Hindawi
Complexity
Volume 2021, Article ID 6654455, 12 pages
https://doi.org/10.1155/2021/6654455

Research Article

Emotion Monitoring for Preschool Children Based on Face Recognition and Emotion Recognition Algorithms

Guiping Yu

Normal College, Eastern Liaoning University, Dandong 118000, Liaoning, China

Correspondence should be addressed to Guiping Yu; 133044@elnu.edu.cn

Received 27 December 2020; Revised 3 February 2021; Accepted 22 February 2021; Published 2 March 2021

Academic Editor: Wei Wang

In this paper, we study the face recognition and emotion recognition algorithms to monitor the emotions of preschool children. For previous emotion recognition focusing on faces, we propose to obtain more comprehensive information from faces, gestures, and contexts. Using the deep learning approach, we design a more lightweight network structure to reduce the number of parameters and save computational resources. There are not only innovations in applications, but also algorithmic enhancements. And face annotation is performed on the dataset, while a hierarchical sampling method is designed to alleviate the data imbalance phenomenon that exists in the dataset. A new feature descriptor, called "oriented gradient histogram from three orthogonal planes," is proposed to characterize facial appearance variations. A new efficient geometric feature is also proposed to capture facial contour variations, and the role of audio methods in emotion recognition is explored. Multifeature fusion can be used to optimally combine different features. The experimental results show that the method is very effective compared to other recent methods in dealing with facial expression recognition problems about videos in both laboratory-controlled environments and outdoor environments. The method performed experiments on expression detection in a facial expression database. The experimental results are compared with data from previous studies and demonstrate the effectiveness of the proposed new method.

1. Introduction

Emotion recognition (ER) is the process of inferring that the other person is in a certain emotional state by observing, analysing, and identifying valid information about the target's emotional state [1]. Humans communicate in their social life through two main channels: auditory and visual. The auditory channel carries language and music, and the visual channel carries facial expressions and body postures [2]. Facial expressions, as a powerful visual channel, play an irreplaceable role in conveying emotional and environmental information to humans, and facial expressions, together with speech and body posture, constitute the main communication system in a social context [3]. Automated facial emotion analysis systems aim to interpret and understand human mental activities by analysing facial expressions. The disciplines related to computer technology and artificial intelligence technology are developing rapidly, generating huge changes in society, and progressing in intelligence [4]. As computers play an increasingly important role in many fields, the need for human-computer interaction (HCI) has become increasingly strong [5]. To make human-computer interaction more natural and intelligent, the development of new technologies has also received extensive attention [6]. It is envisioned that computers can speak, listen, see, understand, and express emotions like real people to achieve natural and barrier-free communication, thus making our lives more convenient [7]. People can convey their emotions through expressions with subtlety and precision, and they can also understand the inner thoughts of others in this regard. In the future, if we want computers to achieve true artificial intelligence and serve us and produce natural and intelligent human-computer interaction with us, then they must have the ability to recognize and express emotions, and they need to have emotions [8]. Emotion recognition is a field related to artificial intelligence, which can help computers to intelligently recognize human emotions [9, 10]. As the field of emotion recognition continues to develop, increased research is being done on

emotion recognition and it has an important place in different application areas such as human-computer communication [11]. The main goal of emotion recognition systems is to interpret input signals from different modalities and use them to convey information about the emotion being interpreted. Nonverbal communication is an important issue to be considered in all these systems [12]. The anatomical structure of the facial features is an important part of the plastic arts and film and television art training. The expression of expressions must be established based on understanding the anatomical structure. The inner emotions of the characters are often expressed through the eyebrows and mouth. It is through these external manifestations that the artist reveals the inner world of the character. An important aspect of such systems is the study of the mechanisms by which humans communicate nonverbally with computers so that applications can interpret and connect with the user's emotions.

Pourshamsi et al. published a FER2013 database containing 28,709 training images [13]. Mittal first pretrained their model on the larger FER2013 training set and fine-tuned the other outdoor datasets using smaller samples [14]. Huang et al. proposed a new face enhancement network boosted deep belief network (BDBN) [15]. Expression recognition is repeated iteratively using three training phases in a unified recurrent framework. The first frame and the last three frames of each image sequence are selected in this experiment to obtain more samples from the CK + database. Numerous experiments using the CK+ and JAFFE databases have demonstrated that their framework is a significant improvement over current state-of-the-art algorithms that have been benchmarked on both databases [16]. Hülsmann et al. trained the network for facial expression recognition, extracting and combining the appearance and geometric features, and trained a video data system and automatic facial action detection [17]. Inspired by the development of this system, many researchers have incorporated facial expressions into their systems as a very successful way to understand a person's mental state. In recent years, many different methods for facial expression detection and recognition have been proposed [18]. Neural classifier-based methods for automatic facial expression recognition have achieved good results. The system is trained using many different images, including various facial poses, to improve the accuracy of the test. Neural classifier involves the computational load of emotion recognition. The researchers proposed a facial expression recognition method based on appearance and shape feature extraction, which first performs decision fusion followed by emotion detection [19]. To improve performance, local descriptors are used for the first time. Dynamic facial movements are not considered in the work. The method that uses facial elements and muscle movements to represent dynamic features eliminates the limitations imposed by methods that use static features, thus improving the correct recognition rate (CRR) [20]. In relative terms, this approach effectively reduces the processing time, yet it is not a real-time video processing method involving multiple frames.

This paper discusses an approach to facial expression recognition in videos with multifeature fusion. The potential of visual patterns (facial images) and audio patterns (speech) is explored. In solving the visual morphology problem, a new feature descriptor called HOG-3D is proposed to characterize facial appearance variations. And an effective geometric deformation feature, which is derived from the deformation changes of facial feature points, is proposed to characterize the changes of facial constructions. The role of audio patterns in emotion recognition is also explored. A multifeature fusion method with multicore learning is further employed to process facial expression recognition in a laboratory-controlled environment and outdoors, respectively. By integrating LSTM networks, a separate deep learning model is proposed: an LSTM model for video-based face verification in the outdoors and a combined deep CNN model and LSTM model to improve spontaneous facial expression recognition. This work aims to improve the classification performance of facial analysis in the outdoors by reducing the number of parameters and the number of training samples as well as the training time of the deep model. A method for facial expression recognition based on image sequences using the fusion of two LSTM models is proposed.

2. Face Recognition and Emotion Recognition Algorithm Design Analysis

2.1. Face Recognition and Emotion Recognition Algorithm Design. As two very similar research fields, expression recognition and face recognition have many commonalities, making many theories and techniques about the face recognition field be directly applied to the field of facial expression analysis [21]. The difference between the two is that face recognition mainly identifies who the person is and recognizes a larger number of categories, while expression recognition recognizes human emotions and recognizes fewer categories, and the current expression analysis categories are mainly six types of expressions, so the classifier selection of the two has a large difference and the feature extraction is also different. Figure 1 shows the general process of facial expression analysis.

