Hindawi
Scientific Programming
Volume 2020, Article ID 8845176, 12 pages
https://doi.org/10.1155/2020/8845176

Research Article

A Facial Expression Recognition Method Using Improved Capsule Network Model

Yifeng Zhao and Deyun Chen

School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, Heilongjiang 150080, China

Correspondence should be addressed to Deyun Chen; chendeyun@hrbust.edu.cn

Received 10 September 2020; Revised 25 September 2020; Accepted 14 October 2020; Published 27 October 2020

Academic Editor: Wenzheng Bao

Aiming at the problem of facial expression recognition under unconstrained conditions, a facial expression recognition method based on an improved capsule network model is proposed. Firstly, the expression image is normalized by illumination based on the improved Weber face, and the key points of the face are detected by the Gaussian process regression tree. Then, the 3dmms model is introduced. The 3D face shape, which is consistent with the face in the image, is provided by iterative estimation so as to further improve the image quality of face pose standardization. In this paper, we consider that the convolution features used in facial expression recognition need to be trained from the beginning and add as many different samples as possible in the training process. Finally, this paper attempts to combine the traditional deep learning technology with capsule configuration, adds an attention layer after the primary capsule layer in the capsule network, and proposes an improved capsule structure model suitable for expression recognition. The experimental results on JAFFE and BU-3DFE datasets show that the recognition rate can reach 96.66% and 80.64%, respectively.

1. Introduction

Human facial expression is a kind of representation language which is naturally or deliberately revealed by the complex stimulation of environment, context, and mood in the process of communication and can be perceived by the visual system [1–3]. It is also the expression that human facial muscles produce stress movement under certain semantic stimulation or active movement driven by consciousness. It is generally believed that human facial expressions are controlled and stress-induced. The so-called controllability of consciousness means that human beings (especially actors) can make or suppress any expression at random [4, 5]. The so-called stress convergence refers to that most people will make similar expressions under the stimulation of a specific semantic environment. For example, when people hear interesting events or face beautiful things, they will naturally show the expression of happiness; when people face expired smelly food or ugly bad scenes, they will generally show the expression of disgust; when people face sudden emergencies, they usually show the expression of surprise, etc. [6].

When the human stress expression is contrary to the normal situation, it usually reflects the inhibition of some psychological factors on the expression. Because the human facial expression is rich in psychological and emotional information, has stress convergence, and is controlled by consciousness, it is easy to get the general attention of a large number of scholars in the field of psychology and pattern recognition [7, 8]. Under the existing technical conditions, people began to use pattern recognition technology to establish the mapping relationship between the face image and the facial expression in the image. Through the computer automatic judgment of human facial expression, the research field of expression recognition is proposed. In a broad sense, facial expression recognition is a process of automatic analysis of face image data by computer. The process of automatic image analysis by computer is just the main content of computer vision [9]. As a special subject born out of the traditional pattern recognition technology, computer

vision studies the difficulties faced by the traditional pattern recognition methods in image data or processes and refines the image data to facilitate the indexing, classification, and automatic analysis of image data [10].

In this paper, the practical significance of research on facial expression recognition technology of single face image is stronger than that of expression recognition technology based on image sequence. The reason is that compared with expression recognition technology based on image sequence, expression recognition technology based on single image can reflect the defects of existing image processing technology and recognition technology in a specific application. It is helpful to improve the applicability of the existing image processing and discrimination technology. Therefore, this paper focuses on the expression recognition of a single face image under unconstrained conditions. For the problem of illumination and pose standardization of face image, the existing lighting and pose standardization methods are easy to lose texture details, so it is not suitable for facial expression recognition. This paper considers that under the premise of unconstrained expression recognition, it is necessary to further improve the degree of expression detail reservation. Thus, the subsequent discriminant model can work effectively. To solve the problem of facial expression recognition, a facial expression recognition method using improved capsule network model is proposed. The main contributions of this paper are as follows.

In this paper, we consider that the convolution features used in facial expression recognition need to be trained from the beginning and add as many different samples as possible in the training process. The proposed method focuses on temporal attention. The attention module uses sigmoid as an activation function, which can not only select important features but also suppress irrelevant information. It can also help smooth the mismatch between the training set and test set and improve the final recognition rate.

2. Related Works

For the facial expression recognition under nonconstrained condition, scholars have proposed many methods. For example, reference [11] proposed a hybrid expression recognition method using High-order Joint Derivative Local Binary Pattern (HJDLBP) and Local Binary Pattern (LBP). The model efficiency is improved by removing unwanted areas and preserving the facial area. The study in [12] proposed a facial expression recognition framework combining two-dimensional Gabor and local binary patterns. By extracting salient features of facial expression, the model efficiency was improved. The study in [13] proposed an adaptive model parameter initialization method based on the multilayer maxout network linear activation function, which improved the performance of the model by extracting highly relevant features of the image sequence. The study in [14] proposed an expression recognition method based on Wasserstein generative adversarial network, which improved the model efficiency by suppressing slight changes of the face. The study in [15] proposed a Deep Cascaded Peak-piloted Network (algorithm in reference [15]), which extracts key and subtle details in the image through peak-conducting feature transformation to improve the accuracy of the model. However, these methods do not consider the edge characteristics of the image.

The study in [16] proposed a facial expression recognition method combining multiple facial features and support vector machines. By extracting important facial features, reducing image noise points can improve the accuracy of the model. The study in [17] proposed a deep convolution BiLSTM fusion network facial expression recognition method, which extracts spatial features from each frame through a convolutional neural network and then models the temporal dynamics. Feature fusion improves the model recognition rate. The study in [18] proposed the facial expression recognition method based on a facial video sequence. By extracting features represented by temporal local binary pattern, the efficiency of the model was improved. The study in [19] proposed a conditional convolutional neural network enhanced random forest expression recognition method (algorithm in Reference [19]), which reduces the noise points of the data set and improves the accuracy of the model. However, when the training data is small, these methods are prone to underfitting.

The study in [20] proposed facial expression recognition based on incremental active learning, which improves the accuracy of the model by reducing the noise points of the image. The study in [21] proposed a multifeature fusion facial expression recognition method based on Extreme Learning Machine (ELM), which improved the accuracy of the model by fusing multiple features. The study in [22] proposed a facial expression recognition method based on feature space and principal component analysis. The method encodes the known image through feature space to improve the accuracy of the model. The study in [23] proposed a facial expression recognition method based on the Two-Stream Convolutional Neural Network (T-SCNN), which improved the accuracy of the model by fusing RGB images and temporal features. However, when the amount of data is large, these methods are prone to overfitting. The study in [24] proposed a multilayer perceptron algorithm for facial expression recognition, which increased the accuracy of the model by adding hidden neurons. However, the parameters of the model were difficult to adjust. The study in [25] proposed a facial expression recognition method based on hidden Markov, which improved the efficiency of the model by extracting the more important features of the image. However, under unconstrained conditions, the model is less robust.

