The Visual Computer (2020) 36:483–498
https://doi.org/10.1007/s00371-019-01635-4

ORIGINAL ARTICLE

Facial expression recognition algorithm based on parameter adaptive initialization of CNN and LSTM

Fengping An[1,2] · Zhiwen Liu[2]

Published online: 31 January 2019

Abstract

In view of the high dimensionality, nonrigidity, multiscale variation and the influence of illumination and angle on facial expressions, it is quite difficult to obtain facial expression images or videos using computers and analyze facial morphology and changes to accurately obtain the emotional changes of the subjects. Existing facial expression recognition algorithms have the following problems in the application process: the existing shallow feature extraction model has lost a lot of effective feature information and low recognition accuracy. The facial expression recognition method based on deep learning has problems such as overfitting, gradient explosion and parameter initialization. Therefore, this paper develops a facial expression recognition algorithm based on the deep learning method. An adaptive model parameter initialization based on the multilayer maxout network linear activation function is proposed to initialize the convolutional neural network (CNN) and the long–short-term memory network (LSTM) method. It can effectively overcome the gradient disappearance and gradient explosion problems in the deep learning model training process. At the same time, the convolutional neural network with an LSTM memory unit is used to extract the related information from the image sequence, and the facial expression judgment is based on a single-frame image and historical-related information. However, the top-level structure of the CNN model is a fully connected feedforward neural network, which undertakes the task of expression classification. Therefore, the SVM classification method replaces the top-level classifier to further improve the expression classification accuracy. Experiments show that the facial expression recognition method proposed in this paper not only accurately identifies various expressions but also has good adaptive ability. This is because the method achieves the adaptive initialization of the parameters of the deep learning model construction process and also analyzes the relevance of the expression database expression, thereby improving the accuracy of expression recognition.

Keywords Model parameter initialization method · CNN · Facial expression recognition · Deep learning · LSTM · SVM

1 Introduction

With the rapid development of the economy and technology, the human society of the future will be a society with extremely intelligent characteristics, and the breadth and depth of intelligence will exceed our imagination. As a result, companies in various countries continue to develop new technologies and introduce evermore intelligent products [1–3]. For example, in 2016, Google Machine Translation used long–short-term memory recurrent neural networks (LSTM-RNNs) to realize the memory and reasoning of language information, the translation became smoother and its level has become close to the artificial professional translation standard [4]. In the future development of intelligent society, how to make machines understand, express and exchange emotions like human beings, thus realizing the intelligence and comfort of human–computer interaction, will be paramount. Facial expression recognition technology can provide a reliable technical image for solving such problems, because the facial expression is relatively easy to obtain and can truly reflect the facial information. The reason for solving this problem through the facial expression

✉ Fengping An
anfengping@163.com; anfengping1985@163.com;
anfengping@bit.edu.cn

Zhiwen Liu
liuzhiwen1961@163.com

[1] School of Physics and Electronic Electrical Engineering, Huaiyin Normal University, Huai'an 223300, China

[2] School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China

recognition method is mainly based on the following three reasons: first, facial expression images convey emotional information relatively realistically, providing important information for discerning emotions [5]; second, there is a progress of facial expression classification research; and third, the machine can understand human emotions in real time. It is an important research topic to enable machines to recognize and understand human emotions and thus effectively and accurately perceive human emotions, expanding machine intelligence functions to better serve human beings. Currently, many researchers in the academic field have worked in many application fields and achieved remarkable results. For example, AKA's Musio intelligent robot can communicate with humans not only in understandable language but also through the emotions of the communication objects. The facial expression software developed by Microsoft Corporation can automatically analyze the facial expression of each person in a photograph through image processing and accurately identify and judge facial expressions such as "surprise" and "sadness".

Facial expression recognition technology is mainly divided into two categories: one is the traditional facial expression recognition method and the other is the facial expression recognition method based on deep learning. The traditional facial expression recognition methods mainly include the following: (1) the feature extraction method based on local texture. It includes (1) the Gabor transform method. In 2005, Deng et al. [6] combined the Gabor feature method with the PCA method for facial expression recognition. However, the facial expression recognition method based on Gabor features is inferior in accuracy, especially for facial expression recognition such as rotation and blurring. (2) The local binary patterns (LBP) method. In 2009, Shan et al. [7] conducted a comprehensive experiment based on the expression recognition of LBP features to verify that LBP features have certain efficiency. However, LBP-based facial expression recognition methods also have problems such as low recognition accuracy and weak anti-interference. (3) The Haar method. In 2012, Satiyan et al. [8] used Haar wavelet features for facial description combined with multiscale analysis and statistical analysis for facial expression recognition. However, Haar-based facial expression recognition has the problem of a high false recognition rate and cannot extract facial expression information more completely. (2) Feature extraction method based on gradient features. It includes (1) the histogram of oriented gradient (HOG) method. In 2015, Chen et al. [9] first enhanced the HOG feature direction invariant feature for facial feature description and then used a multithread cascade classifier to train feature vectors to improve the overall recognition accuracy of the algorithm. However, the HOG-based expression recognition method has many problems, such as poor expression of facial expression information and poor recognition effect. (2) Scale-invariant feature transform (SIFT) method. In 2011, Soyel et al. [10] used the improved SIFT operator to describe the face pose and used SVD to extract the principal component information to realize expression recognition. However, the SIFT-based expression recognition method has problems such as low computational efficiency and being prone to dimensionality disasters. (3) Extraction method based on motion characteristics. The expression recognition algorithm based on this kind of method has many disadvantages such as a large number of computation requirements, excessive time consumption and poor real-time performance, so there is less research on expression recognition using this method [11–13]. Another type of facial expression recognition method was developed—a facial expression recognition method based on deep learning. In 2006, Hinton proposed using the layer-by-layer training method to solve the difficult problem of multilayer neural network training [14]; relevant scholars and research institutions invested considerable manpower and resources and developed powerful open-source learning tools such as Torch, Caffe, Deep Learn Toolbox and Cxxnet. Deep learning provides complex function implementations that can approximate high-dimensional data spaces, thus facilitating the high-dimensional feature learning of images. From the beginning, deep learning has been used for object image classification, face verification and ultimately, face recognition [15–17]. In 2014, Liu et al. [18] used a DBN deep network for feature learning combined with boosting for classification and achieved better results than traditional feature classification on a CK expression database. In 2015, Lopes et al. used convolutional neural networks for feature learning and data enhancement for image data. Experiments on the CK expression database showed that convolution feature learning is beneficial for expression classification. At the same time, the increase in the amount of data improves convolutional network performance [19]. In 2016, Zhang et al. proposed a new method based on a deep learning method, principal component analysis network and convolutional neural network for attitude-invariant facial expression recognition. A large number of experiments conducted on two public databases show that our method has significant improvement over other traditional expression recognition methods [20]. In 2017, Zhang et al. [21] proposed a facial expression extraction algorithm based on deep learning, analyzed the current research state and compared these methods. The research shows that the deep learning method can effectively extract the layered features and use the expressions to classify the images. The recognition accuracy is significantly improved compared with the traditional methods [11, 22–24]. However, the facial expression recognition method based on deep learning has problems such as overfitting, gradient explosion and parameter initialization of the deep learning model. At the same time, the existing algorithms still have problems such as lack of effective use of dynamic sequence information of expression images and poor robustness of the algorithm application.

In view of the problems existing in facial expression methods, the research in this paper focuses on the accuracy and robustness of expression recognition. An adaptive model parameter initialization based on the MMN linear activation function is proposed to initialize the CNN and the LSTM methods. The proposed parameter initialization based on the MMN linear activation function can effectively overcome the gradient disappearance and gradient explosion problems in the deep learning model training process and provides a more theoretical guarantee for effectively training the network model based on the MMN activation function. At the same time, the convolutional neural network with LSTM memory unit is used to extract the related information from the image sequence, and the facial expression judgment is based on the single-frame image and historical-related information. However, the training of the CNN deep learning model is not a process of convex optimization. The model parameter group easily falls into local extremum, and the fully connected layer containing most parameters of the network is difficult to optimize. At the same time, the top-level structure of the CNN model is a fully connected feedforward neural network, which undertakes the task of expression classification. Therefore, the SVM classification method replaces the top-level classifier to further improve the expression classification accuracy. Section 2 of this paper will introduce the parameter adaptive initialization method in the training process of the CNN model. Section 3 will elaborate on the facial expression recognition method based on adaptive parameter initialization CNN-LSTM. Section 4 analyzes the facial expression recognition algorithm proposed in this paper and compares it with the mainstream expression recognition algorithms. Finally, the full text is summarized and discussed.