Most of the current expression recognition effects are validated on publicly available face databases. First, face detection and preprocessing are performed on the original dataset images, i.e., the face regions in the images are detected and localized using computers, and then the face images are cropped to the required size, which mainly contains face localization, face alignment, grayscale, scale normalization, etc. Then, the expression features of the preprocessed face image are extracted. To avoid the large dimensionality of the extracted features, the dimensionality reduction of the expression features is also involved. Finally, appropriate classification methods are selected to classify the extracted features based on the differences between facial expression features. Face alignment is used to better eliminate the effect of different poses or views before extracting face features. These new technologies have infiltrated and merged with each other while developing themselves, and they have also infiltrated with traditional technologies, forming a variety of composite technologies, which further

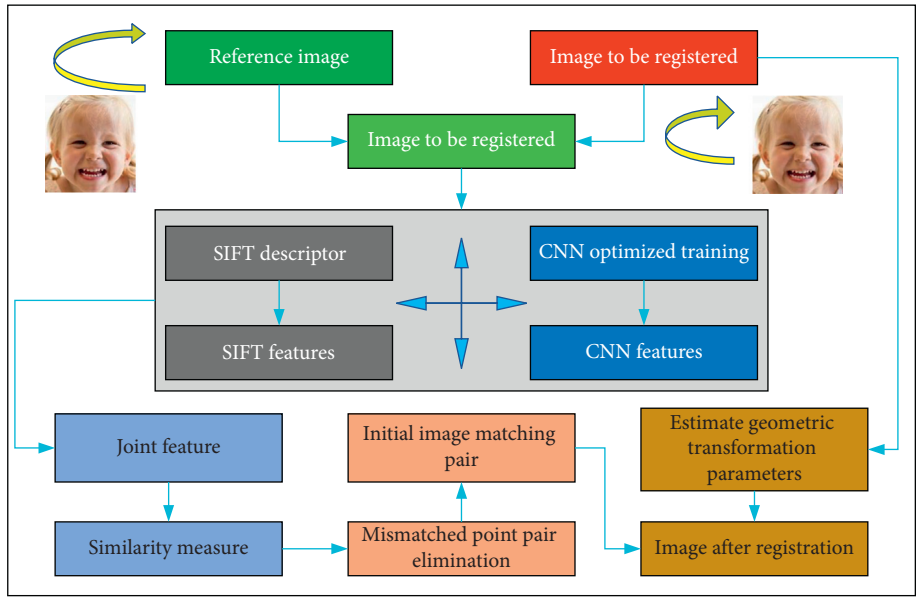

Figure 1: General flow of facial expression analysis.

promoted a new climax in the development of science and technology [22], followed by an inverse deformation of the original image. The selected points can be marks on the face contour or the centre of the eyes or nose, maintaining the robustness of the transformation without changing the facial expression. In addition to geometric normalization using affine transformation (similarity transformation), normalization can also be performed using segmental affine transformation (PWA).

After the affine transformation using the Procrustes shown in the middle column, the pixels within the triangular grid region are deformed separately to fill the reference shape. When using artificial features, PWA proved to be necessary for estimating low-intensity AUs. Although deep CNNs are invariant to rotations and translations, face normalization works well for fast convergence and avoiding overfitting problems with too-small database training. For face expression analysis in extreme poses such as rotations of 90°, it is not possible to obtain near-front faces by affine transformation or PWA. In this case, face alignment becomes the main problem for recovering a single face from an arbitrary pose into a positive face. One approach for frontal face synthesis in recent years is the two-channel generative adversarial network (TP-GAN) [23], trained to preserve both global structure and local appearance details through two codec structures, achieving better results.

$$L(x, y) = \sqrt{(x_1 + y_1)^2 - (x_2 - y_2)^2}, \tag{1}$$

$$f_i = \beta(W_i \cdot [m_i, \chi_{i-1}] + k_i + l). \tag{2}$$

The input gate is used to extract information from the candidate state.

$$C_i^j = \tanh\left(W_i \cdot [m_i, \chi_{i-1}] + k_i + l_j\right), \tag{3}$$

$$l_j = \beta\left(W_i \cdot \left[m_j, \chi_{ij-1}\right] + k_j\right). \tag{4}$$

In this paper, we propose an end-to-end dens exception network model to simultaneously predict 26 discrete categories and 3 continuous dimensions. As shown in Figure 2, dens exception consists of two parts: a feature extraction network and a feature fusion network. The feature extraction network consists of three subnetworks that extract features for face, action, and context, respectively. The feature fusion network uses fully connected fused three-way features to predict 26 discrete categories and 3 continuous dimensions.

Each subnetwork consists of 5 dens exception blocks, as shown in Figure 2. In the middle of each block, a 1×1 convolution is placed as a transformation layer to reduce half of the channels of the feature map. Each dens exception block is the basic module for feature extraction. The original image information is the input of the deformation convolution layer with 3×3 convolution kernels, so that the same object may show different sizes, poses, angle changes, and even nonrigid deformations in the image. The first three dens exception blocks are followed by a 1×1 conversion layer and a 2×2 averaging pool. The dens exception blocks are densely connected using the dense net approach.

$$L_{\text{total}} = L_{\text{disc}} + L_{\text{cont}}, \tag{5}$$

$$L_{\text{disc}} = \frac{1}{I \times J} \sum_{i=1}^{I} \sum_{j=1}^{J} w_i^J \left[\alpha_i l_{i+j} \ln\, l_{i+j} + \chi_{i-j}\left(1 + l_{i+j}\right) \ln\left(1 - l_{i+j}\right)\right]. \tag{6}$$

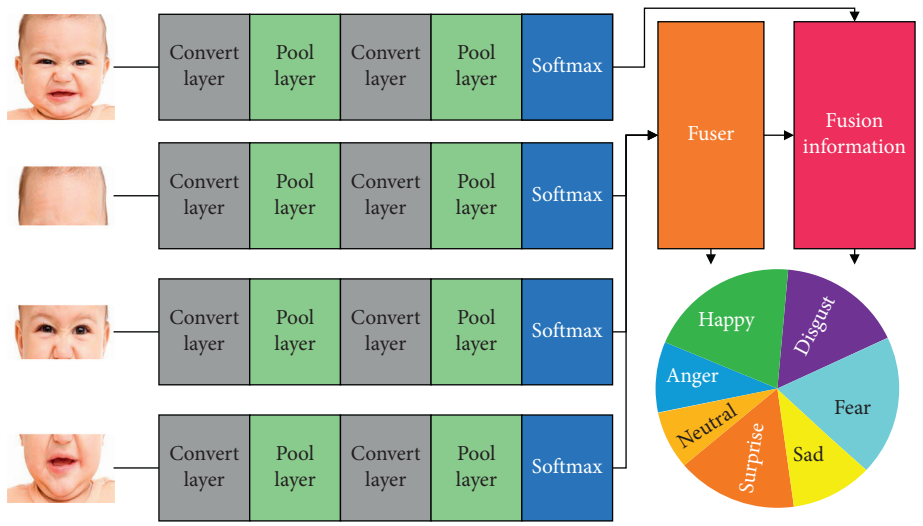

Figure 2: Dens exception network model.

