Hindawi
Complexity
Volume 2021, Article ID 5221950, 15 pages
https://doi.org/10.1155/2021/5221950

Research Article

Research on a Microexpression Recognition Technology Based on Multimodal Fusion

Jie Kang,[1] **Xiao Ying Chen**,[1] **Qi Yuan Liu**,[1] **Si Han Jin**,[1] **Cheng Han Yang**,[1] **and Cong Hu**[2]

[1]*College of Mechanical & Electrical Engineering, Sanjiang University, Nanjing 210012, China*
[2]*Guangxi Key Laboratory of Automatic Detecting Technology and Instruments, Guilin University of Electronic Technology, Guilin 541004, China*

Correspondence should be addressed to Jie Kang; kang_jie@sju.edu.cn

Received 9 July 2021; Revised 28 September 2021; Accepted 28 October 2021; Published 15 November 2021

Academic Editor: Kai Hu

Microexpressions have extremely high due value in national security, public safety, medical, and other fields. However, microexpressions have characteristics that are obviously different from macroexpressions, such as short duration and weak changes, which greatly increase the difficulty of microexpression recognition work. In this paper, we propose a microexpression recognition method based on multimodal fusion through a comparative study of traditional microexpression recognition algorithms such as LBP algorithm and CNN and LSTM deep learning algorithms. The method couples the separate microexpression image information with the corresponding body temperature information to establish a multimodal fusion microexpression database. This paper firstly introduces how to build a multimodal fusion microexpression database in a laboratory environment, secondly compares the recognition accuracy of LBP, LSTM, and CNN + LSTM networks for microexpressions, and finally selects the superior CNN + LSTM network in the comparison results for model training and testing on the test set under separate microexpression database and multimodal fusion database. The experimental results show that a microexpression recognition method based on multimodal fusion designed in this paper is more accurate than unimodal recognition in multimodal recognition after feature fusion, and its recognition rate reaches 75.1%, which proves that the method is feasible and effective in improving microexpression recognition rate and has good practical value.

1. Introduction

The term microexpression was introduced in 1996 by Haggard and Isaacs after an in-depth study. Subsequently, in order to be able to observe microexpressions with the naked eye, Ekman [1] developed METT (Microexpression Training Tool), a microexpression training tool. However, with the help of this tool, the recognition rate of facial microexpressions hovered around 50% at most, and the recognition was not guaranteed. In the past, due to the limitation of hardware, microexpressions were mostly studied by traditional methods for experiments, for example, local binary pattern method, optical flow method, etc.

Based on the local binary pattern referred to as LBP algorithm, its principle is mainly to convert microexpression images from RGB to LBP images, which can attenuate the effect of illumination in a small area. Jiang Wan [2] designed a shallow dual spatiotemporal multiscale neural network TSTNET model to extract the texture properties of SMIC and CASME II microexpression databases using local binary patterns, which are fed into a 3-dimensional convolutional neural network with convolutional long-short-term memory network (LSTM) to extract both temporal and spatial information The model incorporates a discard algorithm and multiplexes the extracted features to reduce the risk of overfitting while learning richer features. The recognition rates of 67.30% and 65.34% are achieved on SMIC and CASME II microexpression databases, respectively, and the model improves the training speed of the network and the recognition rate of microexpressions compared with existing deep learning methods.

The core idea of the optical flow-based method is to extract the information of the optical flow in the vertex frame and the start frame of the microexpression segment and then compare and analyse them. Wen et al. [3] proposed a combination of traditional methods and deep learning methods for the problem of low recognition rate of cross-library microexpressions, in which Apex frame localization is performed in the image preprocessing part; in the feature extraction part, the TVL1 information of Apex frames is first calculated, and then the horizontal and vertical optical flow component features are fused; finally, SVM is used to classify the features. This method has a great improvement over the LPB-TOP (local binary patterns from three orthogonal planes, LPB-TOP) method.

Since the 21st century, computer computing power has developed rapidly, and the research on microexpressions has gradually shifted to the field of CV. The focus of the research content has also gradually transitioned from traditional methods to deep learning direction. Through the analysis of current domestic and foreign microexpression research algorithms, Li [4] improved the optical flow method and convolutional neural network, completed the design of a prototype system for facial microexpression recognition, and proved the feasibility and effectiveness of the method through the comparison of experimental results. Su et al. [5] analysed the motion feature maps of microexpressions, proposed a method for feature reextraction as well as fusion of multiple motion feature maps, extracted different features and temporal features using multichannel CNN and LSTM, tested under CASME database, and achieved excellent results. Khor et al. [6] proposed a network that can handle long sequences (Enriched Long-term Recurrent Convolutional Network, ELRCN), where microexpression picture segments are encoded as they pass through the CNN network for each frame within the microexpression cycle, and then the microexpression category is predicted using a Long-Short-Term Memory Network (LSTM). Ultimately, experimental results show that the method is able to obtain fairly good performance without increasing the data. Liu et al. [7] proposed a local-based deep neural network with two domain adaptation techniques (opposing domain adaptation and motion scaling up and down) that can automatically learn to extract distinguishing features associated with the face, and experiments show that the method at the Second Microexpression Competition (MEGC) achieved a very competitive result. In the field of microexpression recognition, deep learning methods have occupied the majority of the field. However, due to problems such as insufficient data samples, it is difficult for deep learning to bring out its real strength. How to solve this series of problems has become a challenging and difficult work in the field of microexpression recognition.

In this project, a microexpression recognition method based on multimodal feature fusion is proposed after investigation and experimental research, and a microexpression database is established by ourselves. The feasibility and rationality of the designed microexpression recognition method are verified through experiments. Firstly, the volunteer's body temperature was recorded simultaneously with the infrared thermometer while collecting the microexpression fragments, and the multimodal fusion database was established after the image data was organized, and then the data was preprocessed; then the training set and test set were divided according to 6 : 4, and the data was fed into the CNN + LSTM dual-channel neural network for training, and the experimental results were obtained. Finally, the training results are analysed and compared with those using only the separate microexpression database under the same network.