Based on the above analysis, in the field of facial expression recognition, deep learning has good modeling and processing ability for facial expression images, but only when the face illumination and pose are constrained can the model be effectively recognized. Aiming at the problem of facial expression recognition under unconstrained conditions, a facial expression recognition method based on an improved capsule network model is proposed. The improved capsule model can effectively classify facial expressions under unconstrained conditions, which makes up for the deficiency of pure deep convolution network in acquiring

sparse features hidden in discriminative texture, and improves the generalization ability of existing expression classification models for illumination and pose differences.

3. Overall Architecture of the Proposed Method

Through the analysis of the existing deep convolution neural network in the application of expression recognition, this paper thinks that the illumination and posture correction technology has very important application value in alleviating the dependence of deep convolution neural network on the number of samples and improving the quality of perception weight. It can be set by using light treatment technology. We can change the lighting conditions of the face to generate expression samples under different lighting conditions; we can also use the posture correction technology to realize the generalization of the face pose to generate different facial expression samples. According to the idea of local density sampling, this may make the final training model to the real model. The framework of facial expression recognition based on deep learning is shown in Figure 1. In this paper, the illumination and projection analysis technology is still used to analyze, correct, and perceive the face pattern. Under the existing technical conditions, this kind of pretreatment is still a necessary step. In the specific implementation process, a batch of dense sample graphs are generated and input into the deep convolution model for weight training to fully alleviate the problem of insufficient sample number. In the stage of model recognition, illumination and projection analysis are still used to analyze, correct, and perceive the face pattern. In this paper, we add an attention layer after the primary capsule layer in the capsule network and propose that a capsule structure is suitable for expression feature extraction.

4. The Process of the Proposed Method

4.1. Illumination Normalization. A new illumination normalization method based on Weber face (WF) [26] is proposed, which can not only extract illumination-insensitive features effectively but also suppress boundary marks at sudden changes of light. Assume that the lighting component is $I(x, y)$. In WF, all ratios are multiplied by a combination coefficient α and $\alpha \in (0, 1)$, as follows:

$$\mathrm{WF}(x, y) = \arctan\left(\alpha \sum_{i\in A}\sum_{j\in A} \frac{R(x, y) - R(x - i\Delta x, y - i\Delta y)}{R(x, y)}\right). \tag{1}$$

At this time, while enhancing the effective information, the noise will also be enhanced, and the coding value of the area most affected by the light is $(-K_0, -K_1) \cup (K_0, K_1)$. Therefore, to reduce the noise, multiply the interval $(-K_0, -K_1) \cup (K_0, K_1)$ by a suppression factor. The revised WF definition is as follows:

$$\mathrm{WF}(x, y) = \begin{cases} \arctan(\alpha \cdot \arg), & \arg \notin B, \\ \arctan(\alpha \cdot \beta \cdot \arg), & \arg \in B, \end{cases} \tag{2}$$

$$\arg = \sum_{i\in A}\sum_{j\in A} \frac{I(x, y) - I(x - i\Delta x, y - i\Delta y)}{I(x, y)}, \tag{3}$$

$$B = (-K_0, -K_1) \cup (K_0, K_1), \tag{4}$$

where $\beta \in (0, 1)$ is a coefficient that suppresses the influence of light: $A = \{-1, 0, 1\}$.α is used to adjust (increase or decrease) the difference between the WF encoding values of adjacent pixels: the interval B is the interval that is greatly affected by light. The interval has more noise. It can be known from equation (2):

$$\mathrm{WF}(x, y) \in \left[\frac{-\pi}{2}, \frac{\pi}{2}\right]. \tag{5}$$

According to the WF theorem, the minimum perceivable ratio is constant. Therefore, the subinterval of $[-\pi/2, \pi/2]$ is called the low perceptual interval. In the subinterval, since the ratio is smaller than the minimum perceptible ratio, its changes will not be perceivable by the human eye. That is, even if the pixels of the interval are affected by the light, the change is small and can be ignored. k_0 can be defined as follows:

$$k_0 = \tan(K_0). \tag{6}$$

The low-frequency component is regarded as a large-scale feature, which is the part mainly affected by light. The high-frequency component is regarded as a small-scale feature, which is the light-invariant feature. Those close to $\pm\pi/2$ are changing fast and can be regarded as high-frequency components, and those close to 0 are slow changing and can be regarded as low-frequency components. The interval $[-\pi/4, \pi/4]$ is regarded as low-frequency interval. $[-\pi/2, -\pi/4]$ and $[\pi/4, \pi/2]$ are regarded as high-frequency interval. What needs to be suppressed is suitable low-frequency interval, so k_1 can be defined as follows:

$$k_1 = \tan(K_1), \quad K_1 \in \left(K_0, \frac{\pi}{4}\right). \tag{7}$$

4.2. Key Point Detection. For key point detection [27], a model based on the Gaussian process regression tree is proposed. A special kernel function and random partition kernel function are designed. Given a random partition P, the definition of the kernel function is as follows:

$$k_p(a, b) = E[I[\rho(a) = \rho(b)]]_{\rho-p}, \tag{8}$$

where I is the indicator function and $\rho(a)$ refers to the cluster that the partition ρ assigns to a, and the kernel function is defined as the segment that is assigned to the same cluster.

According to Moser theorem, if the function is a semipositive definite function, it is also a kernel function.

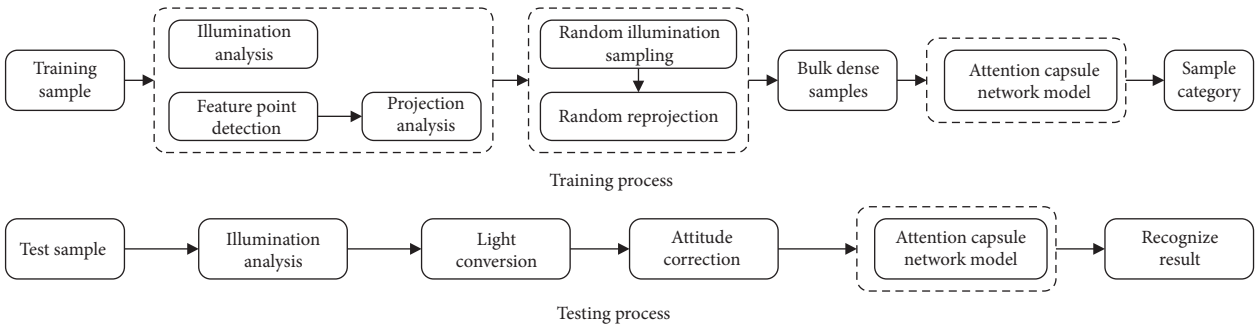

Figure 1: The overall framework of the proposed method.