The size of the original image is 64. Assuming that the output channel of the first dens exception block is N, the size of the output feature map is $6464N$, the feature map will be very large, and the computation will be very heavy. Therefore, we use the average pool after the first three dens exception blocks to reduce the size of the images from 64×64 to 8×8. This greatly reduces the computation and further improves the generalization capability of the network. The last two dens exception blocks pass only the transformation layer. At this stage, the feature maps are small enough. To retain sufficient spatial information, no pooling operation is used. After superimposing the three-way feature map, it is reduced to 26 and 3 by two full connections, and the final features are divided into 26 and 3 classes. Finally, the discrete categories are mapped into [0, 1] by the sigmoid function, and predictions are made for each category.

$$L_{\text{cont}} = \sum_{i=1}^{I}\sum_{j=1}^{J} w_i^J \left[\alpha_i l_{i+j}\log\, l_{i-j} + \chi_{i-j}\left(1 - l_{i+j}\right)\log\left(1 + l_{i+j}\right)\right], \tag{7}$$

$$G_i = \nabla_\theta J\left(\theta_t^i - \theta_t^j\right), \tag{8}$$

$$\theta_t^i = \theta_t^j + \frac{m \cdot G_i}{\sqrt{\sum_{i=1}^{I} G_{ij}^2}}. \tag{9}$$

The input of each layer in dense net comes from the output of all previous layers, that is, for layer L, there will be $L(L+1)/2$ connections in dense net. These 5 layers of exception have 15 connections and a growth rate of R. The number of output features of the dens exception block corresponding to the input feature N is $N+5\times R$. This densely connected approach allows each layer to receive the gradient of the loss function directly from the feature map of the original input, resulting in an implicit deep supervised learning. This structure has the advantages of narrow network scope, few parameters, efficient transmission of gradient and feature information, and easy network training. In convolutional networks, the deeper the network is, the more likely the gradient disappearance problem occurs. Each layer of dense connections directly connects the input and loss functions, alleviating the problem of gradient disappearance and making the network deeper.

2.2. Design of Emotion Detection System for Preschool Children. Image preprocessing plays an important role in the expression recognition system. If the image quality delivered to the expression recognition system is too low or contains a lot of noise, the accuracy of the expression recognition system will be greatly reduced. Usually, the face images after the face detection step are not directly used for expression and emotion recognition due to the problems of lighting, face angle, face pose, different image size, etc. Before performing steps such as expression recognition, the detected images usually need to be preprocessed [24]. The main purpose of image preprocessing is to improve the image quality and thus the accuracy of the algorithm. In this paper, the following kinds of image preprocessing are made: due to the influence of lighting in the image at the time of the acquisition, different brightness and darkness are easy to appear in the image, as shown in Figure 3, and if the grayscale distribution is narrow, it will affect the image contrast and lead to the lack of clear details in the graphics. This will affect the subsequent work, such as feature extraction. To improve the clarity of the image, it is necessary to make the difference of the image grayscale values larger. There are at least 21 types of human facial expressions. In addition to the common 6 types of happiness, surprise, sadness, anger, disgust, and fear, there are 15 types of compound expressions that can be distinguished, such as surprise (happy + surprise) and sadness (sad + anger).

In the original images, there may be some tilted faces. If these tilted faces are directly fed into the expression recognition network, the prediction results obtained often have large deviations from the real ones. The main purpose of geometric normalization is to resize the expression images to

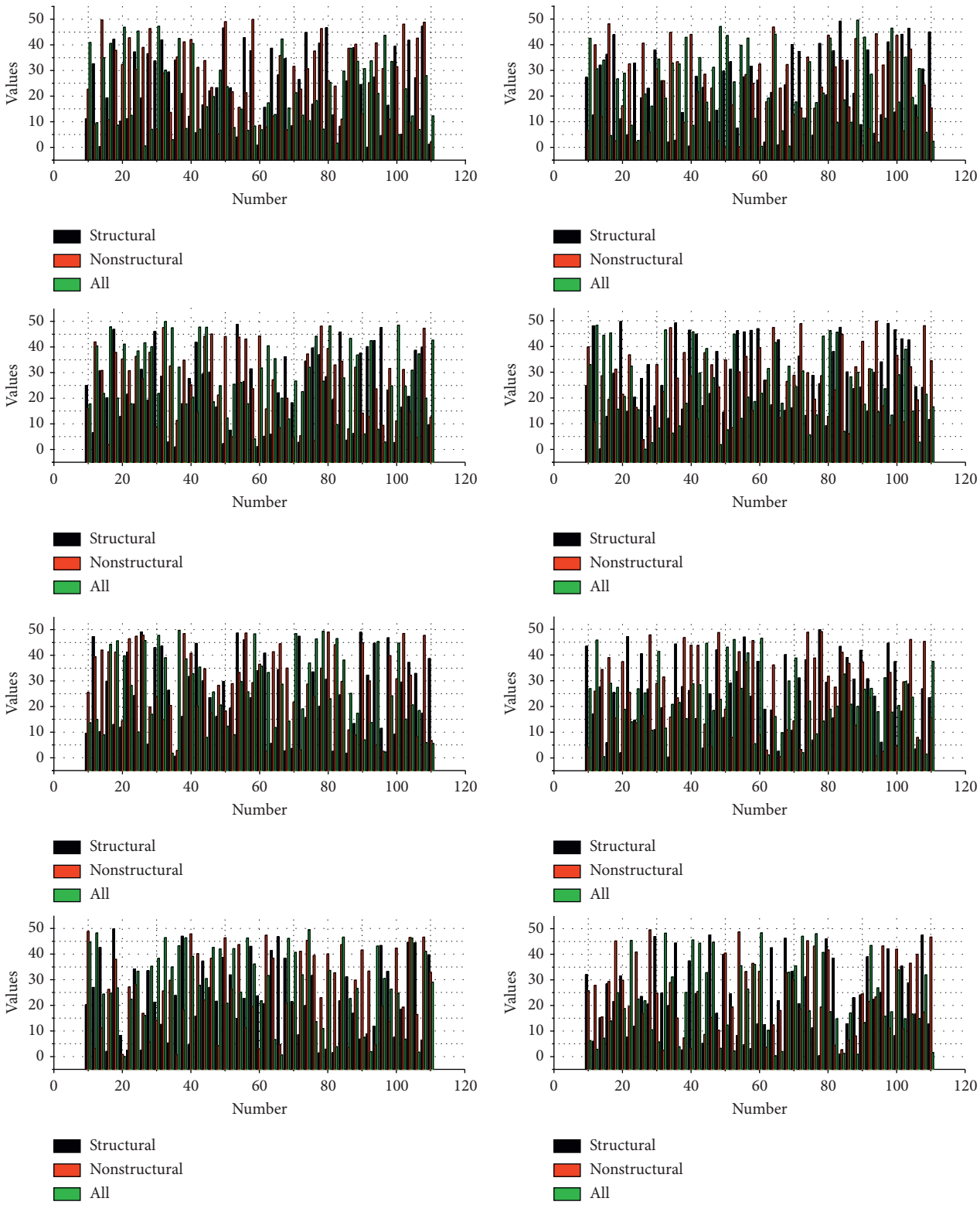

Figure 3: Histogram of images with different luminance.