2. Related Work

2.1. Database Establishment. Since microexpressions were first discovered in the 1960s, along with the development of microexpression psychology and the advancement of computer image analysis technology, the research on microexpressions has made great progress. At present, many teams at home and abroad are working on microexpressions, mainly the teams of Ekman, Matsumoto et al. [8], and Shreve et al. [9] in the United States, Porter et al. [10] in Canada, Zhao and Pietikainen [11] in Finland, and Wu et al. [12] and Luo et al. [13] in Fudan University in China. It should be emphasized that the sample sizes of these databases are very small. There are less than 800 publicly published microexpression samples so far, which is a typical small sample problem. This causes that the current deep learning-based methods cannot fully play their power in the microexpression problem. In fact, it is very difficult to build a database of microexpressions. One reason is that microexpressions are difficult to elicit; researchers often ask subjects to watch emotional videos that elicit their emotions while asking them to disguise their expressions. Some subjects may not show microexpressions or may show them only rarely.

Based on many difficulties of traditional database establishment, this paper proposes and builds a microexpression database based on multimodal feature fusion based on the summary of previous research results. We construct the database by fusing the microexpression image information with the corresponding thermographic temperature text information with multimodal features. The method of coupling multimodal information is applied to the field of microexpression recognition, which is a new attempt. The physiological signals and movements of human body can be coupled to make judgments, but repeated experiments are needed to prove the feasibility of the method.

2.2. Microexpression Recognition Method. Current research on microexpression recognition has focused on optical flow-based algorithms [14], texture feature-based algorithms [15], and deep learning-based algorithms [16]. Most optical flow-based algorithms use dynamic optical flow features to describe the differences in facial expression changes. Optical flow mapping techniques use optical flow change features between consecutive images to analyse the change pattern of microexpressions and then construct an optical flow map of the whole microexpression process as the basis for

microexpression differentiation. The Fast Hyperspectral Optical Flow (FHOF) technique [17] extracts the changes in the parts related to microexpressions by correcting the optical flow method and feeds the optical flow features of the same parts into the classification frame for detection, which effectively excludes the interference generated by changes in facial muscles. The facial dynamic map (FDM) technique [18] aligns the local pixels of microexpressions between images by optical flow method, so as to accurately extract the dynamic change trajectories of each part of the face. The optical flow-based algorithm has the advantages of low computational complexity and easy implementation, but it requires a priori information, such as peak images or positively defined samples, and has some drawbacks in terms of recognition accuracy. Most texture-based microexpression recognition algorithms analyse the differences between facial microexpressions through texture features. Spatiotemporal Gabor (ST-Gabor) filter [19] uses Gabor filtering to extract texture information from images and optical flow techniques to extract features of microexpressions in the time domain. Reisz Phase [20] uses Reisz spectrum theory to detect the peak of microexpressions from the phase information of the image, and this method effectively solves the problem of low importance of microexpressions.

In recent years, deep learning techniques have been extensively studied in problems such as pattern recognition and image processing. Three-Dimensional Fully Connected Dimensional Convolutional Networks (3D-FCNN) [21] designed a 3D convolutional neural network to extract the spatiotemporal features of microexpressions and then input them to a classifier for classification. Using deep convolutional neural networks for learning spatiotemporal domain features of video sequences requires an extremely long training period. Visual Geometry Group Network (VGGNet) [22] uses a convolutional neural network to learn only the local features of microexpression spike images, thus reducing the training time of the network.

After analysing the existing microexpression recognition techniques, it is found that convolutional neural networks (CNN) combined with long-short-term memory (LSTM) artificial neural networks have certain advantages in extracting facial spatial features and temporal features. In this paper, a microexpression recognition method based on multimodal fusion is proposed. The temperature features are passed into the CNN + LSTM neural network as input in the form of text to obtain the feature vector after network convolution; then the microexpression feature vector obtained by the same neural network output is fused with the temperature data features, and after repeated iterations, a new feature vector interval is obtained, and the k-nearest neighbor method is used to classify the fused features to obtain the training model. Finally, the experimental results are tested on the test set of the multimodal database and then compared with the models trained using the microexpression database alone to complete the comparison of the experimental results.

3. Materials and Methods

In recent years, great progress has been made in the combination of psychology and computer technology in the research of microexpression recognition technology. However, the formation mechanism of microexpression is special and the research started late; in particular the establishment level of microexpression database is relatively backward. In the establishment of the microexpression database, there are some problems, such as small number of samples (the largest published database sample number is 247), incomplete types (the number of some emotion samples that are difficult to be captured is very small), and inconsistent acquisition standards (different experimental environments and methods).

Therefore, on the basis of summarizing the previous experience, this study improved the experimental method and combined with the characteristic that the temperature of a specific part of the body changes with the change of human emotions proposed by Lauri Nummenmaa et al. from Aalto University. A microexpression database based on multimode fusion is constructed by fusing temperature data with microexpression image data.

3.1. Existing Microexpression Database. The microexpression database can be divided into active database and passive database according to the induction mode. The active database allows participants to watch emotional videos or pictures and enables participants to generate microexpressions after being stimulated. The collection method of passive database is to ask participants in the experiment to make well-set and weak expressions, but it is difficult for this method to simulate the real and natural emotions of human beings, and it is difficult to meet the training requirements. Table 1 shows 5 public spontaneous microexpression databases.

The CASME II dataset, created by the Chinese Academy of Sciences et al., used a 200 fps camera to capture microexpression clips, with a high frame-rate camera designed to capture subtle changes in the participants' faces. A total of 26 participants took part in the experiment and collected 255 microexpression clips of seven emotions, including disgust, happiness, surprise, sadness, fear, depression, and others. Davison et al. set up the SAMM dataset and used a 200 fps frame-rate camera to capture 195 microexpression clips from 32 participants. SAMM contains seven emotions: anger, contempt, fear, disgust, sadness, joy, and surprise.

It is important to emphasize that because microexpressions range of motion is very small, and relatively regular expressions often face local movement, because the face database on sentiment classification is not very clear, database of the mood of the calibration standard is different, often similar movement was as different kinds of microexpression, and different sport is viewed as a kind of expression. This situation leads to the inconsistency of the results obtained by using various databases to train the microexpression recognition algorithm. In addition, due to the short duration, low intensity, and often local movement of microexpressions, the video quality of many current microexpression databases cannot meet the needs of microexpression recognition and analysis, which requires us to make certain improvements to the database.

Table 1: Existing database of spontaneous microexpressions.

