening. The essence of this method is to stretch the image nonlinearly and redistribute the image pixel values. In this way, the number of pixel values in a certain grayscale range is roughly equal. The local binary mode is an operator that describes the local texture characteristics of the image. It has the advantages of rotation and gray invariance. It can be used to extract local texture features of the image. The specific reprocessing process is shown in Figure 1.

2.2. Improved Convolutional Neural Network Model. With the development of computer processing capabilities, CNN has achieved amazing results in image recognition. The efficiency of CNN-based image recognition methods has also been continuously improved and has gradually replaced the traditional facial expression image recognition methods. In the 2012 ImageNet Object-Oriented Recognition Challenge (ILSVRC), some scholars used the CNN model Alex Net to win the championship. In the 2014 ILSVRC competition, the CNN model Google Net architecture won first place in the classification. Some scholars proposed CNN-VGG Net. The second place in the classification group and the first place in the positioning project group. VGG-Net deepens the number of network layers while avoiding too many parameters, all layers use a 3×3 small convolution kernel, and the convolution layer step size is set to 1. The alternating structure of multiple convolutional and nonlinear activation layers makes extracting deeper and better features than a single convolutional layer structure. In the ILSVRC2015 competition, ResNet, proposed by scholars, won the championship [5]. A connection method called short cut connection in ResNet can theoretically keep the network in an optimal state while the network layer is constantly deepening. There is enough feature information in facial expressions to optimize model parameters to obtain a good recognition effect. Low-pixel face images need to make full use of inconspicuous feature information. On this basis, the article proposes a CNN model for expression recognition of low-pixel facial images to extract facial features better.

The image size of the input CNN model is 32 pixel $\times$ 32 pixel. We increase the number of CNN layers to increase the nonlinearity of the network model. This makes the recognition ability of the decision function stronger. To avoid gradient disappearance and gradient explosion caused by deepening the number of network layers, the network needs to have a more complex structure [6]. Some scholars have proposed a network connection structure high way networks. Some of the features in this structure can pass through certain network layers directly without processing, which makes the structure easier to optimize. Combining this structure and the short cut connection, a short method

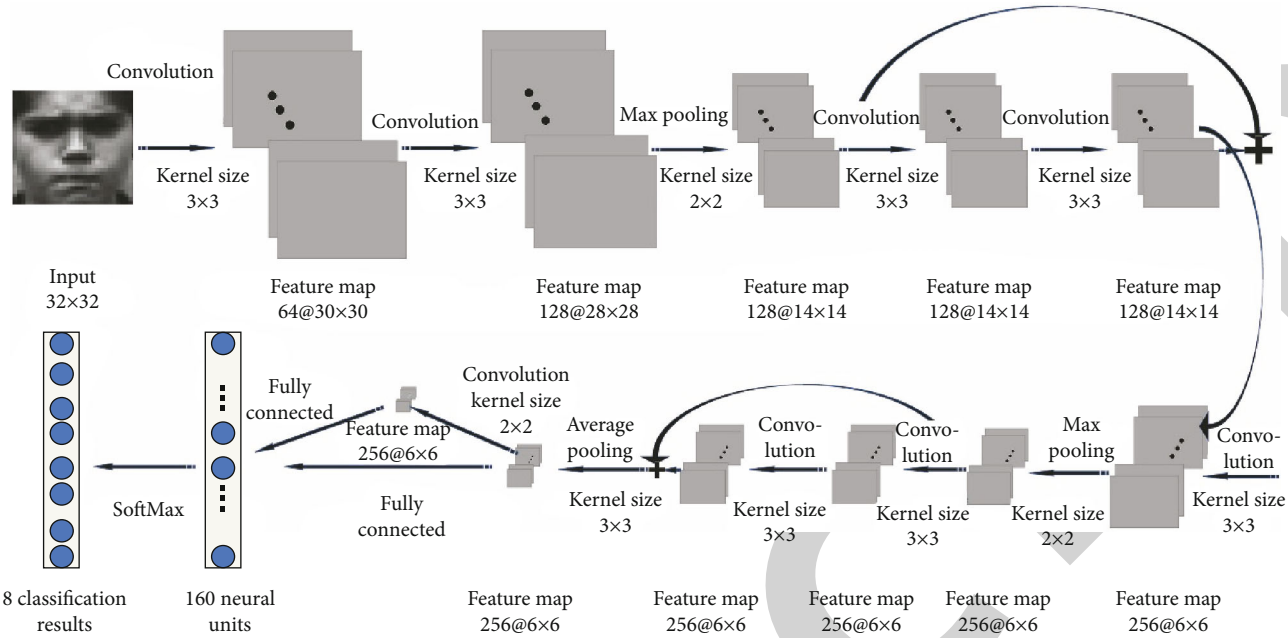

Figure 2: CNN model.