a uniform size, which facilitates the extraction of expression features. In geometric normalization, the face is first adjusted to a uniform size, and then the features are aligned based on some datum points of the face such as nose, mouth, and so on. From this, we can see that the geometric transformation process of a face image can be roughly

divided into the following two steps: firstly, the image is spatially transformed, and these transformations include rotation, uniform size, translation, etc.

The workflow of the expression recognition module of this system is shown in Figure 3. Because of factors such as the number of colour channels of the images and the differences in the skin tones of the faces, preprocessing and face detection of the faces are needed first before face expression recognition. To make up for the lack of data, data enhancement is performed on the dataset in this paper. Before inputting the face images into the expression recognition network, the face images need to be greyed out and histogram equalization, face correction, and cropping of the useless regions in the face images must be carried out. In the training phase, random rotation and horizontal flipping, as well as random horizontal mirroring, are performed on the input expression recognition network to expand the number of data samples.

The FLAW dataset was used to train the face recognition algorithm model. FLAW is a database compiled by the Computer Vision Laboratory at Massachusetts State University, Amherst, USA, which was created specifically to study the face recognition problem in unrestricted environments [25]. The dataset contains more than 13,000 face images, with each face image of a different identity tagged with a different name. All face samples are collected from the network in a nonlaboratory environment and are well suited for the first basic training of face recognition algorithm models. The characteristic line segment can be generated interactively or automatically detected. The image deformation uses the reverse mapping algorithm of bilinear interpolation. Real-time face image capture detection in video surveillance scenarios has greater challenges than static image face detection. Firstly, it is difficult to guarantee the accuracy of the sample: the movement of the object, the delay of the image acquisition equipment, and other factors cause the face in the acquisition image to be blurred; then, there is the complexity of the environment: the face in the image in the video surveillance environment is difficult to be detected by the face in the acquisition image due to the influence of the brightness change of the environment, the influence of the occlusion in the middle of the crowd, the contrast change caused by the hardware factor of the acquisition equipment, the colour change, and other influences. Secondly, it is difficult to accurately detect faces in the captured images due to the relatively small proportion of faces under video surveillance, and it is also difficult to accurately identify them based on accurate detection; lastly, the degree of cooperation with the acquisition: the detected person will not actively cooperate with the acquisition equipment to collect face images, which is different from the static face comparison environment, and it will cause the pose of faces in the captured images to have diversity, as shown in Figure 4.

The degree of blurring indicates the degree of blurring of faces in the face frame of the training set, with a value of 0 for clear faces, 1 for normal faces, and 2 for severe blurring. The degree of occlusion indicates the degree of occlusion of faces in the face frame of the training set, with a value of 0 for no occlusion, 1 for partial occlusion but not severe, and 2 for severe occlusion of large areas. When the face in the face frame appears in the more obvious top view or top view of the camera, it is regarded as an "atypical face," and the value is 1. The brightness indicates the brightness of the face in the face frame, and the value is 0 under normal condition and 1 under over bright condition.

Among them, the emotional mechanism part mainly studies the implied correspondence between emotional states and human physiological behaviours and physiological characteristics, and its underlying theories are emotional psychology, cognitive science, etc. Emotional arousal is often accompanied by changes in multiple physiological traits or behaviours, and conversely, a combination of multiple emotions may cause specific human behaviour and physiological trait. The relationship between human behaviour and physiological characteristics and emotions is very complex. To accurately describe human emotions, we need to uncover the correspondence between physiological characteristics and emotions. The emotion mechanism is the basis for emotion calculation and recognition.

The controller processes the information entered by the user. Responsible for reading data from the view, controlling user input, and sending data to the model, it is the part of the application that handles user interaction. It is responsible for managing the control of the interaction with the user [26]. It receives requests from users and converts their data into parameters and then calls the corresponding functions in the model to perform operations to obtain the returned data and then analyses and passes it to the view for rendering (render) and finally outputs it to the user; the view and the controller together constitute the user interface, as shown in Table 1.

The images after the image preprocessing module mentioned above may store small tilt angles. To increase the applicability and richness of the number, random rotation operation and random horizontal flip, as well as random horizontal mirror operation, are performed on the images input to the expression recognition network in this paper. In the field of classification, many metrics can be used to evaluate the performance of a model. These evaluation metrics are not fixed, and we need to choose an appropriate, simple, and effective evaluation method based on the actual problem we are facing. The best method is to combine LSTM and CNN model algorithms. Two simple and effective evaluation metrics that are commonly used nowadays are error rate and accuracy rate. To evaluate the performance of our model M, we need to compare the predicted results of the model with the true labels.

3. Results and Analysis

3.1. Analysis of Face Recognition Results for Preschool Children. As shown in Figure 5, the model proposed in this paper converges after 150 periods of CK + data. The running time of the model is 2 minutes. Ten cross-validations resulted in an overall facial expression recognition rate of 97.8%. Figure 5 shows the confusion matrix for the CK + database. As shown in Figure 5, the prediction accuracy for fear, anger, and smile expressions is in the 100%