	CASME	CASME II	CAS (ME) 2	SAMM
Number of samples	195	255	53	159
Frame rate	60	200	30	200
Resolution	640 ∗ 480	640 ∗ 480	640 ∗ 480	2040 ∗ 1088
Number of races	1	1	1	13
Number of categories	8	7	N/A	7

3.2. A Microexpression Database Based on Multimodal Fusion. After analysing the advantages and disadvantages of the microexpression database listed in Table 1, we gathered 13 volunteers to participate in the establishment of a microexpression recognition database based on multimode fusion. Next, we will explain the establishment process of the microexpression database.

3.2.1. Microexpression Image Acquisition. Mind Vision industrial camera (MV-UBS31GC) was used to record 255 microexpression clips with a resolution of 1280×720 for 13 volunteers standing in front of a 15.6-inch computer at a frame rate of 75FPS. The segment contains seven emotional categories, namely, happy, surprised, sad, angry, disgusted, scared, and normal. In terms of acquiring microexpression data, in order to collect more accurate microexpression data, we required volunteers to maintain facial neutralization during the test process and use 7 videos of different contents as microexpression eliciting materials. In this study, volunteers experienced high arousal and strong motivation to mask their true emotions. But they were asked not to move their eyes or look away from the screen, to rewatch their own facial movements after the recording, and to point out any facial movements in the video that were not associated with generating emotion for subsequent analysis of the microexpression data.

3.2.2. Corresponding Temperature Data Acquisition. While the volunteers watched the video, the temperature data of each volunteer's head, chest, and shoulder were collected by the HIKMICRO-H11 handheld thermometer, and 765 temperature data corresponding to the self-built database tags were collected after screening and counting, as shown in Table 2.

3.3. Preprocessed. In the multimode microexpression recognition task, the recognition image is preprocessed, including the time domain image interpolation, face detection, feature point location, data expansion, and other steps. The preprocessing of microexpression data is beneficial to the subsequent data feature extraction and classification.

3.3.1. Time Domain Image Interpolation. Under natural conditions, microexpressions are naturally and continuously changing. But each frame of the image contained in the video clip captured by the camera is discontinuous. If we can find a function to fit this continuous line and resample the line more intensively, we can represent the same expression in more images.

Since the microexpression duration period is very short and difficult to observe, Zhou et al. proposed the TIM [23] algorithm to make the whole microexpression period longer without losing microexpression features, which first treats a video clip as a graph and uses nodes in the graph to represent a frame of the image, and adjacent frames in the video are represented in the graph as adjacent nodes as well, and frames in the video that are not the frames that are adjacent in the video are also represented as adjacent nodes in the graph, and the frames that are not adjacent in the video are also not adjacent in the graph; subsequently, the graph is embedded into a low-dimensional model using the graph embedding algorithm, and finally the graph vector is substituted to calculate this high-dimensional continuous curve. The changes that occur in microexpressions after TIM processing and the corresponding temperature and expression changes are shown in Figure 1.

We applied the TIM algorithm to extend the period of microexpression generation, and it can be seen in Figure 1 that the state of the corners of the mouth of the recruited volunteers changed between the 0 ms and 2000 ms moments, and the corresponding head temperature changed from 33.4°C to 34.6°C. This change is identical to the findings of Lauri Nummenmaa et al. of Aalto University [24], where the temperature of specific parts of the body changes when human emotions change. In addition, this has greatly facilitated the production of subsequent datasets. After finishing the video processing, it is also necessary to convert the video into images for the next step of delineating the face regions.

3.3.2. Face Detection and Feature Point Localization. Microexpressions are muscle changes produced by subtle movements of human face muscles. In order to study microexpressions more accurately, it is necessary to first perform face detection on the image, remove interference from regions other than the face, and crop out the face region. This can be achieved by using the Harr plus cascade classifier in OPENCV. The flowchart of the face detection procedure is shown in Figure 2.

Face feature point localization is to detect the shape feature points such as eyebrows, eyes, nose, and lips from the face, which are represented by 68 points. DIIB is a library in OPENCV that can quickly calculate the location of feature points.

The DIIB-based detection method first needs to get the average data of the image of the face feature points as the initial face shape and then get the pixels of the current feature points by calculating the value of the pixels of a random point in the range of the initial feature points and then doing the variance with the average. Finally, we start to construct the residual tree, calculate the size of the difference between the current feature point and the target feature point, select multiple segmentation points using methods

Table 2: Number of each mood type in the database and the corresponding temperature data.

Serial number	Microexpression type	Head temperature (°C)	Thoracic temperature (°C)	Shoulder temperature (°C)	Number of emotions
1	Happy	34.5	35.1	34.1	42
2	Sadness	34.6	34.3	33.1	20
3	Disgusted	35.1	35.2	34.7	56
4	Anger	35.2	35.5	34.2	42
5	Surprise	35.1	35.8	34.9	43
6	Fear	34.8	35.2	34.2	8
7	Normal	33.7	34.1	33.8	44

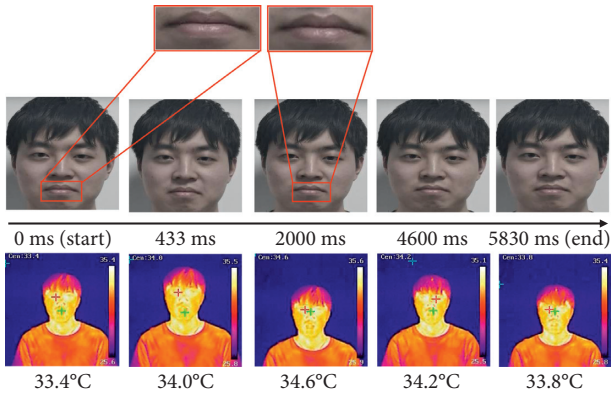


Figure 1: Changes in microexpressions after TIM treatment and corresponding temperature and expression changes.

such as annealing, perform left and right tree division, and select the point with the smallest difference as the best point. The flow of the face feature point localization procedure is shown in Figure 3, and the results of face detection and feature point extraction are shown in Figure 4, respectively.

3.4. Multimodal Fusion Microexpression Recognition Method. For multimodal image fusion methods, a lot of research has been done for this purpose in recent years [25–30]. In 2013, Nummenmaa et al. of Aalto University [24] invited a total of 700 people from all over the world and showed them movies, stories, etc. that evoked different emotions and then used an infrared thermographic camera to measure the change in body temperature in various parts of their bodies. The study found that when human emotions change, the temperature of specific parts of the body also changes. The researchers created a graph of the body temperature distribution of 14 emotions. The experimental temperatures show that each emotion corresponds to a different part of the body.