avoids the gradient disappearance and gradient explosion problems when using deeper networks. The experiment the CNN model used is shown in Figure 2. The number before @ is the number of feature maps, and the number after @ is the size of the feature map (pixel × pixel).

We input the size of the feature map, the size of the kernel of the convolution and pooling operation, the step size, and the size of the output feature map in the convolution operation and the pooling operation. Its mathematical relationship is

$$\omega_2 = \frac{\omega_1 - f + 2p}{s + 1}, \tag{1}$$

where ω_2 is the feature map size after convolution or pooling operation, ω_1 is the size of the feature map before convolution or pooling, f is the size of the kernel for convolution and pooling operations, p is the number of pixels filled with zero, and s is the step size. We added the output tensor of the third layer of the model and the output tensor of the fifth layer to obtain 128 feature maps with a size of 14 pixel × 14 pixel. Then, we pass the ReLU activation function as the input of the sixth layer. The article adds the output tensors of the seventh and ninth layers of the model to obtain 256 feature maps with a size of 6 pixel × 6 pixel. Then, we pass the ReLU activation function as the input of the tenth layer. The twelfth layer is fully connected. We take the output of the tenth and eleventh layers through the ReLU activation function and then concatenate the obtained tensors as the input of the twelfth layer. The output is 160 neurons [7]. The last layer is the SoftMax classifier. The output is eight network nodes, representing the probability that the input image is indifferent expression states. Table 1 is the specific description of the model. The content includes the type of each layer of the model, the corresponding kernel size and step size, and the size of the output feature map of each layer. The CNN includes three basic operations: convolution, pooling, and full connection. Among them, convolution is also divided into inner convolution and outer convolution. In other words, it is the convolution without 0 paddings and the convolution with 0 paddings. The article assumes that the input is the matrix A of $M \times N$. The convolution kernel is moment B, and $M \geq m, N \geq n$ of $m \times n$; then, the output of the inner convolution operation is $C = A * B$. The pixel c_{ij} at the corresponding position can be expressed as

$$c_{ij} = \sum_{s=1}^{m} \sum_{t=1}^{n} a_{i+m-s,j+n-t} \times b_{st}. \tag{2}$$

$1 \leq i \leq M - m + 1, 1 \leq j \leq N - n + 1$ is the corresponding multiplication with the rows and columns of matrix A. Suppose the pixel at the corresponding position of matrix B is b_{st}, and the pixel at the corresponding position of matrix A is $a_{i+m-s,j+n-t}$. Outer convolution is defined as filling A with 0, and the rows and columns of the filled matrix are related to the number of rows and columns of the B matrix. The article makes it an $(M + 2m - 2) \times (N + 2n - 2)$ matrix and then performs inner convolution with B. The formula can be expressed as

$$A \overset{\cap}{*} B = \overset{\cap}{A_B} \overset{\cup}{*} B. \tag{3}$$

We pool matrix A. Suppose it is divided into nonoverlapping blocks, and the size of each block is $\lambda \times \tau$. The matrix $G^A_{\lambda,\tau} ij$ of block ij can be expressed as

$$G^A_{\lambda,\tau}(i, j) = (a_{st})_{\lambda \times \tau}, \tag{4}$$

Table 1: Convolutional network parameters.