Firstly, $k_p(a,b)$ is proved to be a reasonable kernel function. Define

$$k_p(a,b) = I[\rho(a) = \rho(b)]. \tag{9}$$

To prove that k_p is a semipositive definite function, the expectation is decomposed into the limit of summation and each single term is proved to be a semipositive definite function.

$$\begin{aligned} k_p(a,b) &= E[I[\rho(a) = \rho(b)]]_{\rho-p} \\ &= \lim_{n\longrightarrow\infty} \frac{1}{n} \sum_{\rho-p}^{n} I[\rho(a) = \rho(b)] \\ &= \lim_{n\longrightarrow\infty} \frac{1}{n} k_p(a,b). \end{aligned} \tag{10}$$

For any dataset of size N, the covariance matrix of k_p can be arranged into a diagonal matrix:

$$YK_\rho Y^T = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & 0 \\ 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{11}$$

It can be seen that $YK_\rho Y^T$ is a semipositive definite matrix. Therefore, for any dataset, K_ρ is also a semipositive definite matrix and it can be concluded that k_ρ is a reasonable kernel function. Analogously, the kernel function defined in the random partition is applied to the random forest. The kernel function $k(x_1, x_2)$ of the Gaussian process regression tree can be composed of M trees and the distribution of nodes in the tree. The formula is shown in the following formulas:

$$k(x_1, x_2) = \sigma_k^2 \sum_{m=1}^{M} h^m(x_1, x_2), \tag{12}$$

$$h^m(x_1, x_2) = \begin{cases} 1, & \tau^m(x_1) = \tau^m(x_2), \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

τ^m is the split function, and σ_k^2 refers to the scaling parameters of the kernel function.

σ_n^2 and σ_k^2 are the hyperparameters of the Gaussian process regression tree. The maximum likelihood function is solved by the probability density function of the training samples. The formula is shown as follow:

$$P\left(S|X, \sigma_k^2, \sigma_n^2\right) = \frac{1}{(2\pi)^{n/2} K_s^{1/2}} \exp\left\{-\frac{1}{2} S^T K_s^{-1} S\right\}. \tag{14}$$

Take the logarithm of the above formula:

$$\log P\left(S|X, \sigma_k^2, \sigma_n^2\right) = -\frac{1}{2} S^T K_s^{-1} S - \frac{1}{2} \log|K_s| - \frac{n}{2} \log 2\pi. \tag{15}$$

Let $\sigma_r^2 = \sigma_n^2/\sigma_k^2$ and find the maximum value of the maximum likelihood function. Derivate σ_r:

$$\frac{\partial}{\partial \sigma_r} \log P\left(S|X, \sigma_k^2, \sigma_r^2\right) = \frac{1}{2} \mathrm{tr}\left(\left(\alpha\alpha^T - K_s^{-1}\right) \frac{\partial K_s}{\partial \sigma_r}\right). \tag{16}$$

The nonparametric nature of Gaussian process regression results in a large amount of calculation and the computational complexity of K_s^{-1} is $O(N^3)$. Using the reduced-rank approximation method can reduce the amount of calculation. Let $K(X,X) = \sigma_k^2 QQ^T$ and K_s^{-1} can be simplified to the following formula:

$$K_s^{-1} = \sigma_k^{-2}\left(\sigma_r^{-2} I - \sigma_r^{-2} Q K_r^{-1} Q^T\right), \tag{17}$$

where

$$K_r^{-1} = Q^T Q + \sigma_r^2 I. \tag{18}$$

Due to the special construction of the kernel function, $Q = (q_1, \ldots, q_N)^T$, let $q_i = (q_i^1, \ldots, q_i^M)^T$. q_i^m is the index of the leaf node of the sample$i$that falls into the tree$m$. K_r is the matrix of size BM*BM and the computational complexity also changes from $O(N^3)$ to $O((\mathrm{BM})^3)$.

4.3. Face Posture Standardization. Based on the 3D Morphable Model (3DMM), a new face posture normalization method is proposed [28]. The eigenvalue corresponding to the weighting coefficient of each eigenvector is covariance. 0 is the Gaussian probability distribution of the mean and the expression is as follows:

$$\begin{aligned} S &= \overline{S} + \sum_{i=1}^{n} \alpha_i E_i^S, \\ \alpha_i &\sim N\left(0, \lambda_i^S\right), \end{aligned} \tag{19}$$

where $\overline{S}$ is the average three-dimensional face shape, E_i^S is each principal component obtained by the PCA algorithm on the three-dimensional vertex dataset of the face. λ_i^S is the corresponding eigenvalue of E_i^S, and the combination coefficient α_i^S follows a Gaussian distribution with 0 as the mean and λ_i^S as the variance. Extract facial feature point P_i in the model and the corresponding part does not change the linear combination relationship. So:

$$S^{P_j} = \overline{S}^{P_j} + \sum_{i=1}^{k} \alpha_i E_i^{P_j}. \tag{20}$$

Given a face image, obtain the 2-dimensional feature point estimate p_i: (x'', y_i'') of the point P_j in the face image and record P^+ as the position of the camera coordinate system origin in the world coordinate system of S, and E^{p_i} is the matrix formed by merging the feature vector set $\{E^{p_i}\}$ by column. $R_{3:}$ is the third row of the rotation matrix R. M_c is the internal parameter matrix of the camera, and z'' is the common scale factor of the projection imaging. N point energy equation combined with the deformation coefficient estimate is as follows:

$$\begin{aligned} &F\left((M_c, R, z'', P^+), \alpha\right) \\ &= \sum_i \left\| \left(p_i, z''\right)^T - z'' \frac{M_c R\left(\overline{S}^{P_i} - P^+ + E^{p_i}\alpha\right)}{R_{3:}\left(\overline{S}^{P_i} - P^+ + E^{p_i}\alpha\right)} \right\|_2^2 \\ &\quad + \beta \sum_j \frac{\left(\alpha_i - \overline{\alpha}_i\right)^2}{\lambda_j^S}. \end{aligned} \tag{21}$$

In the above formula, the left term controls the estimated residual, and the right term controls the combination coefficient α according to the probability prior. $\overline{\alpha}_i$ is the mean of the probability distribution of α_i, which is not zero when the average shape $\overline{S}^{P_i}$ is updated.

The method of alternating optimization [29] is used to reduce the energy function F. Given α, the minimum energy function F is target estimation (M_c, R, z'', P^+). Given (M_c, R, z'', P^+), the combination coefficient α is obtained by minimizing F.