Based on the above research findings and addressing the current problems of low recognition accuracy and insufficient model generalization ability in microexpression recognition tasks, this study proposes a microexpression recognition method based on multimodal feature fusion for the first time. The method uses CNN + LSTM spatiotemporal depth network model to extract the feature values of each microexpression in a self-built multimodal fusion microexpression database and fuses them in series with the features of body temperature text data during the microexpression change cycle to extract the multimodal standard feature values of each microexpression and then selects test samples to classify microexpressions according to the standard feature values of different kinds of microexpressions. In the following, we will introduce the role of CNN and LSTM neural networks in this method, the fusion method of face feature extraction data and body temperature data, and the multimodal fusion model in detail.

3.4.1. CNN-LSTM. CNN neural network is a feedforward neural network consisting of a combination of superimposed convolutional and pooling layers, which has good learning ability for high-dimensional features and has been widely used in image processing, speech recognition, and other fields [31]. The convolutional layer is the core part in multimodal microexpression image processing, processing a large amount of microexpression image data that passes through this network; as such most of the computational effort is generated here. The pooling layer usually appears periodically between successive convolutional layers during the processing of image data, and it serves to reduce the data dimensionality and is effective in reducing the number of data parameters in the network, reducing the computational effort, and effectively controlling over data fitting. The fully connected layer is used to do weighting on the extracted microexpression features and can also act as a classifier.

LSTM (long-short-term memory network) was proposed by Hochreiter and Schmidhuber in 1997 [32] and is a network based on RNN improvement; this method can be used to solve the problem of gradient disappearance during training of long sequences of temperature text features due to gradient concatenation leading to the inability to update the parameters of the previous neurons, as the neurons between each layer of CNN are not connected. The context-dependent information of the input text is captured, while the RNN will judge the previous information by memory and apply it to the current computation. The traditional RNN is prone to gradient explosion and gradient disappearance problems; LSTM can effectively solve these problems and has become the most commonly used recurrent neural network. The network structure of LSTM is shown in Figure 5.

The LSTM cell has one more hidden state and many structures called cell state Ct and gating structures. The gating structure contains an input gate, an oblivion gate, and an output gate.

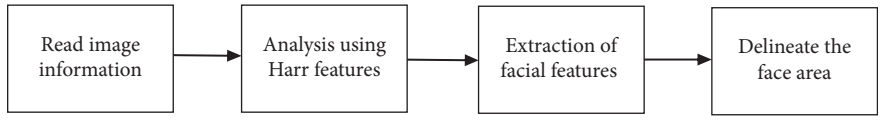


Figure 2: Flow chart of face detection program.

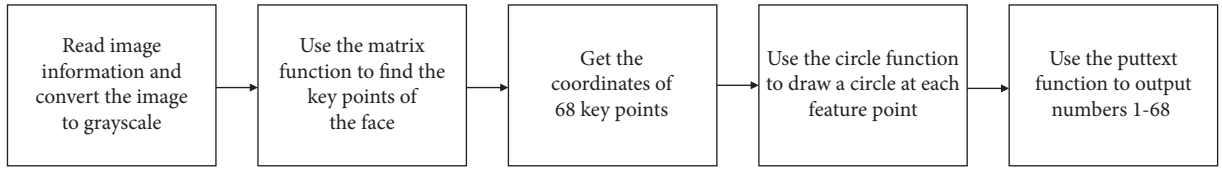
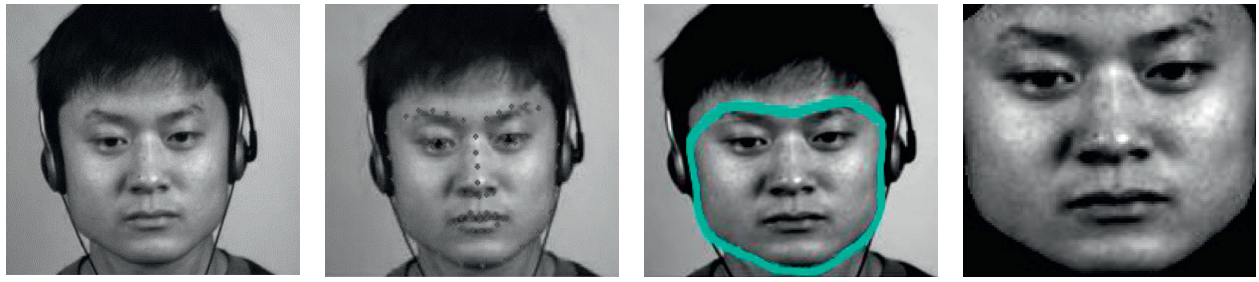


Figure 3: Flow chart of face feature point localization procedure.



(a) (b) (c) (d)

Figure 4: Face detection and face extraction framework. (a) Original image. (b) Face detected with 68 feature points. (c) Facial regions delineated using points 1–27. (d) Thoroughly extracted facial regions.

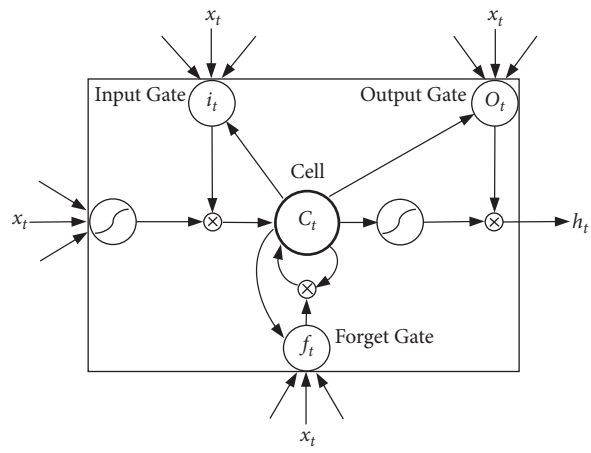


Figure 5: LSTM network structure.

The forgetting gate represents the selective memory of the image information passed from the previous node, i.e., retaining the important information and forgetting the unimportant information. The mathematical representation is

$$f_t = \sigma_g\left(W_f x_t + U_f h_{t-1} + b_i\right), \tag{1}$$

where σ_g is a sigmoid function and $f_t \in [0, 1]$, 0 means "no image information passed," and 1 value means "all image information passed."