Layer	Type	Kernel size (pixel × pixel)	Stride	Output number	Output size (pixel × pixel)
Layer1	Convolution	3 × 3	1	64	30 × 30
Layer2	Convolution	3 × 3	1	128	28 × 28
Layer3	Max pooling	2 × 2	2	128	14 × 14
Layer4	Convolution	3 × 3	1	128	14 × 14
Layer5	Convolution	3 × 3	1	128	14 × 14
Layer6	Convolution	3 × 3	1	256	12 × 12
Layer7	Max pooling	2 × 2	2	256	6 × 6
Layer8	Convolution	3 × 3	1	256	6 × 6
Layer9	Convolution	3 × 3	1	256	6 × 6
Layer10	Average pooling	3 × 3	3	256	2 × 2
Layer11	Convolution	2 × 2	1	512	1 × 1
Layer12	Fully connected	—	—	—	160 × 1
Output	SoftMax	—	—	—	8 × 1

where a_{st} is the element in row s and column t in matrix A, $(a_{st})_{\lambda\times\tau}$ is a matrix block composed of $\lambda \times \tau$ elements in matrix A, and $(i-1)\times\lambda+1 \le s \le i\times\lambda$, $(j-1)\times\tau+1 \le t \le j \times \tau$. Maximum pooling is defined as

$$d\left[G^A_{\lambda,\tau}(i,j)\right] = \max\left(\sum_{s=(i-1)\times\lambda+1}^{i\times\lambda}\sum_{t=(j-1)\times\tau+1}^{j\times\tau} a_{st}\right). \quad (5)$$

Average pooling is defined as

$$a\left[G^A_{\lambda,\tau}(i,j)\right] = \frac{1}{\lambda\times\tau}\sum\left(\sum_{s=(i-1)\times\lambda+1}^{i\times\lambda}\sum_{t=(j-1)\times\tau+1}^{j\times\tau} a_{st}\right). \quad (6)$$

We use overlapping blocks of size $\lambda \times \tau$ to downsample the maximum pooled $A_{\max}$ and the average pooled A_{mean}, respectively. The formula can be expressed as

$$\begin{aligned} D_{\lambda,\tau}(A_{\max}) &= d\left[G^A_{\lambda,\tau}(i,j)\right], \\ D_{\lambda,\tau}(A_{\text{mean}}) &= a\left[G^A_{\lambda,\tau}(i,j)\right]. \end{aligned} \quad (7)$$

Each output A of the fully connected layer can be seen as the r node a_r in the previous layer multiplied by its weight coefficient ω_r, plus a bias value b_h. For example, the input of the fully connected layer is $256\times2\times2$ nodes. That is, the input feature map is 256@2 × 2, and the output has 80 nodes. A total of $256\times2\times2\times80 = 81920$ weight coefficients and 80 offset parameters are required. Then, a single element d_h in its output vector D can be expressed as

$$d_h = \sum_{r=1}^{k}\omega_r \times a_r + b_h. \quad (8)$$

In the formula, k is the number of input feature maps.

3. Expression Recognition Experiment of Low-Pixel Facial Images

3.1. Data Set Preparation. The experiment uses the CK+ data set. This data set is used to evaluate the facial expression recognition (FER) system, and it is also a relatively common data set for facial expression recognition. The content contains 593 video sequences from 123 subjects. The duration ranges from 10 to 60 frames [8]. The data set shows a series of images ranging from calm to violent. The number of original images on different expressions is unevenly distributed. The neutral expression image is the image at the beginning or end of the expression. According to the original number distribution, we select the last 1 ~ 3 expression images of each expression sequence. A total of 686 images were used for modeling. 80% is used as the training set, and 20% is used as the test set. The peak images of the same person with the same expression will not appear in the training set and the test set simultaneously.

The article adds data to the training set. We will test the data gained and ungained test sets in the same trained model. Since the research found that the difference in recognition accuracy is small, no data gain processing is done on the test set. The final image size of the training set is 71370, and the image size of the test set is 137. We perform histogram equalization and local binary mode on all images to obtain three data sets of the same size, including the original image [9]. Table 2 shows the number of images of 8 expressions in each test set and training set. The original image of the eight expressions, the image after the histogram equalization, and the image with the local binary mode are shown in Figure 3.

3.2. Evaluation Criteria. The main evaluation criteria of facial expression recognition methods are recognition accuracy and recognition speed. The recognition accuracy rate is the ratio of correctly recognized expression samples in the test set to the number of samples in all the test sets.

Table 2: Facial expression database.