After updating the position P_i' and rotation matrix R of each face feature point P_i under the camera coordinate, the deformation estimation technology of the N point mapping can be used to calculate the eigenvector combination coefficient α of the modified three-dimensional model S based on the mean of the probability $\overline{\alpha}$ of the feature vector combination coefficients. After updating S, the mean of the probability $\overline{\alpha}$ of the feature vector combination coefficients is updated. When determining the combination of the projection parameters $(\widehat{M}_c, \widehat{R}, \widehat{z''}, \widehat{P}^+)$ and the 3DMMs deformation model coefficient vector $\overline{\alpha}$, the paper establishes the solved 3D human face:

$$\widehat{S} = S_0 + \sum_{i=1} \overline{a}_i E_i^S, \tag{22}$$

and color reference relationship of faces in I:

$$c\left(\widehat{S}^P\right) \approx I\left[\left(\frac{\widehat{z}''}{\widehat{M}_c \widehat{R}_{3:}\left(\widehat{S}^P - \widehat{P}^+\right)} \widehat{M}_c \widehat{R}\left(\widehat{S}^{P_i} - \widehat{P}^+\right)\right)_{1:2}\right]. \tag{23}$$

Given the standard posture projection parameter combination of S_0 as $(\widetilde{M}_c, \widetilde{R}, \widetilde{z''}, \widetilde{P}^+)$, then a standard pose face image I^+ can be generated by the color reference relationship of the three-dimensional model of the human relative to I:

$$I^+\left[\left(\frac{\widetilde{z}''}{\widetilde{M}_c \widetilde{R}_{3:}\left(\widetilde{S}^P - \widetilde{P}^+\right)} \widetilde{M}_c \widetilde{R}\left(\widetilde{S}^P - \widetilde{P}^+\right)\right)_{1:2}\right] = c\left(\widehat{S}^P\right). \tag{24}$$

The normalized face must have symmetry, and the symmetrical point of S^P in the three-dimensional face shape $\widetilde{S}$ is $S^{\overline{P}} = \widetilde{R}^{-1} I_{-x} \widetilde{R} S^P$, where I_{-x} is the diagonalized form of $(-1, 1, 1)$. The quality $q(S^P)$ of $c(S^P)$ is inversely related to the number of references $C_{\#I}(S^P)$ of S^P in I. The number of references of S^P is at least 1, then the following correspondence can be given:

$$q\left(S^P\right) = e^{\gamma\left(1 - C_{\#I}\left(S^P\right)\right)}. \tag{25}$$

By the symmetry of the standard pose face image, the reference colors with higher quality can be given as follows:

$$c^+\left(S^P\right) = q\left(S^P\right) c\left(S^P\right) + \left(1 - q\left(S^P\right)\right) c\left(S^{\overline{P}}\right). \tag{26}$$

4.4. Dense Sampling and Preprocessing. According to the idea of local dense sampling [30], the illumination and posture correction technology has very important application value to alleviate the dependence of deep convolutional neural network on the number of samples and improve the quality of perceived weights. After completing the lighting and projection analysis of the face samples, firstly, 4 random affine transformations of a small range are performed on the light-analyzed image, and then the three Euler angles, scale adjustment parameters, and origin coordinates of the three-dimensional rotation matrix are randomly changed to generate 16 batches of dense sample images as a Mini-Batch for deep convolution model for weight training. This training method expands the sample size by 16 times, which fully alleviates the problem of the insufficient sample. Besides, due to the denseness between the same batch of samples, the probability that the perceptual model overfits the data is reduced. In the process of recognition, as face images with arbitrary lighting and attitude may appear, the paper still uses the technologies of lighting and projection analysis to analyze, correct, and then perceive the face pattern.

4.5. Attention Capsule Network Model. The proposed attention capsule network model has five gated convolution

modules. Each gated convolution module consists of two layers of a gated convolutional network and maximum pooling. Each layer of the gated convolutional network includes linear function and sigmoid activation function. Compared with the traditional CNN, the gated convolutional network replaces the modified linear unit with a gated linear unit. The learnable gate can control the amount of information passed from the current layer to the next. Gated linear units can reduce the disappearance of the gradient. It is achieved by using a sigmoid activation function to preserve the nonlinear capability of the neural network while using a linear function to provide a linear path for the gradient. The maximum pooling operation can reduce the spatial dimension of features.

The output features through the five gated convolution modules are sent to the primary capsule layer. The primary capsule layer consists of a convolution module, remodeling module, and squashing module. After the input features go through the convolutional layer, add the bias, and go through the ReLU nonlinear activation function, it is reshaped into a three-dimensional tensor with $T \times V \times U$ and compressed with squashing function. T is the time dimension before remodeling, V is the dimension inferred from other variables, and $U = 4$ is the size of the capsule. The output of the primary capsule layer has T time slices. Each time slice has V capsules, and each capsule is a tensor with $1 \times 1 \times U$.

The V capsules of each time slice are input into the advanced capsule layer. The calculation is performed between the primary and the advanced capsule layer using a dynamic routing algorithm. The dynamic routing algorithm matches V low-level capsules representing image frames with J high-level capsules representing expression categories. When multiple image frames predict the same event, the expression category of the image is determined. Then, feedback is used to increase the weight between image frames related to the image expression category and reduce the weight of image frames not related to the image expression category to learn the weights between all image frames and image expression categories accurately. With each training, the weight of the routing algorithm is updated, and the final weight is saved at the end of the algorithm. Use the dynamic routing algorithm to calculate the output vector v_j and then calculate the Euclidean length of the output vector v_j. The Euclidean length composition vectors of J categories at each moment t are used as the output of the advanced capsule layer, denoted as $o(t)$.

The V capsules of each time slice are input into the attention layer. The attention layer allows the network model to focus more on finding salient frames of the input image related to the image expression category. The sigmoid activation function of this layer can predict the importance of each frame. The output of the attention layer at each moment t is $z(t)$ and the value of $z(t)$ is between 0 and 1. The attention layer selects saliency frames while suppressing the irrelevant frame of image expression category. The time attention mechanism is realized through the output of the attention layer. Finally, the fusion layer combines the output $o(t)$ of the advanced capsule layer with the output $z(t)$ of the attention layer. The time attention mechanism is realized by selecting significant frames of time slices. Time slices with large attention factors correspond to class-related significant image frames and time slices with small attention factors correspond to class-irrelevant image frames. The final predicted output y_j is obtained by calculating the weighted sum of the output $o(t)$ and the attention factor$z(t)$ of the advanced capsule layer. y_j represents the predicted value of the J image event and the expression is as follows:

$$y_j = \frac{\sum_{t=1}^{T} o_j(t) z_j(t)}{\sum_{t=1}^{T} z_j(t)}, \tag{27}$$

where $o_j(t)$ represents the Euclidean length of the output vector v_j of the j capsule at time t, and $z_j(t)$ represents the $j$attention factor at time t, $j = 1, \ldots, J$, $t = 1, \ldots, T$. $z(t)$ controls the salient image frames of the $o(t)$ transmitted information. Choose a probability threshold τ. When $y_j > \tau$, the output is the j image activity event. The overall framework of the attention capsule network model is shown in Figure 2.