The input gating updates the cell state using the current input x_t, mathematically represented as

$$i_t = \sigma_g\left(W_i x_t + U_i h_{t-1} + b_i\right), \tag{2}$$

where σ_g is also a sigmoid function and $i_t \in [0, 1]$ denotes the probability of remembering the input image information at the current moment. From this, the cell state can be updated according to the forgetting gate and the input gate for the purpose of whether to retain the image information or not. The symbol $*$ denotes multiplication by bit and is mathematically represented as

$$c_t = f_t * c_{t-1} + i_t * \sigma_c\left(W_c x_t + U_c h_{t-1} + b_c\right). \tag{3}$$

A new hidden state h_t is then generated by c_t:

$$\begin{aligned} o_t &= \sigma_g\left(W_o x_t + U_o h_{t-1} + b_o\right), \\ h_t &= o_t * \sigma_h\left(c_t\right), \end{aligned} \tag{4}$$

where σ_g is a sigmoid function, $o_t \in [0, 1]$, σ_h is generally a tanh function, and the tanh function is multiplied with o_t to obtain the final hidden layer h_t, and then the final output can be obtained from h_t as follows:

$$h_t = o_t * o_h\left(f_t * c_{t-1} + i_t * \sigma_c\left(W_c x_t + U_c h_{t-1} + b_c\right)\right). \tag{5}$$

From the above equation, it can be seen that the value of h_t is related to c_t and the value of c_t is independent of W_c, which is the important reason for the disappearance of gradients in the network.

The CNN-LSTM model combines the advantages of CNN and LSTM. The CNN model extracts locally relevant features of microexpression image data layer by layer through local connectivity, weight sharing, and pooling mechanisms. The LSTM model can effectively retain the historical information features of temperature text data sequences contained in human temperature images due to its excellent performance in temporal dependencies. Therefore, we construct a multimodal recognition neural network with a parallel combination of CNN and LSTM, as shown in Figure 6. The LSTM branch of the multimodal recognition neural network consists of two LSTM layers and flatten layer, and the text information of temperature data is converted into a feature vector named β in flatten layer after two LSTM layers; the convolutional neural network branch consists of four convolutional layers, four pooling layers, and one flatten layer, and the microexpression image is converted into a feature vector named β after four convolutional and pooling layers in the convolutional neural network branch. The microexpression image is converted into a feature vector named α in the flatten layer after 4 convolutions and pooling. After fusing the two feature vectors in series, the multimodal series fusion feature vector γ is obtained by full-connected layer FC-1 dimensionality reduction, and feature classification is performed to predict the classification results.

3.4.2. Implementation of Multimodal Data Fusion. In this subsection, the self-collected pairs of data are preprocessed according to Chapter 3. The first step uses the TIM algorithm to process the data set to extend the microexpression cycle and expand the data sample; the second step converts the fragment into a frame by frame form for normalization. The feature vector of a picture in the CNN + LSTM network in the self-built database is a matrix of 190 rows and 198 columns, as shown in Figures 7(a) and 7(b). The temperature images containing the head, chest, and shoulder temperature data of the volunteers are shown in Figure 7(c). After manually reading the temperature data of specific parts of the volunteers, the feature vector of these temperature data in the CNN+LSTM network is a 3-row and 2-column matrix, as shown in Figure 7(d).

After obtaining the body temperature data for each part of each training sample, the mean values of the temperature of each part of the body corresponding to each category of microexpressions were calculated. The final mean values of infrared body temperature measurements for the seven categories were obtained as shown in Table 2.

As can be seen from the table, the temperature of the head, chest, and shoulder is stable around 34.7°C when the emotions are happy. These mean values are the standard values for each category of emotion and will be fed into the CNN + LSTM network along with the microexpression images for training.

Next, let α be the feature vector of extracted microexpression images after network, $\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{im}$ representing the matrix of m images, respectively, and let β be the feature vector of body temperature data after network, $\beta_{i1}, \beta_{i2}, \ldots, \beta_{in}$ representing the head, chest, and shoulder temperature of n groups, respectively. This algorithm fuses the two-feature information in series, and the mathematical representation is

$$\begin{aligned} \alpha &= [\alpha_{i1}, \alpha_{i2}, \alpha_{i3}, \ldots, \alpha_{im}], \\ \beta &= [\beta_{i1}, \beta_{i2}, \beta_{i3}, \ldots, \beta_{in}], \\ \gamma &= \alpha + \beta = [\alpha_{i1}, \alpha_{i2}, \alpha_{i3}, \ldots, \alpha_{im}, \beta_{i1}, \beta_{i2}, \beta_{i3}, \ldots, \beta_{in}], \end{aligned} \tag{6}$$

where the temperature data are in text form, and when different text, numbers, English, and punctuation are fed into the network, different feature vectors are obtained. Tandem fusion can add the feature vector α of the microexpression image with the feature vector β of the three body temperature data to obtain a fused feature vector γ, as shown in Figure 8.

The advantage of fusing diverse features by tandem method is that this method is additive and no information is lost in the process. However, this method has an obvious disadvantage that when the amount of data is too large, data filtering or dimensionality reduction is required before fusion; otherwise it will inevitably bring a burden to the later training and classification, resulting in excessive computation, slow convergence of the model, overfitting, and other problems.

The k-nearest neighbor algorithm is one of the common classification algorithms used in data classification. k-nearest neighbor algorithm is the core idea that when there is a new data input with no category in a given labeled sample, the k-nearest neighbor algorithm will find k similar and related data in the training set, and if these k data belong to a category, the new data will belong to that category. The new data belongs to this category.

Let the given sample be $X = \{x_1, x_2, \ldots, x_n\}$; in n samples there are k-nearest neighbors; if $k_1, k_2, \ldots, k_m$ are the number of samples belonging to class $w_1, w_2, \ldots, w_c$ in k-nearest neighbors, respectively, then the discriminant function can be defined as

$$\begin{aligned} &g_i(x) = k_i, \\ &\sum_{i=1}^{c} k_i = k, \quad i = 1, 2, \ldots, c. \end{aligned} \tag{7}$$

The k-nearest neighbor algorithm calculates the distance between the points in the data set through known categories and the current point and selects the k points with the smallest distance from the current point. Then the probability of occurrence of the category in which the first k points are located is determined, and the category with the highest frequency of occurrence of the first k points is returned as the predicted category for the current point.