2 Multilayer maxout activation function and parameter adaptive initialization algorithm

2.1 Multilayer maxout activation function

To increase the feature representation ability of the activation function and overcome the problems of the traditional activation function in the deep neural network model, this paper proposes a trainable multilayer maxout network activation function.

Assuming that the input $x \in R^d$ (x is the original input vector or the state vector of the previous hidden layer) is known, the calculation process of a neuron node activation function is as follows:

$$f_{i,j_1} = \max_{j_0 \in [1,k_0]} x^T W_{\dots i,j_0} + b_{i,j_0} \tag{1}$$

$$f_{i,j_n} = \max_{j_{(n-1)} \in [1,k_{(n-1)}]} f^T_{i,j_{(n-1)}} W_{\dots ij_{(n-1)}} + b_{ij_{(n-1)}}, n \in [2, N] \tag{2}$$

$$h_i = \max_{j_N \in [1,k_N]} f_{i,j_N} \tag{3}$$

where k_{n} is the number of neurons in the nth layer, N is the number of layers of the MMN activation function, and h_i is the output of the fth hidden layer node of the MMN activation function.

Since the maxout activation function is only a local linear function, only arbitrary convex functions can be fitted. The MMN activation function has the ability to fit arbitrary functions due to its multilayer structure, which overcomes the disadvantage that the maxout activation function cannot fit nonconvex functions. The MMN activation function has a multilayered structure compared to the maxout activation function. The difference between the MMN activation function and the maxout activation function is shown in Fig. 1.

The MMN activation function has powerful feature representation capabilities that can fit arbitrary functions, such as the usual ReLU activation functions. Even more complex nonlinear activation functions are needed to extract potential features, such as w-shaped nonconvex functions as shown in Fig. 2. It can be fitted through two layers of MMN activation functions, each layer requiring two linear function combinations to fit. When you need to fit a more complex function, you only need to increase the number of linear function combinations per layer of the MMN activation function.

The MMN activation function has the following characteristics.

1. The multilayer structure of the MMN activation function increases the nonlinearity of the neural network model so that the MMN activation function has the ability to fit arbitrarily complex potential distributions and improves the feature representation ability of the deep neural network model.
2. The piecewise linear function combination of each layer of the MMN activation function overcomes the shortcoming of the traditional activation function being easily saturated. It can prevent the gradient disappearance and gradient explosion problems occurring during the gradient propagation and accelerate the training convergence speed of the neural network model.
3. The activation function has a trainable feature that can be combined with the parameters of the neural network model for end-to-end joint optimization, which can better fit the potential distribution in the data. It can avoid

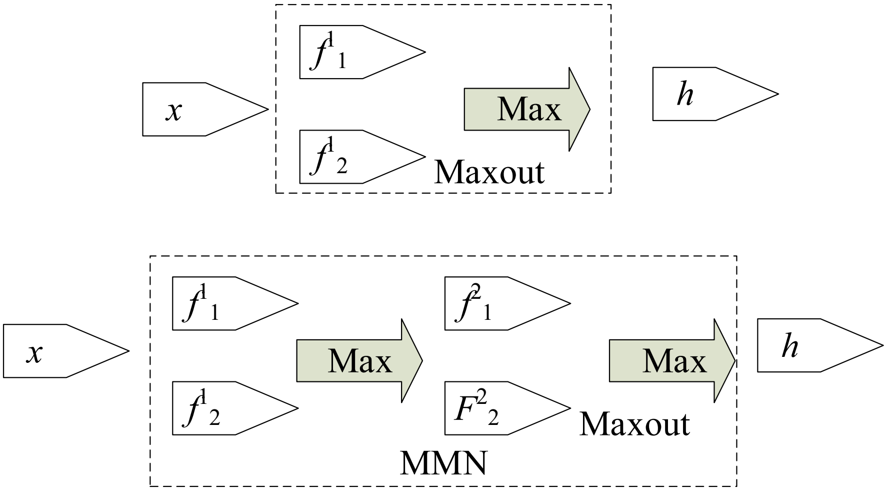


Fig. 1 Difference between MMN activation function and maxout activation function

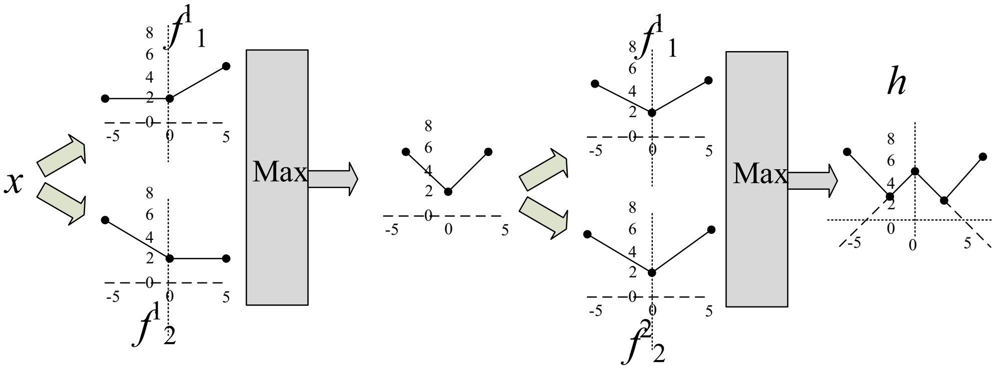


Fig. 2 MMN activation function fitting w-shaped function

the influence of improper selection of nonlinear activation functions on the performance of the model.

Although the MMN activation function has the above characteristics, the MMN activation function has its own parameters, which additionally add neural network model parameters. The model training process requires more computational overhead and storage overhead, which makes the training period longer.

2.2 Model parameter adaptive initialization method based on multilayer maxout activation function

The MMN activation function not only alleviates the gradient disappearance and gradient explosion problems that occur during backpropagation but also increases the feature extraction ability and feature representation ability of the neural network model. It can further improve the image classification accuracy of the deep neural network model by jointly optimizing with other parameters in the deep neural network model. The existing model parameter initialization methods are not applicable to the deep neural network model using the MMN activation function or the maxout activation function because the theoretical derivation assumptions are based on the traditional activation function To initialize the parameters of the neural network model using the MMN activation function or the maxout activation function more quickly and efficiently, in this paper, a neural network model parameter adaptive initialization method based on the MMN activation function and maxout activation function is proposed.

Since the MMN activation function is a multilayered maxout network, the model parameter initialization method based on the multilayer maxout activation function is equally effective for the deep neural network model using the MMN activation function. The theoretical analysis and summary of the model parameter initialization method based on the maxout activation function are carried out from the forward propagation process and the backpropagation process of the deep convolutional neural network model.

2.2.1 Forward propagation process

To facilitate the theoretical derivation of the forward propagation process of the deep convolutional neural network model, the following hypothesis is proposed: all input vectors s and parameter vectors W are independent of each other and obey the same distribution. The initial distribution of the parameter vector W is symmetric about the zero point. The offset b of each layer is always equal to zero.

Here, the l indicates the lth hidden layer of the deep convolutional neural network model, and the response of the lth convolutional layer in the deep convolutional neural network model is:

$$z_l = x_l^T W_l + b_l \tag{4}$$

where $x_l \in R_d$; x_l is the state vector of the original input vector or the previous hidden layer, and the original input vector is preprocessed. The mean value is zero,

$$d = p^2 c \tag{5}$$

where d denotes the number of all input nodes connecting one neuron node, p denotes the size of the convolution kernel (the convolution kernel is square) and c denotes the number of channels input. The output of each neuron node through the maxout activation function is calculated by formula (6), so the variance of z_l is shown in formula (7):

$$f(x) = \max\left(w_1 x + b_1, w_2 x + b_2, \ldots, w_n x + b_n\right) \tag{6}$$

where n is the number of linear functions in the combination. When $w_1 = 1$ and b_1, w_2, b_2, …, w_n, b_n are all equal to zero, the maxout activation function is equivalent to the ReLU activation function. The local linear nature of the maxout activation function can alleviate the gradient disappearance problem, but at the same time, due to the introduction of additional parameters, it means that more computing resources and storage resources are needed in the training process.

$$\mathrm{Var}\left[z_l\right] = d_l \mathrm{Var}\left[W_l x_l\right] \tag{7}$$

Because the weight W_l of the lth hidden layer obeys the Gaussian distribution of zero mean, and the weight W_l and the state vector x_l are independent of each other, so

$$\mathrm{Var}\left[z_l\right] = d_l \mathrm{Var}\left[W_l\right] E\left[x_l^2\right] \tag{8}$$

where $E\left[x_l^2\right]$ is the expectation of x_l^2. To simplify, consider only the maxout activation function consisting of two linear functions, i.e.,

$$x_l = h_{l-1}\left(x_{l-1}\right) = \max\left(z_{l-1,1}, z_{l-1,2}\right) \tag{9}$$

Since the offset b_{l-1} is always equal to zero, and the mean of the weight W_l is also equal to zero, $z_{l-1,1}$ and $z_{l-1,2}$ are both symmetric about the zero point, and the mean is equal to zero.