5. Experimental Results and Analysis

To verify the effectiveness of the proposed facial expression recognition method using CNN and improved capsule network model, the experimental evaluation was performed on the BU-3DFE and the JAFFE dataset. The proposed algorithm is compared with that proposed in reference [23], reference [15], and reference [19] through experiments.

5.1. Experimental Datasets

5.1.1. JAFFE. The JAFFE dataset contains a total of 213 images. 10 Japanese female students were selected, and each person made 7 expressions. In the preprocessing stage, all images are uniformly normalized to 150×110 pixels, and then feature extraction is performed on the images. Figure 3 shows example images of the JAFFE dataset.

5.1.2. BU-3DFE. The dataset covers 2D images modeled in 3D datasets of 7 typical expressions. The dataset includes 100 subjects, of which 56% are female and 44% are male. Images also vary in age, celebrity, and ethnic origin. Figure 4 shows example images of the BU-3DFE dataset.

5.2. Experimental Setup. In the proposed facial expression recognition method using CNN and improved capsule network model, the feature bin is merged by multiple subsequent residual blocks without the single-layer convolution truncated in the residual block chain (RBC). The single-layer convolution consists of 32 convolution kernels with $256 \times 7 \times 7$ and a step size of 2. In this way, the dimensions of the 8 parallel convolutional layers in the feature bin are all $32 \times 8 \times 8$. The reason for taking 8 parallel convolutional layers in the feature bin is to establish the middle-level convolutional feature of each expression. Therefore, the number of parallel convolutional layers in the feature bin

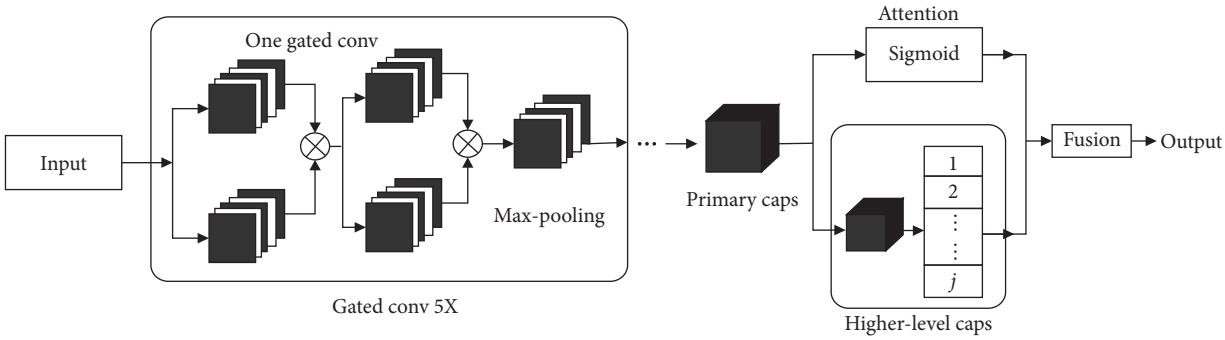

Figure 2: The overall framework of the attention capsule network model.

Figure 3: Example of the JAFFE dataset.

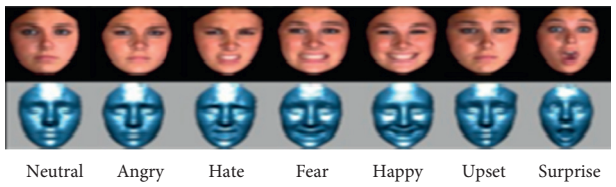

Figure 4: Example of the BU-3DFE dataset.

should be greater than and close to the number of 7 expression classifications. However, using 7 convolutional layers directly is too harsh, so the restrictions are slightly relaxed, and the number of convolutional layers in the feature bin is set to 8. The length of each class vector in the class vector bin is 16. To reconstruct the activation class vector, the activation class vector is first converted into the feature block of $8 \times 8 \times 32$ using a full link method, and then the image is restored by three distributed interleaved convolution modules. 128, 32, 32, 32 distributed interleaved convolution kernels with scales of $6 \times 6 \times 32$, $9 \times 9 \times 128$, $6 \times 6 \times 32$, $10 \times 10 \times 32$, and steps of 2, 1, 2, 2 are adopted. Width and heigth are set to the uniform scale of the model input image 128×128.

5.3. Analysis of Parameter Performance. To verify the value of parameters dynamic routing times, suppression illumination coefficient β, and combination coefficient α of the proposed facial expression recognition method using CNN and improved capsule network model, the experiments were performed on JAFFE and BU-3DFE datasets. The change range of dynamic routing times in the experiment is 1~10. The value of illumination coefficient is 0.1~1, and the value of the combination coefficient is 0.1~1. After a lot of

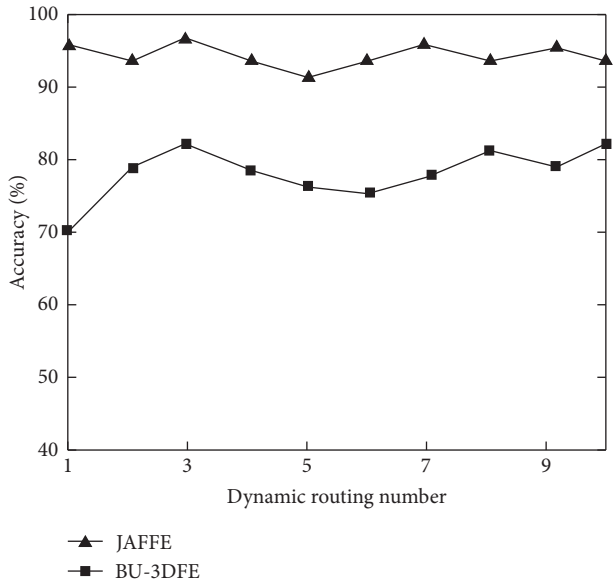

Figure 5: Performance analysis of dynamic routing times.

experiments, the most representative results are shown in Figures 5–7.

It can be seen from Figures 5–7 that when the number of dynamic routing times is 3, the recognition rates on both the JAFFE and BU-3DFE datasets have peaked. The model works best when the light coefficient β and combination coefficient α is 0.2 and 0.4, respectively. Therefore, in the following experiments, the number of dynamic routes is set to 3, the illumination coefficient is set to 0.2, and the combination coefficient is set to 0.4.

6. Results of Key Point Detection

To illustrate the key point detection of the image in this method, the proposed method is compared with existing face key point detection models. Figure 8 shows the comparison of key point detection results of reference [23] algorithm, reference [15] algorithm, reference [19] algorithm, and proposed method on JAFFE and BU-3DFE datasets.