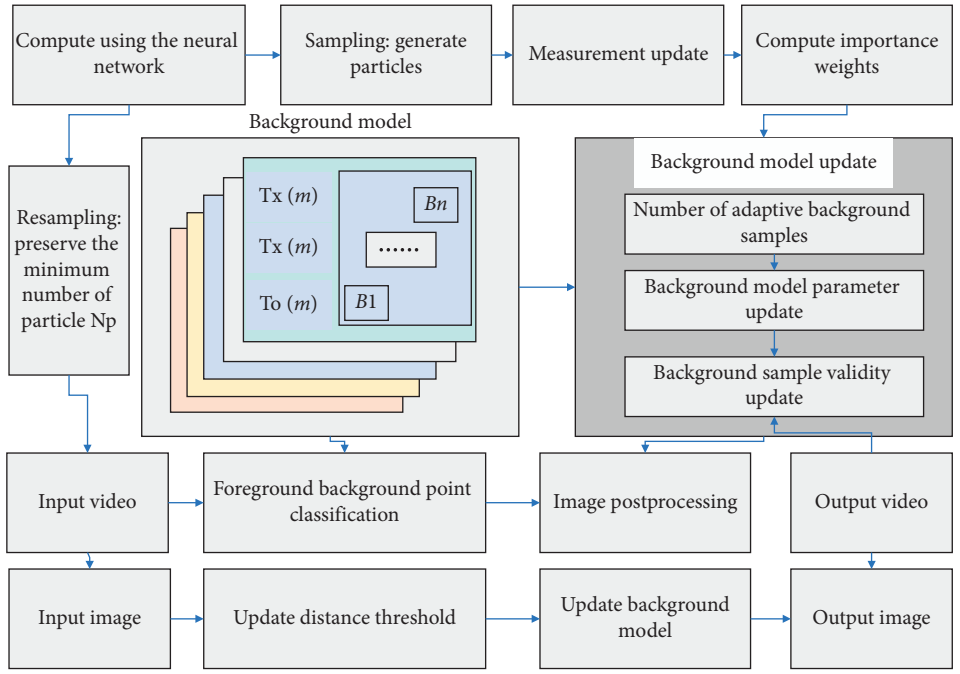

Figure 4: Network structure diagram.

Table 1: Division of the dataset.

Expression	Total number of images	Number of pictures used for training	Number of pictures used for testing
Angry	200	160	40
Disgust	300	240	60
Fear	400	320	80
Happy	200	160	40
Sad	100	80	20
Surprise	120	96	24
Contempt	140	112	28

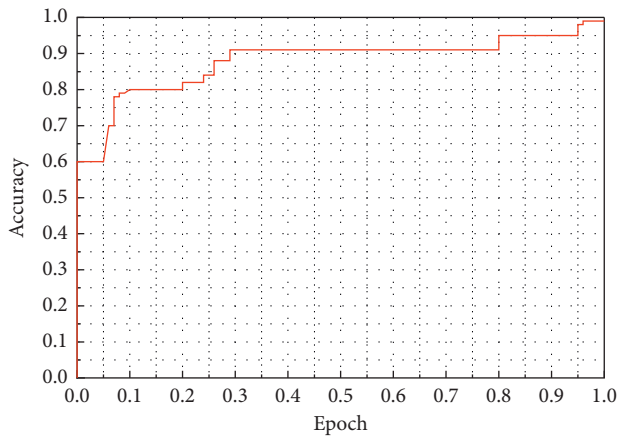

Figure 5: Training accuracy.

range, while the prediction accuracy for disgust expressions is in the lower range (less than 90%). To enhance the performance of the joint appearance and geometric features, the sequential expression recognition performance was obtained for the appearance (LSTM1 output) and geometry (LSTM2 output) features, respectively. The accuracy of the LSTM1 model (appearance model) and the LSTM2 model (geometry model) was lower than that of the joint model proposed in this paper.

The performance of the model is compared with state-of-the-art methods, as shown in Figure 5. The method proposed in this paper shows extraordinary results using two LSTMs with merged layers, and it obtains new benchmarking results relative to existing video-based methods. It can be concluded that such an LSTM model improves accuracy by including appearance and geometric features as sequential pattern inputs. The fine-tuned AlexNet features were extracted because of its ability to demonstrate a specific representation of each significant change in the facial image. After going through the LSTM model alone, it is also demonstrated how appearance and geometry exhibit excellent recognition performance. The method proposed in this paper achieves an average of 97.8% average facial expression accuracy on the CK + database, which is the best performance of current video-based methods, and shows significant performance in the new BP4D database. These slanted faces are sent directly to the facial expression

recognition network, but the predicted results are often different from the actual results because the facial expressions are very complicated and the training data are not much, resulting in different results. The method also achieves 76.16% accuracy on the MMI database, and the proposed method obtains the best accuracy based on the benchmark performance of this database and proves that the method proposed in this paper is very effective in monitoring facial expression changes, as shown in Figure 6.

In the current frame of target tracking, if the maximum drift between the predicted target position and the real position in terms of length and width does not exceed 15%, then the tracking is considered successful. The threshold of confidence score is set to 0.75, and the target is considered to be moved out of the image search range if it does not exceed 0.75, and the target is still in the search range if it exceeds 0.75. In the actual environment test, the combination of the face detection module greatly improves the real-time detection, and based on the principle of uniqueness of the face, the interference of the environment on the accuracy of face recognition is eliminated to a certain extent, and good results are achieved. Roughly speaking, blurred images will reduce image sharpness. This can be done by smoothing the colour transitions between pixels.

In this study, data were collected from preschool children using continuous photography, and the total sample size was 42. One day per week was chosen for the study, and the number of consecutive shots for a single person at the same time was 1500. Because there were many children, the study was divided into three weeks. The sample was processed at the end of the collection to remove image interference, and the sample size was 1000 images for a single person. The training data are used to build the model by backpropagation algorithm to repeatedly update the weights to determine the parameters of the model, the training data in this study use 500 images per person, and the validation data are used to evaluate which model is better in deep learning by using the trained model, and the best one is selected in this step; then, the test data are used to test it because if the model does not work well in the test, it is necessary to choose another model or retrain it next time. In this study, 250 images were used for each of the validation data and test data, as shown in Figure 7.

Finally, from the LMDB file of the training set, the mean file is generated by the built-in calculation of Caffe. The role of the mean file is to normalize the data by subtracting it to reduce the volatility of the sample, which can effectively improve the accuracy rate during training while subtracting the mean value usually makes the brightness of the image decrease, but the brightness is not significant for face recognition, so in this study, the mean file is used to improve the stability of the samples, and Figure 7 shows the mean values of the trained faces.

3.2. System Performance Analysis. Since the direction is not very clear at the beginning of the training, a large learning rate can quickly move in the right direction; however, as the number of iterations increases, the gradient returned will become smaller and smaller because it is close to the relative low point, and too large a learning rate will have too much kinetic energy resulting in a loss value that cannot be effectively reduced to a local minimum value, and it is then necessary to find a local minimum value by dynamically reducing the learning rate. This is where the local minima must be found by dynamically decreasing the learning rate. In this study, the learning rate is reduced by training 10,000 times and decreasing every 2,500 times, and the magnitude of each decrease is shown in Figure 8.