To establish the relationship between the expected $E\left[x_l^2\right]$ and the variance Var[z_{l-1}], x_l is defined as shown in formula (10).

$$x_l = \frac{z_{l-1,1} + z_{l-1,2} + \left|z_{l-1,1} - z_{l-1,2}\right|}{2} \tag{10}$$

Substituting formula (10) into the calculation of expectation $E\left[x_l^2\right]$:

$$\begin{aligned} E\left[x_l^2\right] &= \frac{1}{4} E\left[\left(z_{l-1,1} + z_{l-1,2} + \left|z_{l-1,1} - z_{l-1,2}\right|\right)^2\right] \\ &= \frac{1}{2} E\left[z_{l-1,1}^2 + z_{l-1,2}^2 + \left(z_{l-1,1} + z_{l-1,2}\right)\left|z_{l-1,1} - z_{l-1,2}\right|\right] \\ &= \frac{1}{2}\left(E\left[z_{l-1,1}^2\right] + E\left[z_{l-1,2}^2\right] + \left(E\left[z_{l-1,1}\right]\right.\right. \\ &\qquad \left.\left. + E\left[z_{l-1,2}\right]\right) E\left[\left|z_{l-1,1} - z_{l-1,2}\right|\right]\right) \\ &= \frac{1}{2}\left(\mathrm{Var}\left[z_{l-1,1}\right] + \mathrm{Var}\left[z_{l-1,2}\right]\right) \end{aligned} \tag{11}$$

because $z_{l-1,1}$ and $z_{l-1,2}$ obey the same distribution. Therefore, you can define the variance of $z_{l-1,1}$ as shown in formula (12):

$$\mathrm{Var}\left[z_{l-1}\right] = \mathrm{Var}\left[z_{l-1,1}\right] = \mathrm{Var}\left[z_{l-1,2}\right] \tag{12}$$

Substituting formula (12) into formula (11) gives:

$$E\left[x_l^2\right] = \mathrm{Var}\left[z_{l-1}\right] \tag{13}$$

Substituting formula (12) into formula (8), the relationship between Var[z_l] and Var[z_{l-1}] can be obtained:

$$\mathrm{Var}\left[z_l\right] = d_l \mathrm{Var}\left[W_l\right] \mathrm{Var}\left[z_{l-1}\right] \tag{14}$$

When the deep convolutional neural network model has a total of L hidden layers, the relationship between the variance of the first hidden layer state Var[z_l] and the variance of the last hidden layer state ar[z_{L}] can be expressed as:

$$\mathrm{Var}\left[z_L\right] = \left(\prod_{l=2}^{L} d_l \mathrm{Var}\left[W_l\right]\right) \mathrm{Var}\left[z_l\right] \tag{15}$$

To reduce the internal covariate shift of each hidden layer in the neural network model, it must ensure that the state of each layer of neurons obeys the same distribution, namely

$$\mathrm{Var}\left[z_L\right] = \mathrm{Var}\left[z_1\right] \tag{16}$$

The initialization of the neural network model parameters needs to satisfy the sufficient conditions as shown in formula (17):

$$d_l \mathrm{Var}\left[W_l\right] = 1, \forall l \tag{17}$$

When $l = 1$, since the activation function does not directly act on the input vector, the above sufficient condition (17) still holds.

In summary, based on the neural network model, each hidden layer node state obeys the assumption of the same distribution. The model parameter initialization method proposed in this paper requires the deep convolutional neural network model in which each hidden layer parameter W_l satisfies the Gaussian distribution as shown in formula (18).

$$W_l \sim N\left(0, \frac{1}{d_l}\right). \tag{18}$$

2.2.2 Backpropagation process

In the backpropagation process of the deep convolutional neural network model, it is necessary to address the gradient obtained by each convolutional layer parameter, as follows:

$$\Delta x_l = \hat{W}_l \Delta h_l \tag{19}$$

Δx_l and Δh_l represent gradients $\frac{\partial \text{Loss}}{\partial x_l}$ and $\frac{\partial \text{Loss}}{\partial h_l}$. *Loss* is the loss function of the neural network model, Δx_l is the $c \times l$-dimensional vector, Δh_l is the amount of $\hat{d} \times 1$, where $\hat{d} = p^2 e$, where e represents the number of convolution filters. W and $\overset{\Delta}{W}_l$ can be transformed by changing dimensions since $\overset{\Delta}{W}_l$ is a $c \times \hat{d}$ matrix. Here, similar to forward propagation, it requires the relevant assumptions Δh_l and W (or $\overset{\Delta}{W}_l$) are independent of each other, W (or $\overset{\Delta}{W}_l$) obeys the initialization distribution with respect to 0 symmetry; for all l, $E[\Delta x_l] = 0$.

This article only considers the case of the maxout activation function $h_l = \max(z_{l,1}, z_{l,2})$, which can get:

$$\Delta z_{l,k} = f'(z_{l,k}) \Delta x_{l+1}, k \in \{1, 2\} \tag{20}$$

Among them, the probability that half of $f'(z_{l,k}) = 1$ and $f'(z_{l,k}) = 0$ each appear. Because $f'(z_{l,k}) = 1$ and Δx_{l+1} are independent of each other, they are satisfied for any $k \in \{1,2\}$.

$$E[\Delta h_l] = E[\Delta z_{l,k}] \tag{21}$$

Simultaneously,

$$E\left[(\Delta h_l)^2\right] = \text{Var}[\Delta h_l] = \frac{1}{2}\text{Var}[\Delta x_{l+1}] \tag{22}$$

Therefore, the variance of the gradient Δx_l is:

$$\text{Var}[\Delta x_l] = \frac{1}{2}\hat{d}_l \text{Var}[W_l] \text{Var}[\Delta x_{l+1}] \tag{23}$$

Formula (23) establishes the connection between Var[Δx_l] and Var[Δx_{l+1}]. When the deep convolutional neural network model has a total of L hidden layers, it derives the relationship between Var[Δx_2] and Var[Δx_{L+1}] as follows:

$$\text{Var}[\Delta x_2] = \text{Var}[\Delta x_{L+1}] \left(\prod_{l=2}^{L} \frac{1}{2} \hat{d}_2 \text{Var}[W_l] \right) \tag{24}$$

For the gradient to be smoothly backpropagated to the previous hidden layer, the sufficient condition for the network model parameter initialization to be satisfied is

$$\frac{1}{2}\hat{d}_l \text{Var}[W_l] = 1, \forall l \in [2, L] \tag{25}$$

Formula (25) still applies when the first hidden layer is initialized because the first layer has no activation function that acts directly on the input vector.

In summary, the initialization of the neural network model parameter W needs to obey the Gaussian distribution as shown in formula (26):

$$W_l \sim N\left(0, \frac{2}{\hat{d}_l}\right) \tag{26}$$

After the previous theoretical derivation, the initialization methods derived from the forward propagation process and the backpropagation process are obtained, but the two cannot be satisfied at the same time. To this end, this paper comprehensively considers the above problems in the following optimization problems:

$$\min_{\tau_l} (\tau_l - d_l)^2 + \left(\tau_l - \frac{1}{2}\hat{d}_l\right)^2 \tag{27}$$

Among them

$$W_l \sim N\left(0, \frac{1}{\tau_l}\right) \tag{28}$$

The optimization solution yields sufficient conditions for optimization based on these two sufficient conditions, as shown in the following formula:

$$W_l \sim N\left(0, \frac{4}{2d_l + \hat{d}_l}\right). \tag{29}$$

3 Facial expression recognition based on parameter adaptive initialization CNN-LSTM

Facial expression is a dynamic process. In addition to recognizing human emotions from a static state at a certain moment, emotional state information in real life also includes the process of continuous expression image changes, and a short-lived video sequence can more accurately reflect the emotional changes. Self-encoders, confidence networks and convolutional neural networks have strong feature learning capabilities but have weaker ability to capture contextual timing information. Therefore, this paper uses the CNN + LSTM recognition network to obtain richer and more discriminative expression information from successive human expression image sequences, thereby eliminating the influence of individual differences and the external environment and improving the accuracy of recognition. It combines a CNN network that extracts deep visual features with an LSTM network that learns to identify and synthesize temporal dynamic sequence information into a network model. This kind of network focuses on the influence between microexpressions and has more application value than single-frame microexpression recognition. The CNN network is currently the best deep learning model for 2D image applications. It can extract the depth features of images, which is the most important step in identifying expressions. LSTM is the most widely used and best network in the improved model of recurrent neural networks. In this paper, the two networks are combined to take advantage of their respective processing information for the recognition of facial expression sequences.