Category	Train database	Test database
Angry	14040	27
Disgust	12220	24
Fearful	7800	15
Happy	14300	28
Sad	8970	15
Surprised	8580	16
Scorn	5460	12
Neutral	8970	15

Recognition speed is when it takes to recognize each test set sample after the recognition model is established. The time it takes is the ratio of the time it takes to identify the test set to the number of samples in the test set.

$$A = \frac{1}{p}\sum_{1}^{p} g[f(x_b) = y_b],$$
$$t_d = \frac{T}{p}, \tag{9}$$

where A is the recognition accuracy rate, p is the total number of samples in the test set, g is the indicator function, x_b is the given sample, $f(x_b)$ is the output after the sample passes the model, y_b is the label of a given sample, t_d is the recognition speed [10], and T is the total time spent. We can get it by subtracting the time before the first test sample was recognized by the time after the last test sample was recognized.

3.3. Experimental Process. Because the input CNN image pixels are low, the output recognition effect will fluctuate slightly, so we introduce the decision fusion and final image recognition. In the testing phase, we use five trained network models to judge the test set data, respectively. Then, we use the SoftMax average voting (SAV) method to fuse the judgment results of the five models. Finally, we get the final result to improve the recognition effect. The test steps are shown in Figure 4.

It can be seen from Section 2.2 that the output of CNN is a 1-dimensional vector. The value of each element in the vector is the probability that the image may be a certain category. SoftMax average voting is to average the output results of five trained CNNs. Take the average of three experiments from the most likely result at the end of the article. The graphics card is NVIDAGe Force 940MX. The main frequency is 1122 MHz, and the memory is 2.00 GB. The operating system is Linux Ubuntu 16.04. The software is Python 3.6, NVIDIA CUDA, and cuDNN libraries. We adopted the training strategy to improve the recognition accuracy is to add batch normalization (BN) and ReLU activation functions after each convolutional and pooling layer. This can overcome the disappearance of the gradient and speed up the training speed.

We selectively add L2 regularization and dropout to alleviate overfitting. The learning rate decay strategy is adopted. We choose a larger value at the beginning of the learning rate. After N rounds of iteration, the attenuation is 1/10 of the initial learning rate [11]. We use the Adam optimization algorithm during optimization, which enables the network to find the global best advantage faster. The data sets are the original image, the image with the histogram enhancement, and the local two. The CNN model is trained when the value pattern feature map is used, and at the same time, we adjust the parameters of the network.

According to the accuracy of the test, we first determine whether we need to add the L2 regularization and dropout layer and then determine which layer to place. We determine the approximate range of the learning rate according to the loss during training, training loss, and test accuracy. Two dichotomies obtain the learning rate.

3.4. Result Analysis. To find the optimal situation of the facial expression recognition system, we input the original image data set, the local binary pattern feature map data set, and the histogram equalized data set in the CNN model. The average recognition accuracy and speed obtained as a result are shown in Table 3. The experiment is the average of three experiments.

It can be found that the accuracy of the input data set after histogram equalization is better than that of the original image data set. The accuracy of the data set the input as the local binary pattern feature map is the worst. There is no obvious difference in speed between the three. The recognition speed of the data set the input as the local binary pattern feature map is slightly faster, but the speed of 0.29 s has a small advantage compared with the accuracy of 3.6%. Analysing the different data sets in Figure 3, we believe that certain data enhancements will strengthen the image information and improve the recognition accuracy. Because the original pixels of the data are too low, we directly extract the feature map of its local binary mode, which will enhance the texture information while losing more information [12]. Therefore, the speed is the fastest when the accuracy is the lowest. The experiment finally selects the data set enhanced by histogram equalization to input the CNN model. We add L2 regularization and dropout after the twelfth layer. The dropout parameter is 0.7, and the initial value of the learning rate is 0.0001. After 1000 iterations, it decays to 1/10 of the initial learning rate. Table 4 shows the three experimental results of the improved CNN model on the CK+ data set for the eight-expression recognition.