Then, the program regulates the weights through the training data and generates the loss values through the training data and the validation data, and the accuracy is

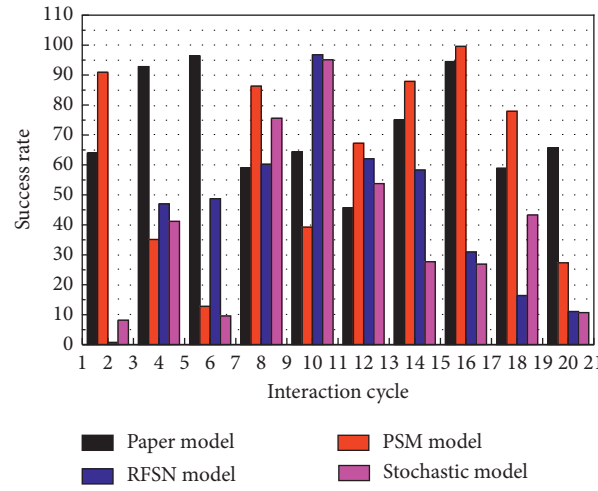

Figure 6: Algorithm accuracy test results.

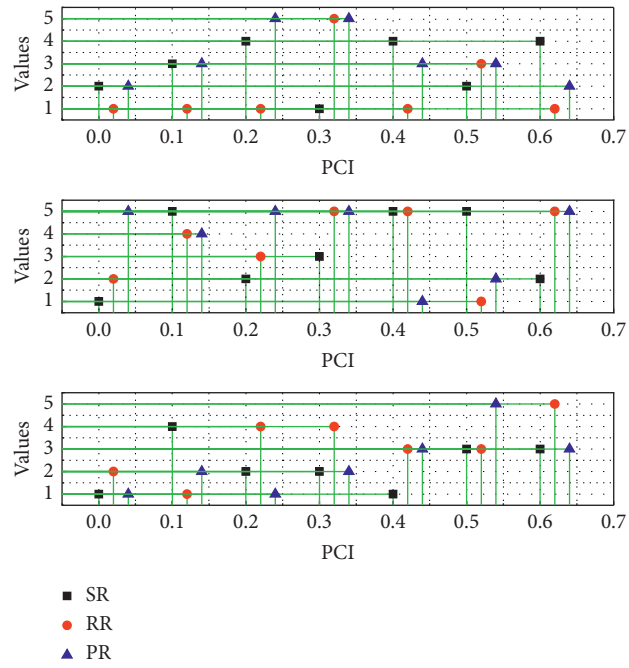

Figure 7: Comparison of databases in terms of reading performance.

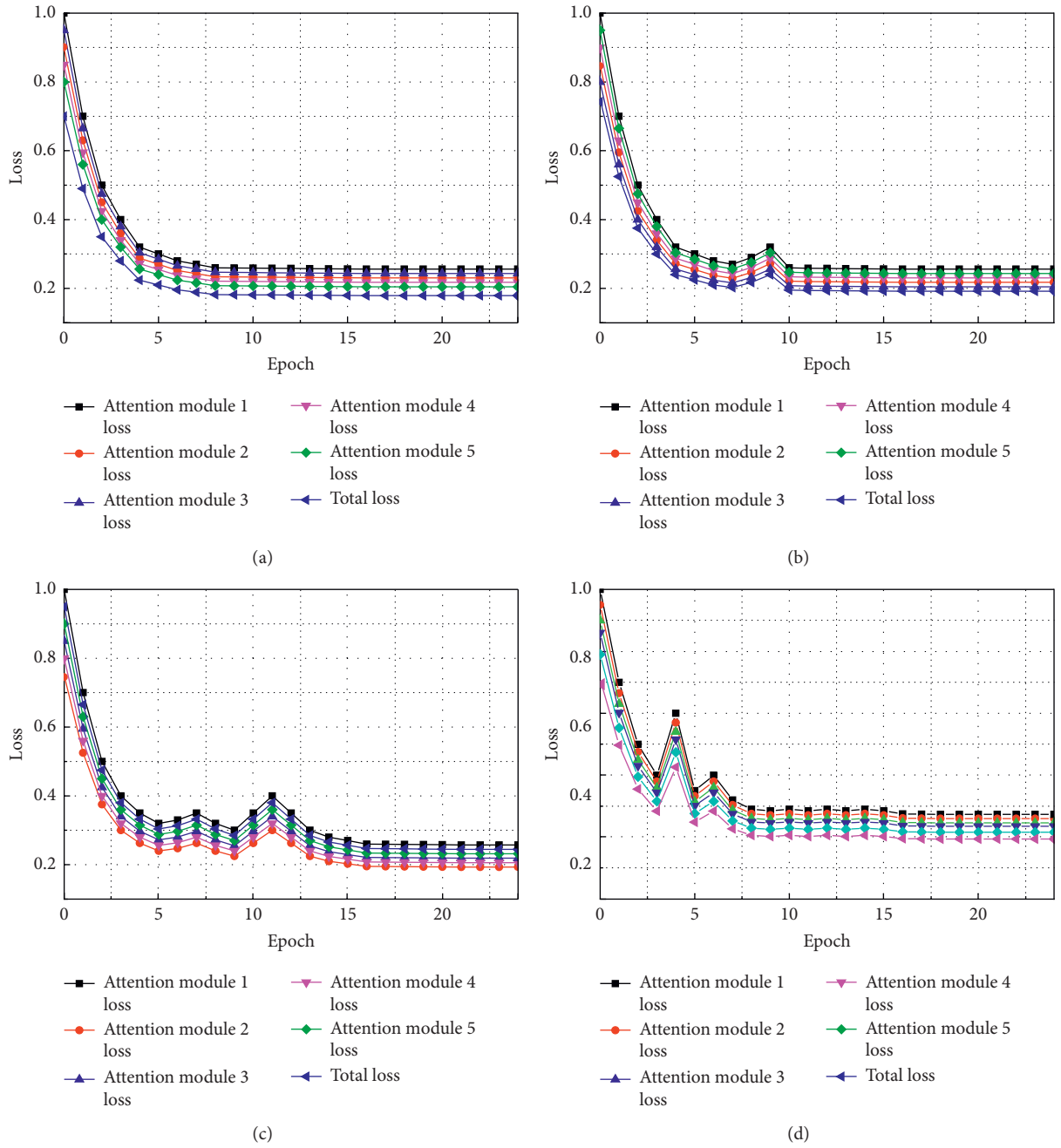

Figure 8: Learning rate change vs. training loss value line graph.

obtained by validating the model through the validation data once every 100 training overlaps. Figure 8 shows the accuracy and loss values after each training when the initial learning rate is 0.01. Due to the high learning rate, the number of steps of stochastic gradient descent is too many to stay accurately at the local minima, which causes the situation of no convergence. The validation loss in the figure is 87.3392 because it is too close to the training loss to be covered by it. Many models take a week to get the basic model, and the network cannot be trained once, and we also need to fine-tune the model for further improvement, which is only a training model, which also needs to adjust various hyperparameters, so the training time of the whole model is very long, and the cost of using GPU training is also very high. So, a major problem for deep learning is that the training time is too long. The solution includes two aspects; on the one hand, when designing the network, it is important to reasonably reduce the number of parameters and save computational resources without losing accuracy. This is also related to whether the designed network structure can be applied on the ground, as shown in Figure 9.