3.1 LSTM memory cell

LSTM is a special RNN network. LSTM introduces memory cell and control gate technology to memorize information. By setting the corresponding gate structure and controlling the flow of information in the network, LSTM can retain information in complex and sophisticated network elements for a long time. It implements the function of remembering the previous information network and updating the hidden layer parameters for the new input network. It solves the problem of gradient disappearance in BPTT training. Its structure divides a layer of the recurrent neural network into four layers and interacts in a special way. It trains the LSTM neural unit, plays the role of information filtering, remembers some of the required inter-sequence information and forgets some useless information, thus achieving the extraction and utilization of regular information. The basic LSTM unit consists of a memory unit and three control gates. The structural unit is shown in Fig. 3. The structure and calculation process of each door are described in detail below.

1. Input gate

   The input to the input gate is the output h_{t-1} of the previous hidden layer and the input x_t at this moment; x_t represents the newly received data, and h_{t-1} represents the historical information transmitted at the previous moment. Here, the tanh function is used to combine h_{t-1} and x_t and use it as the final current input. σ is the sigmoid activation function. The specific formula is as follows.

   $$i_t = \sigma\left(W^{xi}x_t + W^{hi}h_{t-1} + b^i\right) \tag{30}$$

2. Forget gate

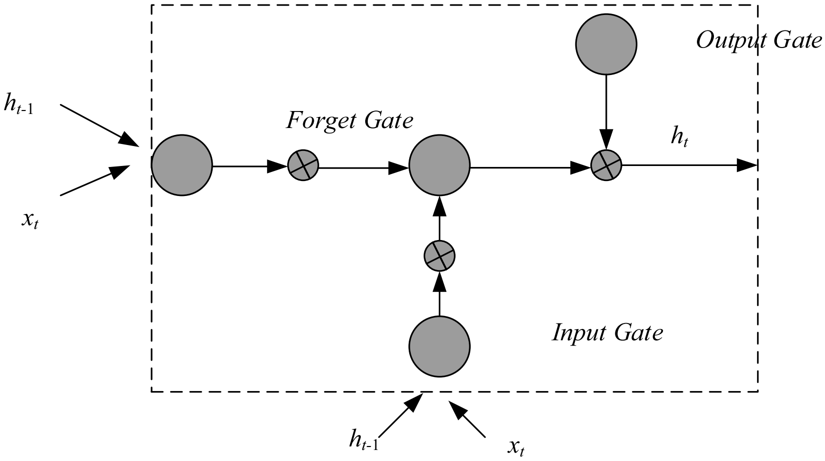
   This gate is used to control the forgetting of the current node history information. It uses the sigmoid activation function to choose what information the network remembers. In this way, the redundant historical information is cleaned up, and it refines effective associated information and efficient management system capacity information. When the value of the function is closer to 1, it means that the value of the memory information at the current time is higher, and the memory information is retained as much as possible for the next stage. When the function value is close to 0, it means that the value of the memory information at the current time is low, and more memory information is discarded. That is, 0 means no pass at all, and 1 means a complete pass.

   $$f_t = \sigma\left(W^{xf}x_t + W^{hf}h_{t-1} + b^f\right) \tag{31}$$

3. Memory cell

   This part is used to store state information, that is, to retain long-term historical information. In general, a memory block contains multiple memory cells, and all memory cells share three control gates. The three gates are used to manage the corresponding information quan-

Fig. 3 LSTM basic unit structure diagram



tity flow, discard or retain information, and the value range is [0, 1], 0 means blocking, and 1 means release. These memory units enable LSTM to learn very complex and long-term time dynamics that are not available in traditional RNN networks.

$$c_t = f_t \cdot c_{t-1} + i_{t\cdot\tanh}\left(W^{xc}x_t + W^{hc}h_{t-1} + b^c\right) \tag{32}$$

4. Output gate

   The output gate is used to control and adjust the node output information. If the node information represents the main feature, the output effect is increased; otherwise, the output information is attenuated. At the same time, it also controls the previous memory update output, which determines the size of the information output to the next moment.

$$o_t = \sigma\left(W^{xo}x_t + W^{ho}h_{t-1} + b^o\right) \tag{33}$$

$$h_t = o_{t-1} \cdot \tanh\left(c_t\right) \tag{34}$$

where i_t, f_t, c_t, o_t and h_t represent the input control gate, the forgetting control gate, the neuron activation, the output control gate, and the hidden layer vector at time t, respectively, W is a weight matrix connecting different gates, and b is the deviation vector.

3.2 Parameter adaptive initial CNN-LSTM facial expression recognition model

The built-in LSTM unit effectively solves the RNN network gradient problem and strengthens the long-term memory capability of the RNN network. In this paper, the combination of CNN and LSTM is used to learn the feature information of the facial dynamic expression sequence, and the intrinsic law in facial expression time series is deeply explored to realize facial expression recognition. The essence is to first extract an image of one person in the database into N frames and extract the features of each expression image corresponding to N CNNs to form a CNN feature sampling layer. Each layer of the CNN is connected to a single-layer LSTM, and the upper and lower N-layer LSTMs are connected to each other to form an LSTM feature learning layer; finally, a feature classification layer based on an optimized support vector machine (SVM) is formed. The specific model is shown in Fig. 4.

1. Input CNN feature sampling layer

   This layer is responsible for the preprocessing of the expression image, and then, the feature learning of the generated feature vector is performed by the parameter adaptive initialization CNN. It includes image data preprocessing, face detection, face localization, convolutional feature learning and feature sampling. First, input the original image sequence $x = \{x_1, x_2, \ldots, x_T\}$ and preprocess the image to eliminate the influence of illumination; at time $t = 1$, use the fast-CNN method to quickly locate the face image and segment it. At subsequent times, a fast tracking of the face image is performed and segmented. Then, the face image enters the parameter adaptive initialization CNN network learning and generates image abstract features through cross-convolution

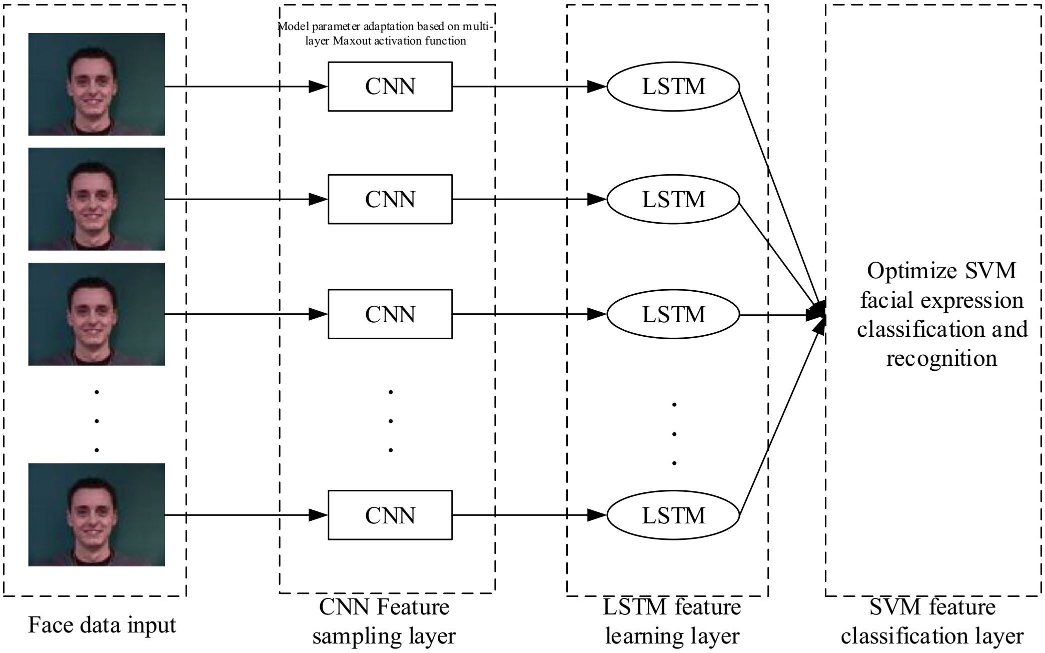

Model parameter adaptation based on multi-layer Maxout activation function
CNN
CNN
CNN
CNN
LSTM
LSTM
LSTM
LSTM
Optimize SVM facial expression classification and recognition
Face data input
CNN Feature sampling layer
LSTM feature learning layer
SVM feature classification layer

Fig. 4 Parameter adaptive initialization CNN-LSTM facial expression recognition model

and pooling. Finally, input to the K-average sampling layer, which averages the K consecutive image features of x_{T-K+1}, x_{T-K+2}, and x_T, and enters the loop network learning as an input feature. This layer feature learning does not need to be too abstract for three reasons. (1) Avoiding the improper initialization of model parameters leads to overfitting problems. (2) It reduces the number of calculations and obtains edge-level features. (3) It uses the K-average sampling layer to remove random interference and increase stability. Its basic schematic is shown in Fig. 5.