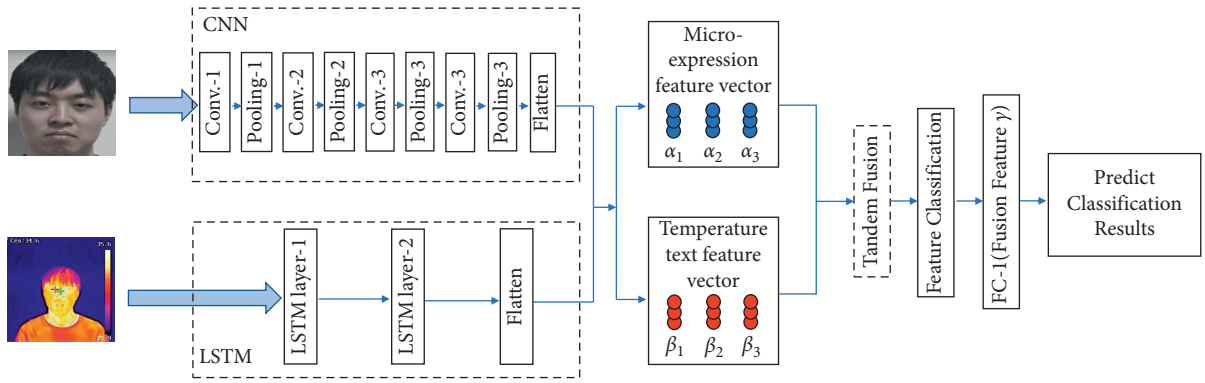


Figure 6: Multimodal recognition neural network model.

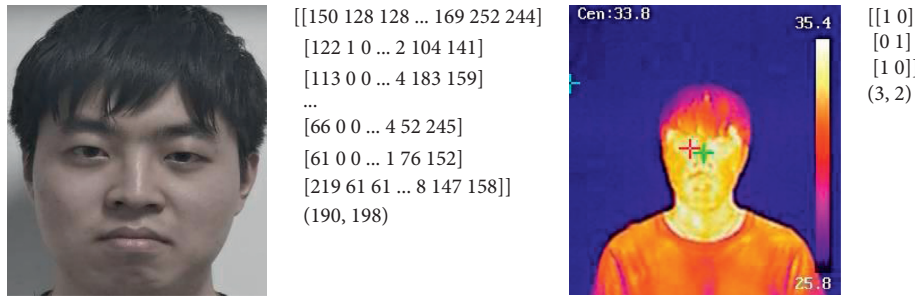


Figure 7: Feature vectors (b, d) obtained from the text data contained in images (a) and (c) after CNN + LSTM neural network.

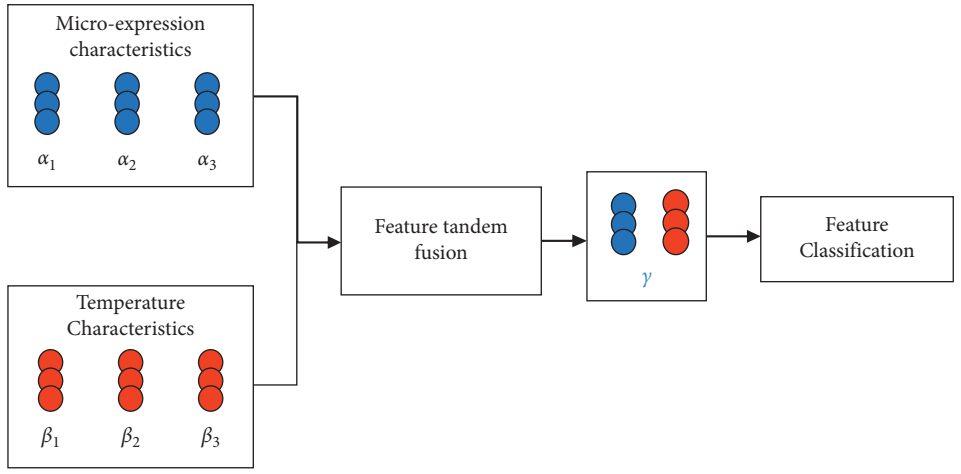


Figure 8: Multimodal feature vector fusion framework.

4. Experiment

4.1. Experimental Software and Hardware Environment. Operating system: Microsoft Windows 10 64-bit.

Development language: Python, MATLAB.

Development Environment: TensorFlow, Keras.

Data sets: SAMM data set, CASME data set, CASME II data set, self-built data set.

Hardware: MSI Laptop GL62VR-7RFX (Intel(R) Core(TM) i7-7700HQ CPU@2.80 GHz), Hikvision Thermometer H11, Mind Vision industrial camera.

Memory: 16 GB.

Graphics card model: NVIDIA GeForce GTX 1060.

4.2. LBP. This section uses the LBP algorithm to realize microexpression recognition on the public database, so that it can be compared with the recognition accuracy of the following LSTM and CNN + LSTM. The implementation of this algorithm is divided into three steps; firstly, data preprocessing: this step has been done in Chapter 3; secondly, the features of each image are extracted using the LBP

algorithm; finally, the classification is performed using the classifier. After the experimental study, it is found that LBP has many advantages, for example, the image grayscale value is constant after extracting features using LBP and the features will not change when the image is rotated and shifted. However, LBP can only process images singly and cannot process long sequence data.

4.2.1. Feature Extraction. The core idea of the LBP algorithm is to compare a pixel point in an image with the grayscale value of its neighboring pixel points computationally, and if the neighboring pixel value is larger than the pixel value of that point, the value obtained is 1, and otherwise it is 0. The mathematical representation is as follows:

$$\mathrm{LBPP}, R(xc, yc) = \sum_{P=0}^{P-1} s(iP - ic)2^P, \tag{8}$$

where (xc, yc) is the pixel at the center point, ic represents the luminance, and s is the sign function satisfying the following relationship:

(i) $s(x) = 1,\ x \geq 0,$

(ii) $s(x) = 0,\ x < 0.$

The difference between a pixel point and its neighboring pixels is preserved in the LBP algorithm. External environmental factors, such as changes in brightness and contrast, change the pixel value of the image, but the magnitude of the LBP value remains unchanged, so LBP can avoid a series of problems arising from the difference relationship between pixel values in an image.