It can be seen from Figure 8 that the proposed method is better than the three existing methods of reference [23] algorithm, reference [15] algorithm, and reference [19] algorithm significantly. It shows that when using the original face image for training in facial expression recognition, there will be a large error using rough geometric constraints as the real face key points. Therefore, mapping the features to high dimension space by random partitioning is needed.

6.1. Effect Verification of Illumination and Posture Preprocessing. To verify the effects of illumination and posture preprocessing, training was performed on the JAFFE and BU-3DEFE datasets, and validation was performed on the CK+ dataset and multi-PIE datasets. In each process of sampling, a single expression image is combined with illumination and posture processing technology to convert the camera projection perspective and illumination condition, and then, 32 training data of Mini-Batch are generated. The experimental results are shown in Table 1.

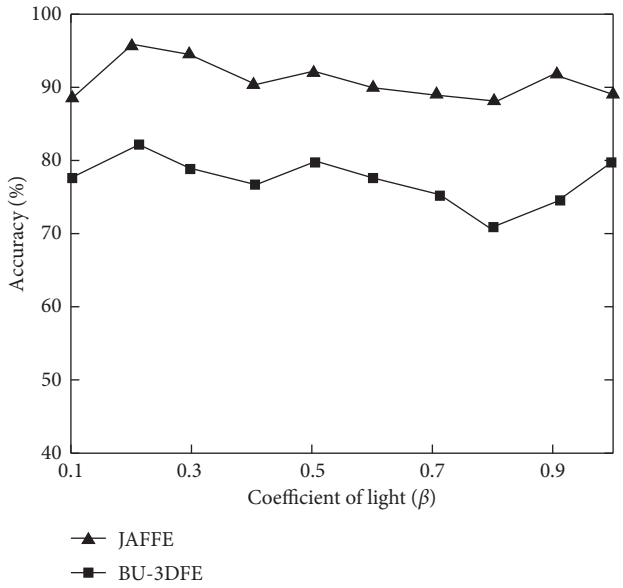

Figure 6: Analysis of the value of the illumination coefficient.

It can be seen from Table 1 that after the preprocessing of illumination and posture, the accuracy of cross-dataset recognition of several deep learning methods has been greatly improved. The results verify the effectiveness of the proposed illumination and posture normalization method.

6.2. Recognition Result. To verify the effectiveness and superiority of the algorithm in the paper, Tables 2 and 3 show the recognition rates of various expressions of different pose angles on the BU-3DFE and multi-PIE datasets. Figures 9 and 10 show the confusion matrix with the best performance

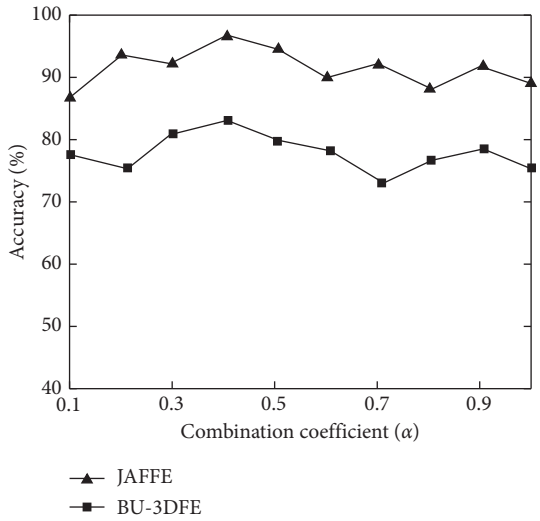

Figure 7: Analysis of the value of the combination coefficient.

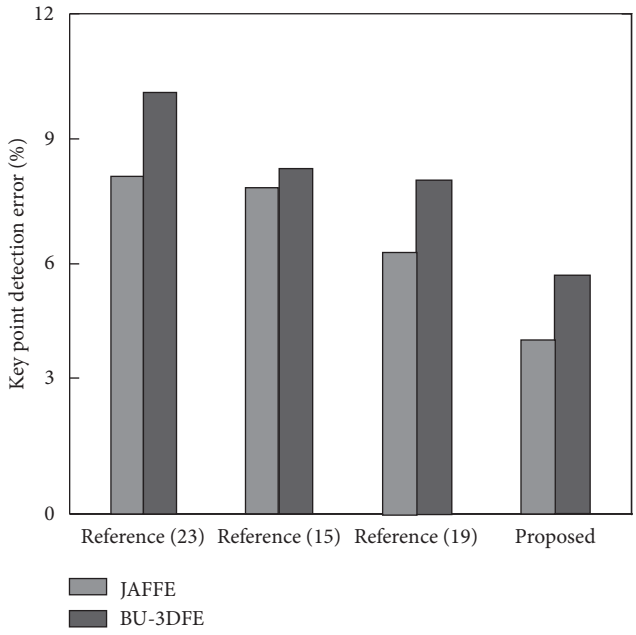

Figure 8: Comparison with existing key point detection methods on JAFFE and BU-3DFE datasets.

Table 1: Comparison with existing methods on the CK+ and multi-PIE datasets before and after illumination and posture preprocessing (%).

Methods	Preprocessing	CK+	Multi-PIE
Reference [23] algorithm	No	55.33	44.53
	Illumination	57.26	49.15
	Illumination and posture	59.16	52.62
Reference [15] algorithm	No	54.41	45.43
	Illumination	57.28	48.29
	Illumination and posture	61.05	55.17
Reference [19] algorithm	No	59.67	56.24
	Illumination	62.38	58.31
	Illumination and posture	63.51	60.68
Proposed	No	63.47	59.63
	Illumination	65.28	62.45
	Illumination and posture	67.31	65.31

Table 2: Expression recognition results of different postures on the JAFFE dataset (%).

Angle	−30°	−15°	0°	15°	30°	Average
Angry	98.74	97.64	96.81	97.38	98.33	97.78
Hate	96.84	94.82	91.87	95.21	97.41	95.23
Fear	98.47	97.81	97.32	98.11	99.34	98.21
Happy	96.76	96.46	94.27	97.88	97.18	96.11
Neutral	98.24	97.26	95.14	98.45	98.81	97.58
Upset	95.87	94.79	93.76	95.37	96.81	95.32
Surprise	97.58	96.51	94.28	96.94	98.14	96.69
Average	97.50	96.47	94.78	97.05	98.00	96.66

of each expression of the proposed method on the JAFFE dataset and the BU-3DFE dataset.

It can be seen from Tables 2 and 3 that the recognition rate of each expression of different poses on the JAFFE dataset has reached more than 90%. In the BU-3DFE dataset, the highest recognition rate of the single expression also reached 86.13%, and under different posture angles, the recognition rate of each expression is higher than that at 0° because the proposed method improves the smoothness of the image and reduces the distortion of the face texture to increase the recognition rate.