2. LSTM feature learning layer

   This layer is responsible for learning the features of the expression image sequence. It includes the RNN (recurrent neural networks) main loop and LSTM information memory. First, the convolution sampling feature vector is obtained as input from the upper layer (feature sampling layer); then, according to the time series, the loop unit LSTM extracts the information generation state information; finally, the feature vector is output to the next layer for classification.

3. SVM feature classification layer

   This layer is responsible for classifying the eigenvector output sequences learned by the cyclic network layer. The convolutional neural network has powerful distortion-invariant feature extraction capability. It maps two-dimensional data into one-dimensional data and retains most of the information by performing convolution, pooling, and other layer-by-layer operations on the original data. However, the training of CNN is not a process of convex optimization, and the model parameter group easily falls into local extremum. Moreover, the top-level structure of CNN is a fully connected feedforward neural network, which can also classify facial expressions. Therefore, this paper intends to use a support vector machine to classify facial expressions.

The support vector machine (SVM) is a classifier with structural risk minimization based on statistical learning theory. It has strong generalization ability and structural sparsity. Compared with other machine learning methods, SVM has the following characteristics. First, SVM has good generalization ability and shows obvious advantages in solving small sample problems. Second, for the linear separable problem, two kinds of sample geometric interval maximizations are used as the optimization target. For the nonlinearly separable problem, the support vector machine introduces the kernel function to convert the original input into a high-dimensional space to make it linearly separable. It maintains the complexity of the algorithm independent of the sample dimension and is only related to the number of samples. Third, in terms of noise immunity, the support vector machine uses soft interval technology to allow a small number of samples to be misclassified and imposes penalties on the misclassified samples, further enhancing the robustness of the classification. Fourth, the support vector machine

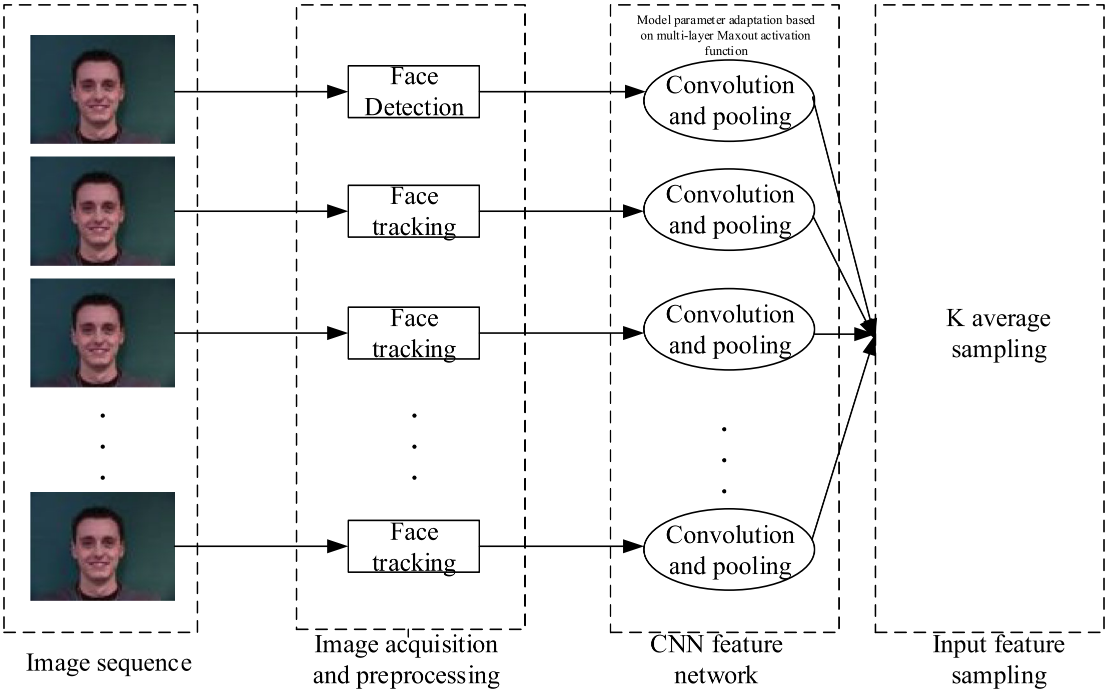


Fig. 5 Input feature sampling schematic

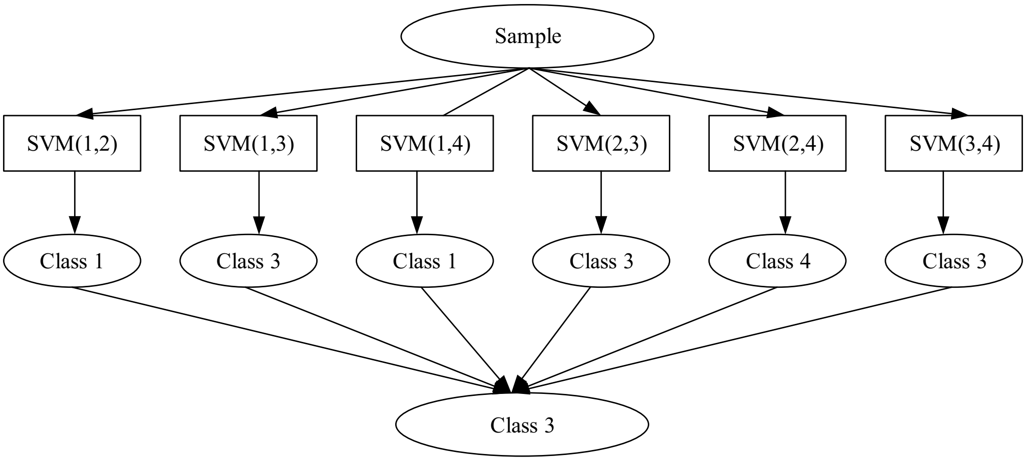


Fig. 6 Schematic diagram of the "one-to-one" support vector machine structure

is similar in structure to the three-layer feedforward neural network. The number of hidden layer nodes is determined by the support vector, and the extreme point that can guarantee the convergence of the model is the global maximum point. It can obtain both the number of hidden layer nodes and the weight vector. In addition, the support vector machine has extended the multiclassification task, including the "one-to-one" (OAO) and "one-to-many" methods. Considering the advantages of the "one-to-one method," such as high precision, low sample size and less training time, this paper selects this method for facial expression classification and recognition. OAOSVM requires more classifiers, and its implementation in the four-category problem classification is shown in Fig. 6.

For a class k classification problem, OAOSVM needs to construct $k(k-1)/2$ classifiers. After each classifier outputs the result, it is usually a common voting mechanism; given a test sample, all classifiers output their classification results for this sample. If the sample is determined to be a class t, then the class t gets one vote, and finally, the class is determined based on the most votes.

4 Experimental results and analysis

4.1 Overfitting verification experiment

In order to verify the effectiveness of the proposed method, the model parameter adaptive initialization method proposed in this paper is compared with other method parameter initialization on the CIFAR-10 dataset. The specific experimental results are shown in Table 1.

It can be seen from Table 1 that the test error rate obtained by this method is the lowest. This is because the model parameter adaptive initialization method proposed in this paper improves the image classification performance of the deep convolutional neural network model using the MMN activation function.