4.2.2. Classification Method. Classification is the process of classifying the extracted features and outputting the classification results. Corinna Cortes et al. proposed in 1995 the use of support vector machines (SVM), to solve problems such as classification. For binary classification problems, the core idea of SVM is to construct an optimal flat space to classify the data into two classes; in case of high-dimensional data, the algorithm constructs a mathematical representation of the hyperplane space as

$$f(x) = w^T x + b, \tag{9}$$

where both w and x are vectors. The algorithm classifies the data according to the value of $f(x)$, comparing the data to small black dots on the hyperplane. $f(x) = 0$ when the dots are in the central hyperplane M0; M1 and M2 hyperplanes exist on both sides of M0, and the values of $f(x)$ of the data distributed in M1 and M2 are 1 or −1, respectively.

SVM is stable and efficient compared with other algorithms. Although the operation speed of random forest (RF) is faster than SVM, it is easy to overfit when dealing with noisy data of microexpression fragments; kmeans algorithm is not applicable to high-dimensional data such as human face. Therefore, SVM is still the mainstream choice for microexpression recognition tasks at present.

4.2.3. Analysis of Experimental Results. In the experiments to study the LBP algorithm, the number of samples is too small, so the experiments use SVM classifier to avoid overfitting. The face region also needs to be divided into chunks when LBP extracts features, and the final results are different for different division methods. We divide the face region into $n \times n (1 \leq n \leq 8, n \in Z)$ regions and recognize each image. The recognition rates of different face region divisions are shown in Figure 9.

From Figure 9, we can see that the recognition rate increases with the increase of the number of regional blocks, and the highest recognition rate of 62.8% is achieved when the number of regional blocks is 7×7. However, due to the increasing number of regional chunks, the dimensionality of feature vectors is also increasing, and the computation volume is also increasing, which leads to the slow convergence of the model, so the recognition rate of the model is instead lower than that at 7×7 when the number of chunks is 8×8. The LBP algorithm has been developed for decades and has a pivotal position in the microexpression recognition task, and the traditional excellent algorithms are also progressing and in the deep learning is still occupying a place in the era of hot learning. Experimental results show that feature extraction using LBP is simple and efficient, but there is still much room for improvement at present. Below we will compare the recognition accuracy obtained in this section with the microexpression recognition accuracy obtained by the experimental method below and select the best method for multimodal feature fusion experiments.

4.3. CNN and LSTM. This experiment will verify the effect of microexpression recognition in LSTM alone and in the combination of CNN + LSTM. Both approaches use the CASME II dataset and divide the training set and test set according to 6 : 4.

CNN + LSTM can be pretrained with CNN first and then train the CNN + LSTM model by fine-tuning the parameters and other methods when the network starts to converge, which can make the network converge more rapidly. If the two neural networks are trained directly at the same time at the beginning, it will probably lead to the network not converging or converging too slowly due to the small change in the value of the loss function, which is time-consuming and labor-intensive.

Algorithm framework is shown in Figure 10.

During training, both methods stabilized after about 3000 iterations, and the most obvious feature is that the value of the loss function stabilized at about 1.0, and the training process is shown in Figure 11. If the training is continued, it will lead to overfitting of the model. After the training is completed, the accuracy of the model is tested and the confusion matrix is obtained, as shown in Figure 12.

As can be seen from Figure 12, the accuracy of the four categories of normal, disgusted, surprised, and happy in this model is higher than the recognition rates of several other categories, which is due to the difficulty in evoking and collecting certain categories when acquiring digital microexpression images, resulting in too few samples, too low

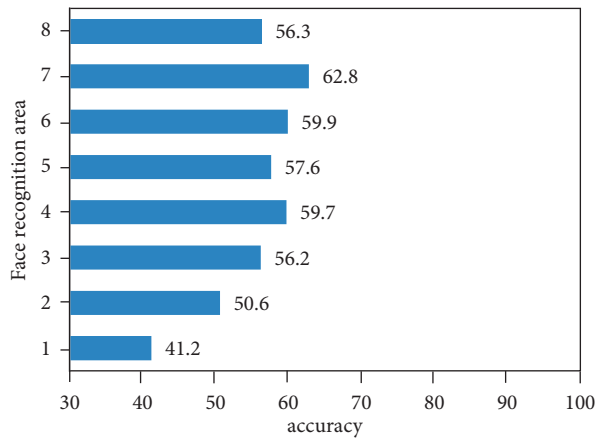


Figure 9: Recognition rate of microexpression images with different number of blocks.

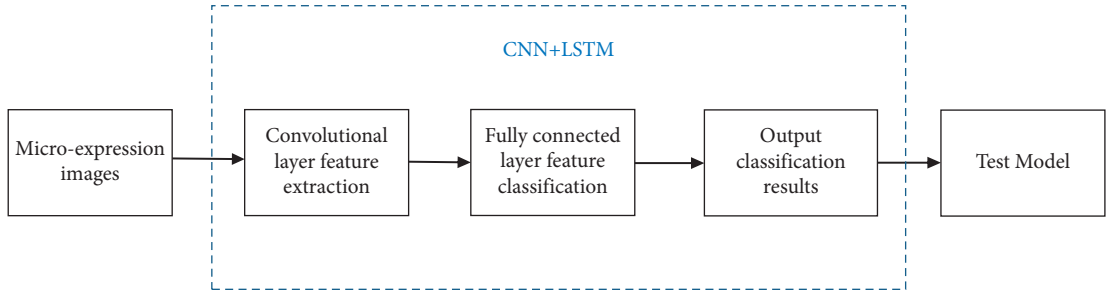


Figure 10: CNN + LSTM algorithm framework.