It can be seen from Table 4 that the recognition rates of happy and surprised expressions are higher. Fearful's recognition rate is low and fluctuates greatly. On the one hand, it may be that the first two characteristics are more distinct, while fearful and sad have similar characteristics to some extent. On the other hand, in the CK+ data set, the amount of raw data of the first two expressions is abundant, and the amount of fearful data is small. A total of 15 images were in the fearful test set. There are only five different expression images in the original data. This leads to unequal training times. To verify the effect of the established recognition

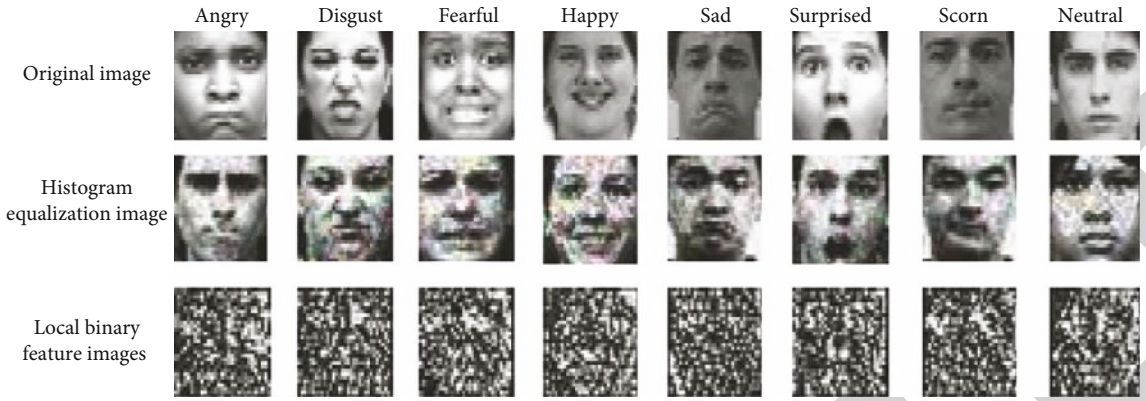

Figure 3: Sample images of 8 emoticons.

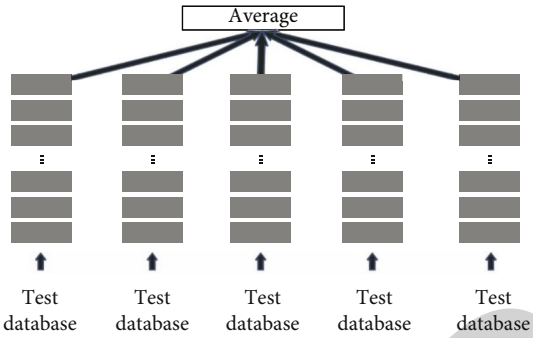

Figure 4: Schematic diagram of SoftMax average voting.

Table 3: Comparison of recognition accuracy and time.

Model	Original database	Local binary pattern database	Histogram equalization database
Accuracy (%)	89.5	86.9	90.5
Time (s)	1.54	1.27	1.56

Table 4: The recognition accuracy of the improved CNN model.

Category accuracy	First experiment	Second experiment	Third experiment	Total
Angry	96.3	88.9	96.3	93.8
Disgust	91.7	91.7	95.8	93.1
Fearful	60	86.7	60	68.9
Happy	100	100	100	100
Sad	86.7	80	86.7	84.5
Surprised	100	100	93.8	97.9
Scorn	83.3	83.3	75	80.5
Neutral	93.3	93.3	93.3	93.3
Total	90.5	91.2	89.8	90.5

model, we input the JAFFE data set as the test set to identify other expressions except for the contempt expression. The average recognition accuracy of the three results is 82.4%.

Table 5: Comparison of recognition accuracy and time between the two models.

Model	LeNet-5	Without decision level
Accuracy (%)	74.6	87.9
Time (s)	0.59	0.31

To prove that the proposed method and decision fusion method are pertinent to the expression recognition of low-pixel facial images, we conduct experiments in two cases, respectively [13]. One is to replace the improved CNN model with the classic shallow convolutional neural network LeNetG5. The second is to experiment without using decision fusion. Table 5 shows the average recognition accuracy and comparison of the three experiments in the two cases above.

It can be seen from Table 3 that the improved CNN model has an increase of 15.9% in recognition accuracy compared with the LeNetG5 network. This proves that this method is more suitable for expression recognition of low-pixel face images [14]. The recognition accuracy after decision fusion is 2.6 percentage points higher than that of the network without decision fusion. The main reason is that the experimental effects in the three experiments are unstable, and the recognition accuracy of the two experiments is about 90.0%. The result of one experiment was 83.9%. However, each experiment result of this method is obtained by averaging five trained network models, and its overall stability is relatively high. This proves that this method is effective and feasible in practice.