Table 1 Test error rates for different initialization methods on the CIFAR-100 data set

Method type	Test error rate (%)
CNN + spearmint [1]	14.98
Conv. probout + dropout [2]	11.35
Method of this paper	7.19 ± 0.12

Table 2 Test error rate for different initialization methods after CIFAR-100 data set enhancement

Method type	Test error rate (%)
CNN + spearmint [1]	9.61
Conv. probout + dropout [2]	8.87
Method of this paper	5.73 ± 0.05

In order to further verify the validity of the proposed activation function and overcome the overfitting problem of deep neural network model training, the training sample is added to the experimental dataset CIFAR-10 by data enhancement method, and then, the model is further compared. In the model training, the image block (24 × 24 pixels) is obtained from the original training image (32 × 32 pixels) and then randomly cropped as the input of the model training, and the model is supervised and trained end-to-end. For model testing, only the image block (24 × 24 pixels) in the center of the test image is used as the input for the model test. The performance comparison results of different models on the data-enhanced CIFAR-10 image classification data set are shown in Table 2.

It can be seen from Table 2 that the CIFAR-100 data set is trained by supplementing the data, which not only reduces the risk of model overfitting, but also enhances the robustness of the model. At the same time, the image classification performance of the deep convolutional neural network model is further improved. In addition, compared with other deep convolutional neural network models, the proposed method can extract the more discriminative feature information in the image, so it can further improve the image classification performance of the deep convolutional neural network model.

4.2 JAFFE facial expression database recognition experiment

To verify the effectiveness of the proposed algorithm, the facial expression recognition experiment was carried out using the international standard expression test database JAFFE to verify the validity of the generated high-dimensional features for classification recognition.

The sample data on the JAFFE emoticon is small, all Asian, and it contains seven expression categories: anger, disgust, fear, happiness, sadness, surprise and neutrality. The 213 images on the JAFFE database were grouped into experiments. The image size was 256 × 256, which was divided into three groups, two of which were used as training sets and the other as test sets, and cross-validation was performed. An example of the sample is shown in Fig. 7. The specific recognition results are shown in Fig. 8. It can be seen from Fig. 8 that the algorithm proposed in this paper can accurately identify other facial expressions in addition to not correctly identifying the neutral facial expression.

To compare the recognition effects of the proposed algorithm and other algorithms, LBP + SVM, Gabor + SVM and CNN were used to identify the experiments in the JAFFE database. The specific results are shown in Table 3.

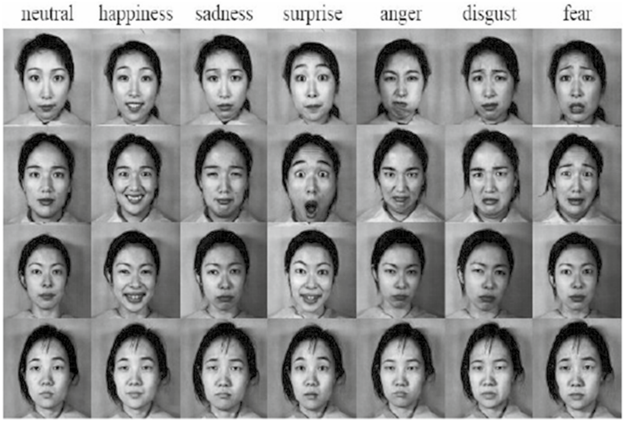
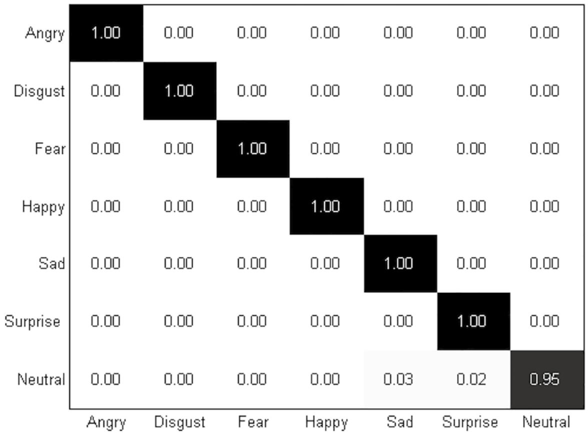


Fig. 7 Example of a facial expression image in the JAFFE database



	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Angry	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Disgust	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Fear	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Happy	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Sad	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Surprise	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Neutral	0.00	0.00	0.00	0.00	0.03	0.02	0.95

Fig. 8 JAFFE database recognition result graph

It can be seen from Table 1 that the recognition rate of the proposed algorithm in the JAFFE database is 99.3%, which is not only superior to the traditional LBP + SVM and Gabor + SVM methods but also better than the CNN method. This is mainly because this article is based on the optimization method proposed by CNN. It not only considers the deep learning of facial expressions but also considers the correlation characteristics of related expressions and adaptively optimizes the parameter initialization during the training of deep learning models. This is also the most important reason for the best expression recognition method in this paper. It also shows that the proposed algorithm is more robust than other algorithms.

4.3 Cohn-Kanade expression database recognition experiment

To further verify that the proposed facial expression recognition method is universal and effective, this paper uses the method of this paper to identify the facial expressions of the Cohn-Kanade expression database. The Cohn-Kanade database is a database of facial expressions made by 210 people aged 18–30. It also contains seven expression categories: anger, disgust, fear, happiness, sadness, surprise and neutrality. It includes approximately 2300 image sequences, each of which is encoded by a single AU or AU combination. The

Table 3 Comparison of different algorithm recognition effects

Method type	Recognition accuracy (%)
LBP + SVM	92.4
Gabor + SVM	94.1
CNN	97.5
Method of this paper	99.3

Fig. 9 Example of seven basic facial expressions in the Cohn-Kanade expression database

objects in the database are composed of people of different races and genders. Some of the images from the database are shown in Fig. 9.

To ensure sufficient sampling of the expression database, images with obvious expression features in the Cohn-Kanade database were collected, and a total of 2370 image samples containing expression category markers were obtained after fully sampling Cohn-Kanade. It includes 435 angry, 265 disgusted, 300 afraid, 406 happy, 344 despised, 343 sad and 277 pleasant. The 2370 samples were subjected to face detection and localization so that a pure face could be obtained to obtain the original expression set used in the experiment.

To improve the reliability of the recognition results, four cross-validation experiments were used in the identification, that is, all the image samples were equally divided into four groups, three of which were used for training each time, and the remaining one group of data was used for testing. Such recognition experiments were repeated 4 times, and the average of the 4 times was taken as the recognition result. The specific recognition result is shown in Fig. 10. It can be seen from Fig. 10 that the expression recognition algorithm proposed in this paper fails to identify sadness and neutrality but can accurately identify other expressions, and the average recognition rate reached 99.6%.

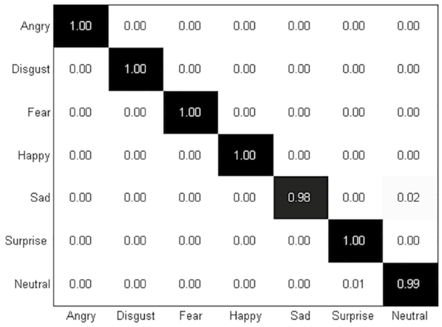


Fig. 10 Cohn-Kanade expression database image sequence recognition rate

To compare the recognition effects of the proposed algorithm and other algorithms, LBP + SVM, Gabor + SVM and CNN were used to identify the experiments in the JAFFE database. The specific results are shown in Table 4.

It can be seen from Table 2 that the recognition algorithm proposed in this paper had a recognition rate of 99.6% on the Cohn-Kanade expression database, which is not only superior to the traditional LBP + SVM and Gabor + SVM method but also higher than the CNN method. This is mainly because this paper is based on the optimization method proposed by CNN. It not only considers the temporal correlation characteristics of related expressions but also has a higher expression recognition rate with SVM as the top-level classifier. This is because the expression features extracted by CNN are basically separable, and the expression classification surface is optimized by the property of SVM structure risk minimization. The sparseness of support vector compensates for the insufficient feature extraction ability of CNN in a small sample. Moreover, it also adaptively optimizes the parameter initialization during the training of the deep learning model. The experimental results fully demonstrate that the proposed algorithm is more robust than other algorithms.