As can be seen from Figures 9 and 10, the accuracy rates on the JAFFE and BU-3DFE datasets have reached 96.66% and 80.64%, respectively. Hate is more difficult to identify on the JAFFE dataset and fear is more difficult to identify on the BU-3DFE dataset. The correct recognition rates are 95.23%

Table 3: Expression recognition results of different postures on the BU-3DFE dataset (%).

Angle	−30°	−15°	0°	15°	30°	Average
Neutral	83.57	81.49	80.46	81.02	84.16	82.14
Angry	85.34	82.17	81.88	82.74	86.07	83.64
Hate	80.44	77.31	76.27	76.34	80.19	78.11
Fear	77.28	75.72	74.15	75.11	76.89	75.83
Happy	79.24	78.21	77.46	78.71	81.83	79.09
Upset	82.84	80.21	80.14	80.39	83.67	81.45
Surprise	86.13	83.18	82.29	83.83	85.67	84.22
Average	82.12	79.76	78.95	79.73	82.64	80.64

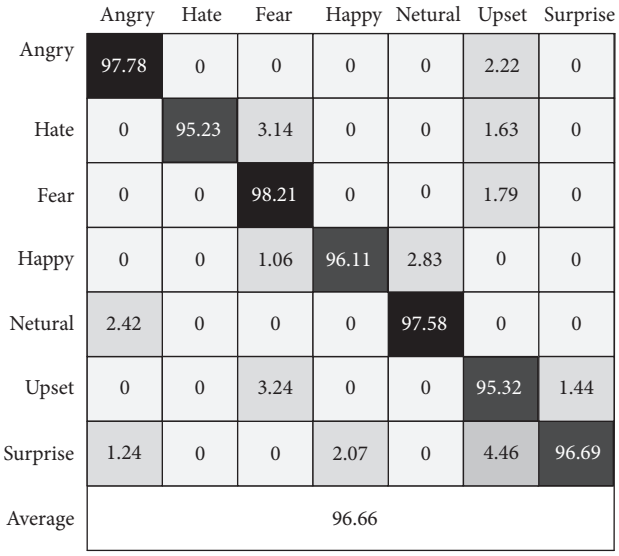

Figure 9: Confusion matrix diagram of the JAFFE dataset using the proposed method.

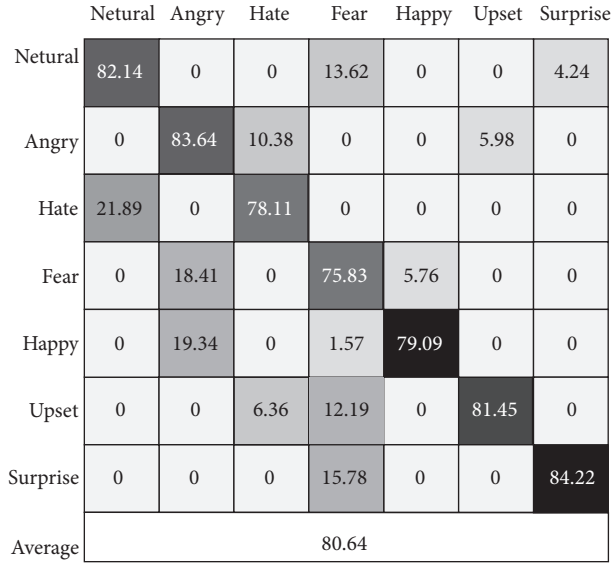

Figure 10: Confusion matrix diagram of the BU-3DFE dataset using the proposed method.

and 75.83%, respectively. The reason is that the two expressions have similar texture changes around the eyes.

To verify the effectiveness and superiority of the algorithm in the paper, a comprehensive comparison is made with the existing methods on the JAFFE dataset and the BU-3DFE dataset. During the experiment, it is best to guarantee that the training object is performed under condition independent of the test object and SVM is used as

Table 4: Comprehensive comparison with existing methods on the JAFFE dataset.

Methods	Angry	Hate	Fear	Happy	Neutral	Upset	Surprise	Average
Reference [23] algorithm	91.28	91.07	94.68	90.23	93.16	89.24	91.08	91.53
Reference [15] algorithm	93.47	92.41	95.87	92.18	94.29	91.17	93.33	93.25
Reference [19] algorithm	96.34	93.27	96.82	95.18	95.27	93.17	94.57	94.95
Proposed	97.78	95.23	98.21	96.11	97.58	95.32	96.69	96.66

Table 5: Comprehensive comparison with the existing methods on the BU-3DFE dataset.

Methods	Natural	Angry	Hate	Fear	Happy	Upset	Surprise	Average
Reference [23] algorithm	79.28	78.21	74.39	70.15	75.12	75.91	79.84	76.13
Reference [15] algorithm	79.64	80.14	75.28	72.18	75.64	77.93	80.15	77.28
Reference [19] algorithm	80.25	81.47	77.48	75.49	77.21	80.39	82.68	79.29
Proposed	82.14	83.64	78.11	75.83	79.09	81.45	84.22	80.64

the classifier. The experimental results are shown in Tables 4 and 5.

It can be seen from Tables 4 and 5 that under the same classifier, the proposed method can obtain a higher recognition accuracy rate than several other expression recognition methods. The reason is that the proposed method extracts light-insensitive features fully, suppresses boundary mark at sudden changes of light, and reduces noise points of image. Meanwhile, mapping features to high-dimensional space through random partitioning helps to distinguish similar-looking expressions. Therefore, the proposed method can improve the recognition accuracy of CNN models effectively.

7. Conclusion

A new facial expression recognition method improved the capsule network model is proposed, which reduces the noise of the image by adaptive preprocessing of the image illumination, reduces the complexity of the model, and improves the accuracy of model by using random partitioning. The improved model adds an attention layer after the primary capsule layer in the capsule network. It can increase the attention to the salient parts by weighting. That is, it can automatically select the most relevant important frames of the audio event class and ignore the irrelevant frames (such as background noise). Our attention layer realizes the attention mechanism by selecting the saliency of time slices. Thus, the overfitting of the model is reduced. Experimental results show that the improved capsule model can effectively classify facial expressions under unconstrained conditions, which makes up for the deficiency of pure deep convolution network in acquiring sparse features hidden in discriminative texture and improves the generalization ability of existing expression classification models for illumination and pose differences.

In the future task of facial expression recognition, it is planned to integrate the attention matrix into the attention capsule network. The attention capsule network is used for weakly labeled semisupervised expression image detection. Also, the attention capsule network is applied to other large-scale data problems with low discrimination.