By comparing the training results of the two models, it can be found that both the improved VGG-19 model proposed in this paper and the typical Resnet18 network

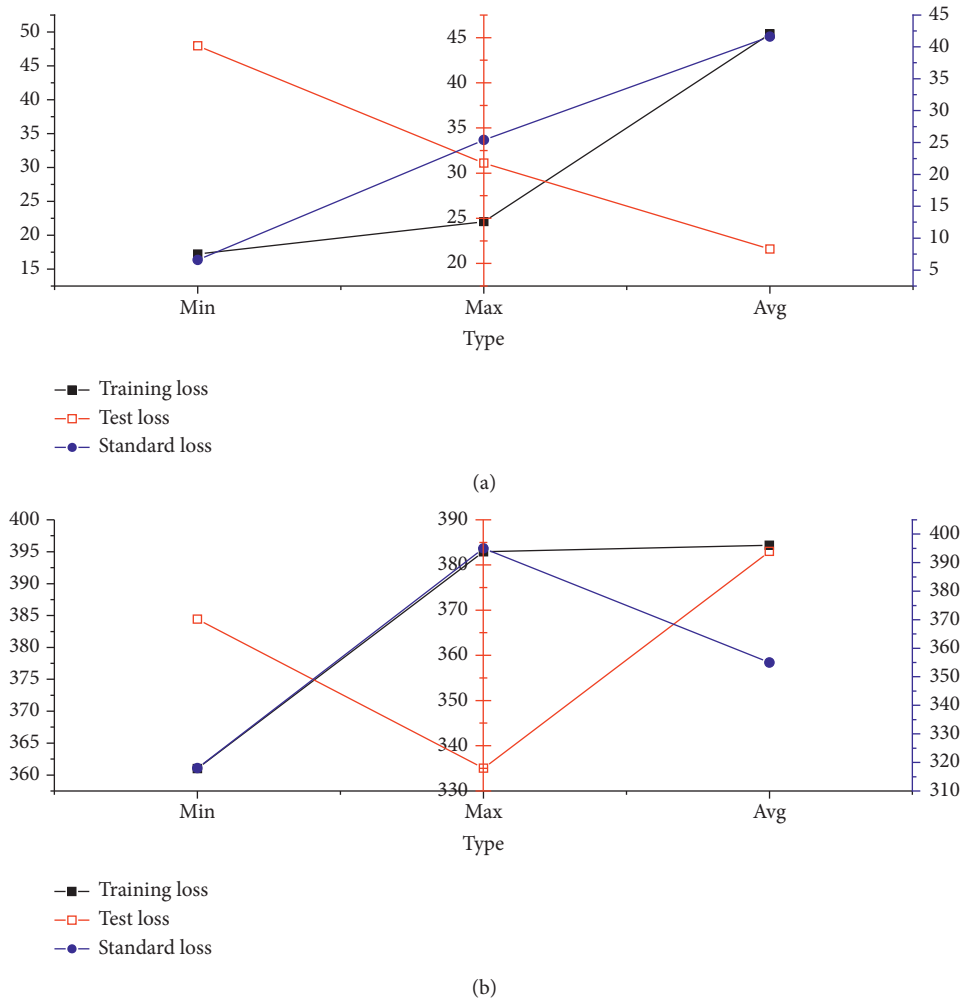

Figure 9: Performance of the improved VGG network on the homemade dataset.

perform satisfactorily, but the accuracy of the improved VGG-19 network is better. On the public test set of the face dataset FER2013, the quasi-curvature of the improved VGG-19 model proposed in this paper reaches 71.486%, and on the private test set of FER2013, its accuracy reaches 73.057%, which has surpassed the model with the highest accuracy in the FER2013 face recognition competition in 2013 (71.161%). Since the faces in the CK + database were collected under laboratory conditions, where the image quality is much better than that in FER2013, the samples are relatively easier to recognize. Therefore, the accuracy of the improved VGG-19 network on the CK + dataset is also much higher compared to that of FER2013. On the CK + dataset, the improved VGG-19 model proposed in this paper achieves an accuracy of 93.535%, and its performance does not lag compared to that of the other models. This also reflects the superiority of deep convolutional networks over traditional computer vision algorithms in the field of facial expression recognition, as shown in Figure 10.

By observing the confusion results trained by the two models, it can be found that both the improved VGG-19 network proposed in this paper and the traditional Resnet18 network have a relatively high misclassification rate when distinguishing the expressions such as sad and angry, which is because the distinction before certain expressions, in reality, is also ambiguous and difficult to distinguish. For some expressions (for example, happy and surprised), the accuracy of the two models is higher; for other expressions (for example, fear and sadness), the accuracy of the two models is lower. This is mainly due to two reasons. The first reason is that facial features become more obvious with expressions of fear, sadness, happiness, and surprise. The second reason is that the number of pictures of various expressions in the dataset is unevenly distributed, and expressions with more samples tend to have higher recognition accuracy, for example, in CK+, the number of pictures of happy and surprise tags is significantly higher than that of other tags. For example, in CK+, the number of images for happy and surprise tags is significantly higher than other

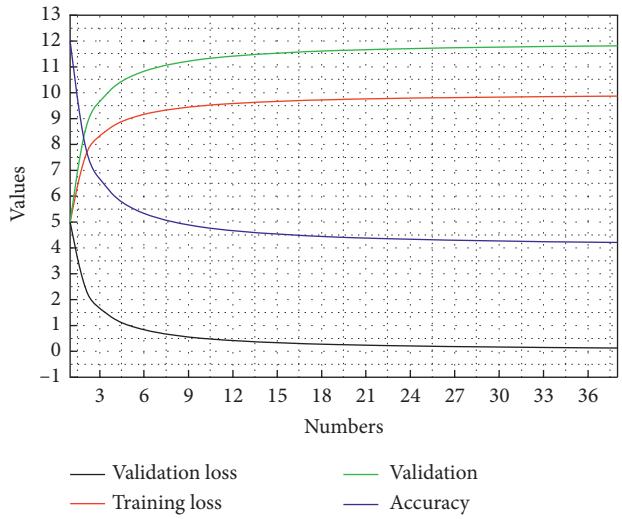

Figure 10: Training status line graph.

tags, which explains the higher recognition accuracy of happy and surprise tags than other tags. Overall, the improved VGG-19 network proposed in this paper meets the requirements of the designed system design.

From the training of deep learning to the fine-tuning of parameters to the online teaching management system built by combining face monitoring, face recognition, and emotion analysis models, this study has changed the unfairness of students' absence or absenteeism caused by clocking in time and the instability of face recognition which may be masked and cannot be recognized correctly. In addition to determining whether students are present in class, it can also confirm students' class status and learning attitude through continuous monitoring.