4.4 Static expression image recognition and image sequence expression recognition experiment

To compare the performance of the static image recognition and temporal expression image recognition module, separate segmentation image blocks and whole images were experimentally performed on the Cohn-Kanade expression database. The specific steps are as follows: (1) design a 16-layer convolutional neural network for static image recognition and (2) use the recognition algorithm proposed in this paper to identify dynamic sequence facial images. The specific recognition results are shown in Tables 5 and 6. As seen in Tables 5 and 6, the recognition rate of the static image recognition algorithm is 92.3%. The recognition rate of the recognition algorithm proposed in this paper is 98.9%, which is an increase of 6.6%. The difference in the recognition effect between the two methods is still relatively large. At the same time, the static expression image recognition algorithm

Table 4 Comparison of different algorithm recognition effects

Method type	Recognition accuracy (%)
LBP + SVM	87.7
Gabor + SVM	88.7
CNN	98.6
Method of this paper	99.6

Table 5 Static expression image recognition results of the Cohn-Kanade expression database

Emoticon type	Number of test samples	Correct identification number						Average recognition rate (%)
		Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	
Neutral	50	41	41	42	43	40	41	82.7
Angry	60	54	54	54	54	54	55	90.3
Disgust	30	28	28	28	28	28	28	93.3
Fear	35	31	31	31	31	31	31	88.6
Happy	45	44	44	44	44	44	44	97.8
Sadness	40	40	40	40	40	40	40	100
Surprised	30	28	28	28	28	28	28	93.3
Total	290	266	266	267	268	265	267	92.3

Table 6 Dynamic expression image recognition results of the Cohn-Kanade expression database

Emoticon type	Number of test samples	Correct identification number						Average recognition rate (%)
		Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	
Neutral	50	48	48	48	48	47	49	96.0
Angry	60	58	58	59	58	58	57	96.7
Disgust	30	30	30	30	30	30	30	100
Fear	35	35	35	35	35	35	35	100
Happy	45	45	45	45	45	45	45	100
Sadness	40	40	40	40	40	40	40	100
Surprised	30	30	30	30	30	30	30	100
Total	290	266	266	267	268	265	267	98.9

can only accurately identify the sad expression and cannot accurately identify other expressions. However, the methods proposed in this paper only have neutral and angry expressions that are not correctly recognized, and other expressions are correctly identified.

From the experimental results of Tables 5 and 6 comparison, the method proposed in this paper is significantly higher than the traditional deep learning recognition method. In this paper, based on the parameter adaptive initialization of CNN and LSTM, 98.9% of the average recognition rate of expression was obtained on the Cohn-Kanade dataset. This is mainly because this paper fully considers that deep learning easily falls into the problems of overfitting and gradient explosion, making full use of the relationship between model parameter initialization and these problems, and comprehensively considering the timing characteristics of the facial expressions. At the same time, it combines the LSTM method to refine these features and obtain a more appropriate facial expression recognition model to achieve more accurate facial expression recognition. This experiment further validates the outstanding advantages of this method in the temporal correlation of facial expression recognition.

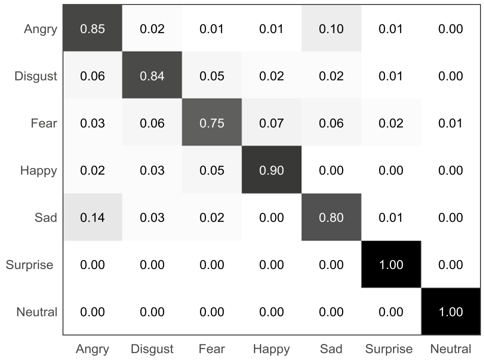
4.5 Other expression database experiments

In order to further verify that the facial expression recognition method proposed in this paper can be universal and effective. The facial expressions of BU-3DFE expression database [25], FER2013 expression database [26] and Oulu-CASIA expression database [27] were separately identified by this method. The BU-3DFE Expression Database covers face scanning point clouds of 100 subjects, including undergraduate, graduate and faculty members at Binghamton University and US State University. This database contains 60% of women and 40% of men. Each subject showed seven different expressions in front of the 3D scanning system: anger, disgust, fear, happiness, sadness, surprise, and neutrality. And some of the expressions are shown in Fig. 1a. The FER2013 expression database includes seven facial expressions of anger, disgust, fear, happiness, sadness, surprise and neutrality. Some expressions are shown in Fig. 1b. The Oulu-CASIA expression database contains six basic expressions from 80 people. They are anger, disgust, fear, happiness, sadness and surprise. This paper selects the data under normal light, and some of the expressions are shown in Fig. 11c.

To improve the reliability of the recognition results, four cross-validation experiments were used in the identification, that is, all the image samples were equally divided into four

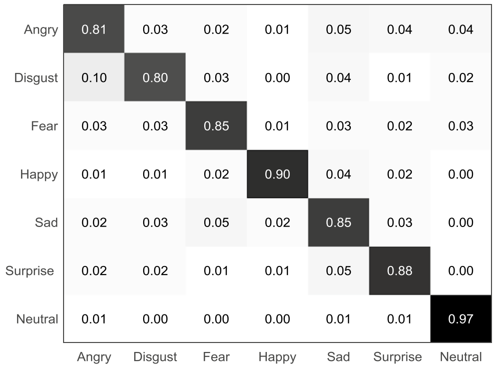
(a) BU-3DFE (b) FER2013 (c) Oulu-CASIA

Fig. 11 Example of a facial expression image from BU-3DFE, FER2013 and Oulu-CASIA databases



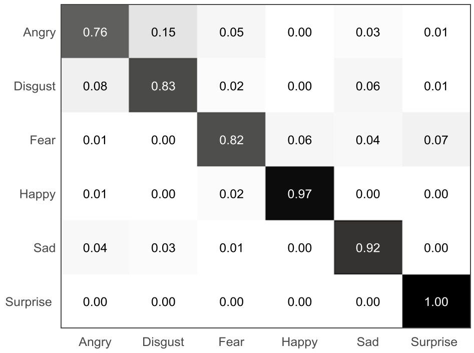
	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Angry	0.85	0.02	0.01	0.01	0.10	0.01	0.00
Disgust	0.06	0.84	0.05	0.02	0.02	0.01	0.00
Fear	0.03	0.06	0.75	0.07	0.06	0.02	0.01
Happy	0.02	0.03	0.05	0.90	0.00	0.00	0.00
Sad	0.14	0.03	0.02	0.00	0.80	0.01	0.00
Surprise	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Neutral	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Fig. 12 Confusion matrix on BU-3DFE



	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Angry	0.81	0.03	0.02	0.01	0.05	0.04	0.04
Disgust	0.10	0.80	0.03	0.00	0.04	0.01	0.02
Fear	0.03	0.03	0.85	0.01	0.03	0.02	0.03
Happy	0.01	0.01	0.02	0.90	0.04	0.02	0.00
Sad	0.02	0.03	0.05	0.02	0.85	0.03	0.00
Surprise	0.02	0.02	0.01	0.01	0.05	0.88	0.00
Neutral	0.01	0.00	0.00	0.00	0.01	0.01	0.97

Fig. 13 Confusion matrix on FER2013



	Angry	Disgust	Fear	Happy	Sad	Surprise
Angry	0.76	0.15	0.05	0.00	0.03	0.01
Disgust	0.08	0.83	0.02	0.00	0.06	0.01
Fear	0.01	0.00	0.82	0.06	0.04	0.07
Happy	0.01	0.00	0.02	0.97	0.00	0.00
Sad	0.04	0.03	0.01	0.00	0.92	0.00
Surprise	0.00	0.00	0.00	0.00	0.00	1.00

Fig. 14 Confusion matrix on Oulu-CASIA

Table 7 Comparison of recognition results of different algorithms in BU-3DFE database

Method type	Recognition accuracy (%)
Lopes [29]	75.6
CNN	84.1
3DCNN [28]	86.8
Method of this paper	87.7

Table 8 Comparison of recognition results of different algorithms in FER2013 database

Method type	Recognition accuracy (%)
CNN	84.5
CNN and CPC fusion method	85.3
Method of this paper	86.6

groups, three of which were used for training each time, and the remaining one group of data was used for testing. Such recognition experiments were repeated 4 times, and the average of the 4 times was taken as the recognition result. The three types of expression libraries are identified by the method of this paper. The specific recognition results are shown in Figs. 12, 13 and 14. At the same time, in order to better verify the advantages of this method, the BU-3DFE expression database is identified by the methods proposed by CNN, 3DCNN [28] and Lopes [29]. The specific recognition results are shown in Table 7. The FER2013 expression database is identified by the CNN, CNN and CPC fusion methods, and the recognition results are shown in Table 8. The Oulu-CASIA expression database was identified by CNN, PPDN [23] and FN2EN [30] methods, respectively. The specific recognition results are shown in Table 9.