Data Availability

The data included in this paper are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] X. Guo, Y. Tie, L. Ye, and J. Yan, "Identifying facial expression using adaptive sub-layer compensation based feature extraction," *Journal of Visual Communication and Image Representation*, vol. 50, pp. 65–73, 2017.

[2] E. Owusu, J.-D. Abdulai, and Y. Zhan, "Face detection based on multilayer feed-forward neural network and Haar features," *Software: Practice and Experience*, vol. 49, no. 1, pp. 120–129, 2019.

[3] Y. Li, Y. H. Wang, J. Liu, and W. Hao, "Expression-insensitive 3D face recognition by the fusion of multiple subject-specific curves," *Neurocomputing*, vol. 275, pp. 1295–1307, 2017.

[4] D. Kreuter, H. Takahashi, Y. Omae et al., "Classification of human gait acceleration data using convolutional neural networks," *International Journal of Innovative Computing, Information and Control*, vol. 16, no. 2, pp. 609–619, 2020.

[5] S. Zhang, G. Zhang, and X. Zhao, "Robust facial expression recognition using improved sparse classifier," *International Journal of Computer Applications in Technology*, vol. 52, no. 1, pp. 59–70, 2015.

[6] X. Zhao, X. Shi, and S. Zhang, "Facial expression recognition via deep learning," *IETE Technical Review*, vol. 32, no. 5, pp. 347–355, 2015.

[7] W. Sun, H. Zhao, and Z. Jin, "An efficient unconstrained facial expression recognition algorithm based on stack binarized auto-encoders and binarized neural networks," *Neurocomputing*, vol. 267, pp. 385–395, 2017.

[8] L. Best-Rowden, H. Han, C. Otto, B. F. Klare, and A. K. Jain, "Unconstrained face recognition: identifying a person of interest from a media collection," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2144–2157, 2014.

[9] G. Hermosilla, J. Ruiz-del-Solar, R. Verschae, and M. Correa, "A comparative study of thermal face recognition methods in

unconstrained environments," *The Journal of the Pattern Recognition Society*, vol. 45, no. 7, pp. 2445–2459, 2012.

[10] K. Sudars, "Face recognition Face2vec based on deep learning: small database case," *Automatic Control and Computer Sciences*, vol. 51, no. 1, pp. 50–54, 2017.

[11] Q. Hu and L. Zhai, "RGB-D image multi-target detection method based on 3D DSF R-CNN," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 33, no. 8, pp. 1954026.1–1954026.15, 2019.

[12] M. U. Nagaral and T. H. Reddy, "Hybrid approach for facial expression recognition using HJDLBP and LBP histogram in video sequences," *International Journal of Image, Graphics and Signal Processing*, vol. 10, no. 2, pp. 1–9, 2018.

[13] Z.-T. Liu, S.-H. Li, W.-H. Cao, D.-Y. Li, M. Hao, and R. Zhang, "Combining 2D gabor and local binary pattern for facial expression recognition using extreme learning machine," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 23, no. 3, pp. 444–455, 2019.

[14] F. An and Z. Liu, "Facial expression recognition algorithm based on parameter adaptive initialization of CNN and LSTM," *The Visual Computer*, vol. 36, no. 3, p. 483, 2019.

[15] G. C. Liu, Z. B. Yu, and Q. S. Liu, "Deeper cascaded peak-piloted network for weak expression recognition," *The Visual Computer*, vol. 34, no. 12, pp. 1691–1699, 2018.

[16] H.-H. Tsai and Y.-C. Chang, "Facial expression recognition using a combination of multiple facial features and support vector machine," *Soft Computing*, vol. 22, no. 13, pp. 4389–4405, 2018.

[17] D. Liang, H. Liang, Z. Yu, and Y. Zhang, "Deep convolutional BiLSTM fusion network for facial expression recognition," *The Visual Computer*, vol. 36, no. 3, pp. 499–508, 2020.

[18] T. B. Abdallah, R. Guermazi, and M. Hammami, "Facial-expression recognition based on a low-dimensional temporal feature space," *Multimedia Tools and Applications*, vol. 77, no. 15, pp. 19455–19479, 2018.

[19] Z. W. Luo, F. Fang, Z. Xie et al., "Conditional convolution neural network enhanced random forest for facial expression recognition," *The Journal of the Pattern Recognition Society*, vol. 84, pp. 251–261, 2018.

[20] M. U. Ahmed, K. J. Woo, K. Y. Hyeon, M. R. Bashar, and P. K. Rhee, "Wild facial expression recognition based on incremental active learning," *Cognitive Systems Research*, vol. 52, pp. 212–222, 2018.

[21] X. Ban, Y. Li, G. Yang, and L. Yang, "Multiple features fusion for facial expression recognition based on ELM," *International Journal of Embedded Systems*, vol. 10, no. 3, pp. 181–187, 2018.

[22] A. Saha and S. N. Pradhan, "Facial expression recognition based on eigenspaces and principle component analysis," *International Journal of Computational Vision and Robotics*, vol. 8, no. 2, pp. 190–200, 2018.

[23] J. Li, Y. Mi, G. Li et al., "CNN-based facial expression recognition from annotated RGB-D images for human-robot interaction," *International Journal of Humanoid Robotics*, vol. 16, no. 4, pp. 504-505, 2019.

[24] H. Boughrara, M. Chtourou, and C. B. Amar, "MLP neural network using constructive training algorithm: application to face recognition and facial expression recognition," *International Journal of Intelligent Systems Technologies and Applications*, vol. 16, no. 1, pp. 53–79, 2017.

[25] A. A. Nashat, "Facial expression recognition using best tree RD-LGP encoded features and HMM," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 16, no. 6, Article ID 1850047, 2018.

[26] A. Pandey and K. Ramesh, "An improved normalization technique for white light photoelasticity," *Optics and Lasers in Engineering*, vol. 109, pp. 7–16, 2018.

[27] L. Jian, J. Zequn, Z. Rui, L. Meiju, and G. Enyang, "Key point location method for pedestrians in depth images based on deep learning," *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*, vol. 36, no. 2, pp. 1143–1151, 2019.

[28] Y. F. Zhou and N. Chen, "The LAP under facility disruptions during early post-earthquake rescue using PSO-GA hybrid algorithm," *Fresenius Environmental Bulletin*, vol. 28, no. 12A, pp. 9906–9914, 2019.

[29] J. Jian, Y. Guo, L. Jiang, Y. An, and J. Su, "A multi-objective optimization model for green supply chain considering environmental benefits," *Sustainability*, vol. 11, no. 21, p. 5911, 2019.

[30] E. El-Sayed, R. F. Abdel-Kader, H. Nashaat, and M. Marei, "Plane detection in 3D point cloud using octree-balanced density down-sampling and iterative adaptive plane extraction," *IET Image Processing*, vol. 12, no. 9, pp. 1595–1605, 2018.