In recent years, facial expression recognition methods for face images with a size of 32 pixel × 32 pixel have been proposed one after another. On the CK+ data set, some scholars have proposed a cross-connect LeNetG5 CNN. We perform seven classifications of images that do not include neutral expressions, and the recognition accuracy rate is 83.74%. Some scholars have proposed a shallow CNN to achieve a seven-class recognition accuracy of 97.38%. This is higher than the recognition accuracy of this method.

4. Conclusion

Aiming at the expression recognition of low-pixel face images, the paper proposes an improved CNN expression

recognition method. The article increases the nonlinearity of the network model by adding a convolutional layer. We can learn from extract image features in more layers and reflect image information.

Data Availability

The data 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 work was supported by the Scientific Research Platform of Chongqing University of Education, Big Data practice platform for sports and health based on Campus IOT, No. 2017XJPT07; 'Future School (infant education)' of National Center for Schooling Development Programme of China, Research on the Construction of Intelligent Kindergarten Based on IOT, No. CSDP18FC3204; and Chongqing Key Research Base of Humanities and Social Sciences 'Chongqing Research Center of Overall Development of Urban-Rural Teachers Education,' No. 18JDZDWT04.

References

[1] F. Kong, "Facial expression recognition method based on deep convolutional neural network combined with improved LBP features," *Personal and Ubiquitous Computing*, vol. 23, no. 3-4, pp. 531–539, 2019.

[2] J. Saeed and A. M. Abdulazeez, "Facial beauty prediction and analysis based on deep convolutional neural network: a review," *Journal of Soft Computing and Data Mining*, vol. 2, no. 1, pp. 1–12, 2021.

[3] X. Zhou, "Video expression recognition method based on spatiotemporal recurrent neural network and feature fusion," *Journal of Information Processing Systems*, vol. 17, no. 2, pp. 337–351, 2021.

[4] X. Liu, Y. Xia, H. Yu, J. Dong, M. Jian, and T. D. Pham, "Region based parallel hierarchy convolutional neural network for automatic facial nerve paralysis evaluation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 10, pp. 2325–2332, 2020.

[5] G. Yolcu, I. Oztel, S. Kazan et al., "Facial expression recognition for monitoring neurological disorders based on convolutional neural network," *Multimedia Tools and Applications*, vol. 78, no. 22, pp. 31581–31603, 2019.

[6] M. Z. Lifkooee, Ö. M. Soysal, and K. Sekeroglu, "Video mining for facial action unit classification using statistical spatial–temporal feature image and LoG deep convolutional neural network," *Machine Vision and Applications*, vol. 30, no. 1, pp. 41–57, 2019.

[7] J. K. Park and D. J. Kang, "Unified convolutional neural network for direct facial keypoints detection," *The Visual Computer*, vol. 35, no. 11, pp. 1615–1626, 2019.

[8] K. S. Yoon and J. Y. Choi, "Compressed ensemble of deep convolutional neural networks with global and local facial features for improved face recognition," *Journal of Korea Multimedia Society*, vol. 23, no. 8, pp. 1019–1029, 2020.

[9] X. Pan, "Fusing HOG and convolutional neural network spatial–temporal features for video-based facial expression recognition," *IET Image Processing*, vol. 14, no. 1, pp. 176–182, 2020.

[10] N. Mehendale, "Facial emotion recognition using convolutional neural networks (FERC)," *SN Applied Sciences*, vol. 2, no. 3, pp. 1–8, 2020.

[11] H. Adachi, K. Oiwa, and A. Nozawa, "Drowsiness level modeling based on facial skin temperature distribution using a convolutional neural network," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 14, no. 6, pp. 870–876, 2019.

[12] P. Barros, N. Churamani, and A. Sciutti, "The facechannel: a fast and furious deep neural network for facial expression recognition," *SN Computer Science*, vol. 1, no. 6, pp. 1–10, 2020.

[13] H. Liao, G. Wen, Y. Hu, and C. Wang, "Convolutional herbal prescription building method from multi-scale facial features," *Multimedia Tools and Applications*, vol. 78, no. 24, pp. 35665–35688, 2019.

[14] A. Satapathy and L. J. Livingston, "A lite convolutional neural network built on permuted Xceptio-inception and Xceptio-reduction modules for texture based facial liveness recognition," *Multimedia Tools and Applications*, vol. 80, no. 7, pp. 10441–10472, 2021.