It can be seen from Tables 3, 4 and 5 that the recognition accuracy of the proposed algorithm in the BU-3DFE expression database, the FER2013 expression database and

Table 9 Comparison of recognition results of different algorithms in Oulu-CASIA database

Method type	Recognition accuracy (%)
CNN	75.3
PPDN [23]	84.6
FN2EN [30]	87.7
Method of this paper	88.3

the Oulu-CASIA expression database are 87.7, 86.6 and 88.3%, respectively. They are not only superior to traditional PPDN methods, but also better than CNN, CNN and CPC fusion, 3DCNN and other methods. This is mainly because this paper is based on the optimization method proposed by CNN. It not only considers the deep learning of facial expressions, but also considers the correlation characteristics of expressions. At the same time, it also avoids problems such as overfitting that deep learning is easy to fall into. This is also the most important reason for the best expression recognition method in this paper. It also shows that the proposed algorithm is more robust than other algorithms.

5 Conclusion

In this paper, an adaptive model parameter initialization convolutional neural network based on MMN linear activation function and a long–short-term memory network are combined to identify facial expressions. The initialization of the adaptive model parameters proposed in this paper effectively overcomes the overfitting problems in the deep learning model training process, and it avoids the problems of poor recognition of the model due to improper initialization of parameters.

Aiming at the lack of effective utilization of the dynamic sequence information of the expression image and the poor robustness of the algorithm, a dynamic expression recognition method combining CNN and LSTM is designed. The method uses CNN to perform the cyclic collection of image sequences, LSTM learning and memory sequence association information and combines single image information and sequence association information for expression discrimination. The small-scale image data are used to quickly locate the face image, and then, the convolutional neural network is used to extract the visual features. The cyclic neural network is used to establish the overall cyclic network structure. The single unit learns the image sequence data from LSTM to acquire and store the correlation characteristics. This paper compares JAFFE, Cohn-Kanade, BU-3DFE, FER2013 and Oulu-CASIA facial expression database and compares it with traditional facial expression recognition method, 3DCNN method, PPDN method, FN2EN method and CNN method. The experimental results show that the proposed method not only recognizes the static expressions of various static images but also better recognizes the facial expressions with time-series characteristics. The recognition effect is better than the existing mainstream facial expression recognition methods and other deep learning facial expression recognition algorithms.

Acknowledgements This paper is supported by National Natural Science Foundation of China (No. 61701188).

References

1. Pransky, J.: The Pransky interview–Martin Haegele, Head of Department Robotics and Assistive Systems. Fraunhofer IPA. Ind. Robot Int. J. **45**(3), 307–310 (2018). https://doi.org/10.1108/IR-04-2018-0060
2. Vouloutsi, V., Verschure, P.F.M.J.: Emotions and self-regulation. Living Mach. Handb. Res. Biomim. Biohybrid Syst. **10**, 327 (2018)
3. Pickett, L.: Don't fear the cobot: collaborative robots, or cobots, are infiltrating factories on a global scale. But can robots and humans really work together in harmony? We asked the experts. Quality **57**(1), 12A (2018)
4. Wu, Y., Schuster, M., Chen, Z. et al.: Google's neural machine translation system: bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)
5. Mehrabian, A.: Communication without words. Commun. Theory **12**, 193–200 (2008)
6. Deng, H.B., Jin, L.W., Zhen, L.X., et al.: A new facial expression recognition method based on local Gabor filter bank and PCA plus lda. Int. J. Inf. Technol. **11**(11), 86–96 (2005)
7. Shan, C., Gong, S., McOwan, P.W.: Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis. Comput. **27**(6), 803–816 (2009)
8. Satiyan, M., Nagarajan, R., Hariharan, M.: Recognition of facial expression using Haar wavelet transform. Trans. Int. J. Electr. Electron. Syst. Res. JEESR Univ. Technol. Mara UiTM **3**, 91–99 (2010)
9. Chen, J., Takiguchi, T., Ariki, Y.: Facial expression recognition with multithreaded cascade of rotation-invariant HOG. In: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, pp. 636–642 (2015)
10. Soyel, H., Demirel, H.: Improved SIFT matching for pose robust facial expression recognition. In: 2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops (FG 2011), IEEE, pp. 585–590 (2011)
11. Yu, Z., Zhang, C.: Image based static facial expression recognition with multiple deep network learning. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM, pp. 435–442 (2015)
12. Jung, H., Lee, S., Yim, J. et al.: Joint fine-tuning in deep neural networks for facial expression recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2983–2991 (2015)
13. Eleftheriadis, S., Rudovic, O., Pantic, M.: Discriminative shared Gaussian processes for multiview and view-invariant facial expression recognition. IEEE Trans. Image Process. **24**(1), 189–204 (2015)

14. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
15. Liu, M., Shan, S., Wang, R. et al.: Learning expression lets on spatio-temporal manifold for dynamic facial expression recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1749–1756 (2014)
16. Maninchedda, F., Oswald, M.R., Pollefeys, M.: Fast 3d reconstruction of faces with glasses. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp. 4608–4617 (2017)
17. Kacem, A., Daoudi, M., Amor, B.B. et al.: A novel space-time representation on the positive semidefinite cone for facial expression recognition. In: ICCV, pp. 3199–3208 (2017)
18. Liu, P., Han, S., Meng, Z. et al.: Facial expression recognition via a boosted deep belief network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1805–1812 (2014)
19. Lopes, A.T., de Aguiar, E., Oliveira-Santos, T.: A facial expression recognition system using convolutional networks. In: 2015 28th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), IEEE, pp. 273–280 (2015)
20. Zhang, F., Yu, Y., Mao, Q., et al.: Pose-robust feature learning for facial expression recognition. Front. Comput. Sci. **10**(5), 832–844 (2016)
21. Zhang, T.: Facial expression recognition based on deep learning: a survey. In: International Conference on Intelligent and Interactive Systems and Applications, Springer, Cham, pp. 345–352 (2017)
22. Zhang, K., Huang, Y., Du, Y., et al.: Facial expression recognition based on deep evolutional spatial-temporal networks. IEEE Trans. Image Process. **26**(9), 4193–4203 (2017)
23. Zhao, X., Liang, X., Liu, L., et al.: Peak-piloted deep network for facial expression recognition. In: European Conference on Computer Vision, Springer, Cham, pp. 425–442 (2016)
24. Cao, C., Weng, Y., Zhou, S., et al.: Facewarehouse: a 3d facial expression database for visual computing. IEEE Trans. Vis. Comput. Gr. **20**(3), 413–425 (2014)
25. Yin, L., Wei, X., Sun, Y., et al.: A 3D facial expression database for facial behaviour research. In: 7th International Conference on Automatic Face and Gesture Recognition, FGR 2006, IEEE, pp. 211–216 (2006)
26. Goodfellow, I.J., Erhan, D., Carrier, P.L., et al.: Challenges in representation learning: a report on three machine learning contests. Neural Netw. **64**, 59–63 (2015)
27. Zhao, G., Huang, X., Taini, M., et al.: Facial expression recognition from near-infrared videos. Image Vis. Comput. **29**(9), 607–619 (2011)
28. Liu, M., Li, S., Shan, S., et al.: Deeply learning deformable facial action parts model for dynamic expression analysis. In: Asian Conference on Computer Vision, Springer, Cham, pp. 143–157 (2014)
29. Lopes, A.T., de Aguiar, E., De Souza, A.F., et al.: Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. Pattern Recognit. **61**, 610–628 (2017)
30. Ding, H., Zhou, S.K., Chellappa, R.: Facenet2expnet: regularizing a deep face recognition net for expression recognition. In: 2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017), IEEE, pp. 118–126 (2017)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Fengping An received B.S. degree from School of Economic and Management, Hefei University, Hefei, China, in 2008 and M.S. degree from School of Economics and Management, Hebei University of Engineering, Handan, China, in 2011 and Ph.D. degree from School of Computer and Communication Engineering, Beijing University of Science and Technology. He has worked in Huaiyin Normal University. His research areas are image processing, artificial intelligence and pattern recognition.

Zhiwen Liu received a Ph.D. degree in communication and electronic systems from Beijing Institute of Technology. In 1997, he was promoted as professor at Beijing Institute of Technology. He has been identified as a young academic leader in school for three consecutive times. He is currently the director of the Department of Signal and Image Processing, the person in charge of the key professional "Electronic Information Engineering" of the National Defense Science and Technology Commission, and the leader of the life information engineering discipline.