4. Conclusion

There are many ways of emotion recognition, such as facial expression recognition, gesture emotion recognition, voice emotion recognition, physiological pattern recognition, and multimodal emotion recognition. This paper proposes multimodal emotion recognition using a deep learning approach, which focuses on emotion recognition research combining face, action, and context. Previous traditional research methods have to perform feature extraction first and then use a suitable classification method; we use a deep learning-based approach that can perform both feature extraction and classification tasks. The current deep learning-based research methods have penetrated various fields and have applications in various tasks of images, such as image segmentation, face recognition, real-time multiperson pose estimation, object tracking, and so on. Primarily, we present datasets related to emotion recognition, various image- and video-based databases available for facial expression analysis, and human gesture analysis and verification. And the classification of the datasets in this paper is explained, as well as the data processing part of the work, introducing the problems of the emetic dataset used in this paper and the proposed processing methods. Although the face tracking algorithm can improve the immunity of the whole system to face dynamic blur, pose, angle, and other factors, it is still insufficient. Based on this problem, the paper proposes an algorithmic model for face blur detection to represent the robustness of faces in video images, which combines the face matching score output from the face detection algorithm model to represent the features of face blur and the image blur features represented by the Laplace operator filtering together to represent the face blur in video images.

Data Availability

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

Conflicts of Interest

The author declares that there are no conflicts of interest.

References

[1] L. Liu, O. De Vel, Q.-L. Han, J. Zhang, and Y. Xiang, "Detecting and preventing cyber insider threats: a survey," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 2, pp. 1397–1417, 2018.

[2] M. Azad-Manjiri, A. Amiri, and A. Saleh Sedghpour, "ML-SLSTSVM: a new structural least square twin support vector machine for multi-label learning," *Pattern Analysis and Applications*, vol. 23, no. 1, pp. 295–308, 2020.

[3] L. Jiang, L. Yan, Y. Xia, Q. Guo, M. Fu, and K. Lu, "Asynchronous multirate multisensor data fusion over unreliable measurements with correlated noise," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 53, no. 5, pp. 2427–2437, 2017.

[4] H. Wu, Z. Zhang, C. Jiao, C. Li, and T. Q. S. Quek, "Learn to sense: a meta-learning-based sensing and fusion framework for wireless sensor networks," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8215–8227, 2019.

[5] P. Ghamisi, R. Gloaguen, P. M. Atkinson et al., "Multisource and multitemporal data fusion in remote sensing: a

comprehensive review of the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 1, pp. 6–39, 2019.

[6] H. Zhang, X. Zhou, Z. Wang et al., "Adaptive consensus-based distributed target tracking with dynamic cluster in sensor networks," *IEEE Transactions on Cybernetics*, vol. 49, no. 5, pp. 1580–1591, 2018.

[7] X. Yuan and Y. Pu, "Parallel lensless compressive imaging via deep convolutional neural networks," *Optics Express*, vol. 26, no. 2, pp. 1962–1977, 2018.

[8] D. Nada, M. Bousbia-Salah, and M. Bettayeb, "Multi-sensor data fusion for wheelchair position estimation with unscented Kalman Filter," *International Journal of Automation and Computing*, vol. 15, no. 2, pp. 207–217, 2018.

[9] V. Radu, C. Tong, S. Bhattacharya et al., "Multimodal deep learning for activity and context recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 4, pp. 1–27, 2018.

[10] Q. Zhou and Y. Zheng, "Long link wireless sensor routing optimization based on improved adaptive ant colony algorithm," *International Journal of Wireless Information Networks*, vol. 27, no. 2, pp. 241–252, 2020.

[11] Z. Zhao, X. Wang, and T. Wang, "A novel measurement data classification algorithm based on SVM for tracking closely spaced targets," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 4, pp. 1089–1100, 2018.

[12] M. A. Al-Jarrah, A. Al-Dweik, M. Kalil et al., "Decision fusion in distributed cooperative wireless sensor networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 1, pp. 797–811, 2018.

[13] M. Pourshamsi, M. Garcia, M. Lavalle, and H. Balzter, "A machine-learning approach to PolInSAR and LiDAR data fusion for improved tropical forest canopy height estimation using NASA AfriSAR campaign data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 10, pp. 3453–3463, 2018.

[14] N. Mittal, U. Singh, R. Salgotra, and B. S. Sohi, "An energy efficient stable clustering approach using fuzzy extended grey wolf optimization algorithm for WSNs," *Wireless Networks*, vol. 25, no. 8, pp. 5151–5172, 2019.

[15] W. Huang, Y. Ling, and W. Zhou, "An improved LEACH routing algorithm for wireless sensor network," *International Journal of Wireless Information Networks*, vol. 25, no. 3, pp. 323–331, 2018.

[16] S. Nakayama, G. Blacquière, and T. Ishiyama, "Automated survey design for blended acquisition with irregular spatial sampling via the integration of a metaheuristic and deep learning," *Geophysics*, vol. 84, no. 4, pp. P47–P60, 2019.

[17] J. Hülsmann, J. Traub, and V. Markl, "Demand-based sensor data gathering with multi-query optimization," *Proceedings of the VLDB Endowment*, vol. 13, no. 12, pp. 2801–2804, 2020.

[18] A. Belhadi, Y. Djenouri, J. C.-W. Lin, and A. Cano, "Trajectory outlier detection," *ACM Transactions on Management Information Systems*, vol. 11, no. 3, pp. 1–29, 2020.

[19] A. Zappone, M. Di Renzo, and M. Debbah, "Wireless networks design in the era of deep learning: model-based, AI-based, or both?" *IEEE Transactions on Communications*, vol. 67, no. 10, pp. 7331–7376, 2019.

[20] A. B. Hamida, A. Benoit, P. Lambert et al., "3D deep learning approach for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 8, pp. 4420–4434, 2018.

[21] A. Farasat, G. Gross, R. Nagi, and A. G. Nikolaev, "Social network analysis with data fusion," *IEEE Transactions on Computational Social Systems*, vol. 3, no. 2, pp. 88–99, 2016.

[22] H. A. Pierson and M. S. Gashler, "Deep learning in robotics: a review of recent research," *Advanced Robotics*, vol. 31, no. 16, pp. 821–835, 2017.

[23] L. Li, K. Ota, and M. Dong, "Deep learning for smart industry: efficient manufacture inspection system with fog computing," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 10, pp. 4665–4673, 2018.

[24] A. Jalali and H. Farsi, "A new steganography algorithm based on video sparse representation," *Multimedia Tools and Applications*, vol. 79, no. 3-4, pp. 1821–1846, 2020.

[25] H. Song, J. J. Thiagarajan, P. Sattigeri, and A. Spanias, "Optimizing kernel machines using deep learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 11, pp. 5528–5540, 2018.

[26] S. De, L. Bruzzone, A. Bhattacharya et al., "A novel technique based on deep learning and a synthetic target database for classification of urban areas in PolSAR data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 1, pp. 154–170, 2017.