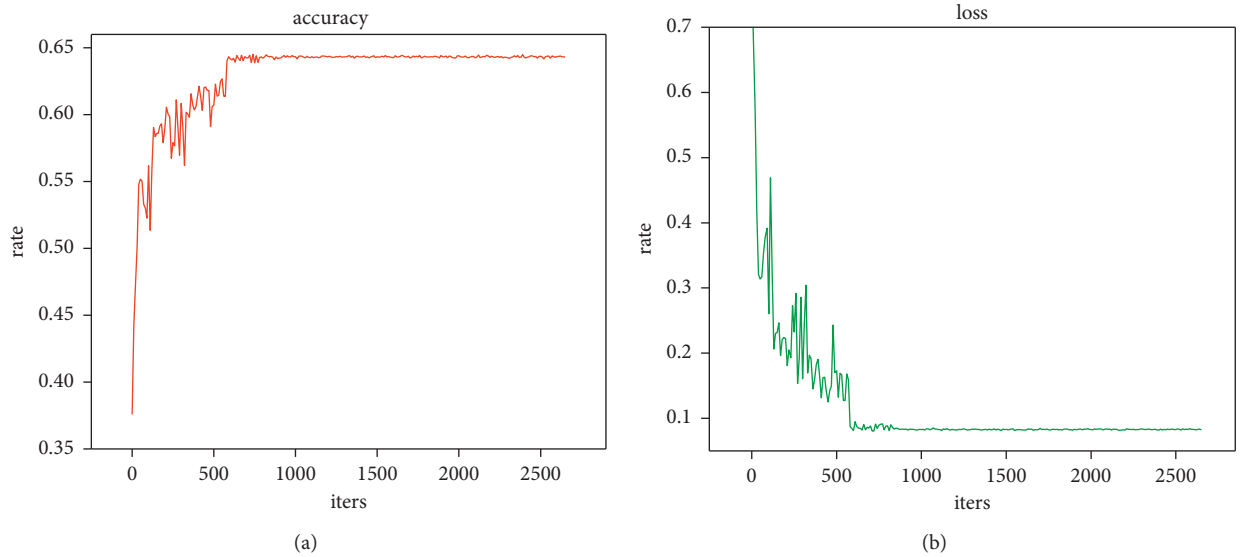


Figure 11: (a) Accuracy of CNN + LSTM under CASME II, (b) loss function of the training model.

recognition rates, and perhaps even overfitting problems. The recognition rates obtained by testing on each of the three datasets with the two methods are shown in Figure 13.

From Figure 13, we can see that the recognition rate of CNN + LSTM is higher under the three datasets than under the LSTM model alone. The reason for this is that after the CNN convolutional layer the features contain redundant information that has been filtered once and then filtered by the LSTM, the redundant information in the features is discarded, and the remaining information belongs to the

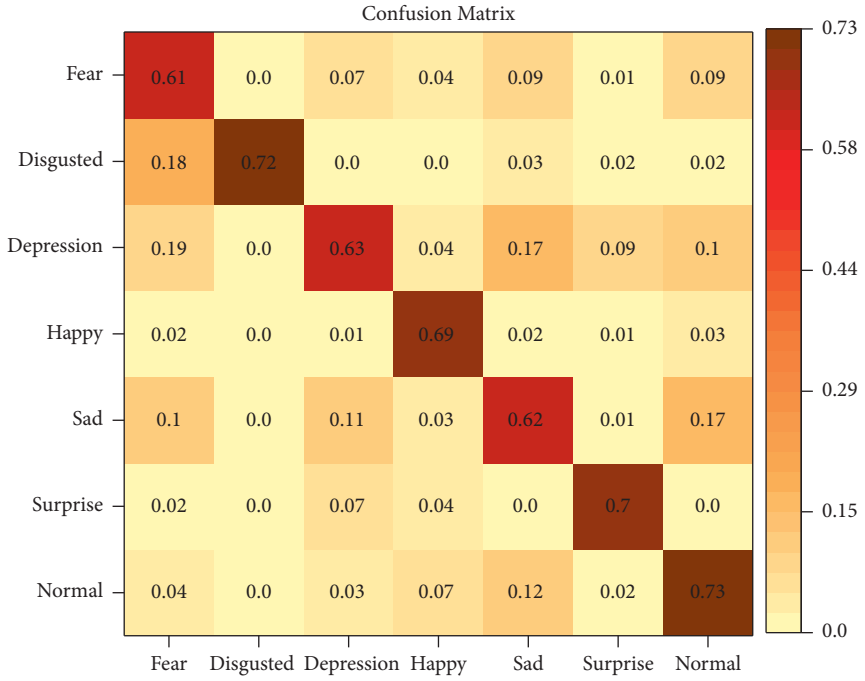


Figure 12: Confusion matrix of CNN + LSTM under CASME II.

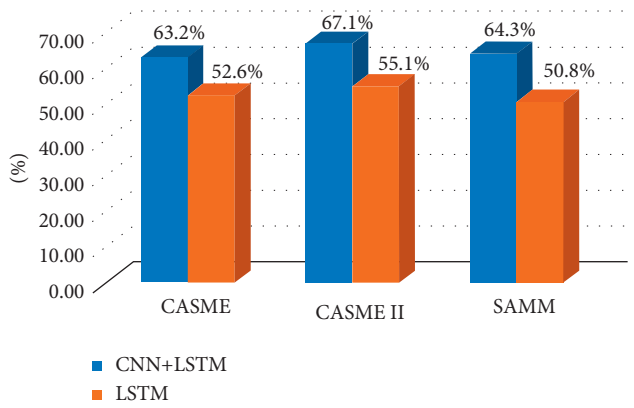


Figure 13: Recognition rates of different algorithm models under three datasets.

essence, while when trained with the LSTM alone, no features are available in the network at this time and further training is required. The experimental results do not differ much under the three datasets. In the microexpression recognition task, the recognition accuracy is higher using CNN + LSTM neural network.

4.4. Recognition Effect of CNN + LSTM under Multimodal Fusion Database. Recognition effect under multimodal fusion database using CNN + LSTM is measured by comparing the LBP algorithm with CNN and LSTM algorithms. We selected the CNN + LSTM neural network with higher recognition rate and added the corresponding human temperature data to the individual microexpression images for training and then compared the recognition rate of the model trained by the individual microexpression images without the temperature data with that of the model trained under the multimodal fusion database. During the first training with the multimodal fusion database, the accuracy of the model and the value of the loss function are shown in Figure 14.

As can be seen from the figure, the model accuracy during this training stabilized at 0.69 after 60 iterations, with almost no change, and the loss value also stabilized at around 0.9. However, after evaluating the model, it is found that the effect is very unsatisfactory, which is due to the inclusion of Early Stopping function (Early Stopping) in the network structure, the purpose of which is to prevent overfitting. During this training, the value of the loss function dropped quickly and the curve was almost stable, indicating that the model was stuck in a local optimum that could not be jumped out, which is one of the drawbacks of the Early Stopping method.

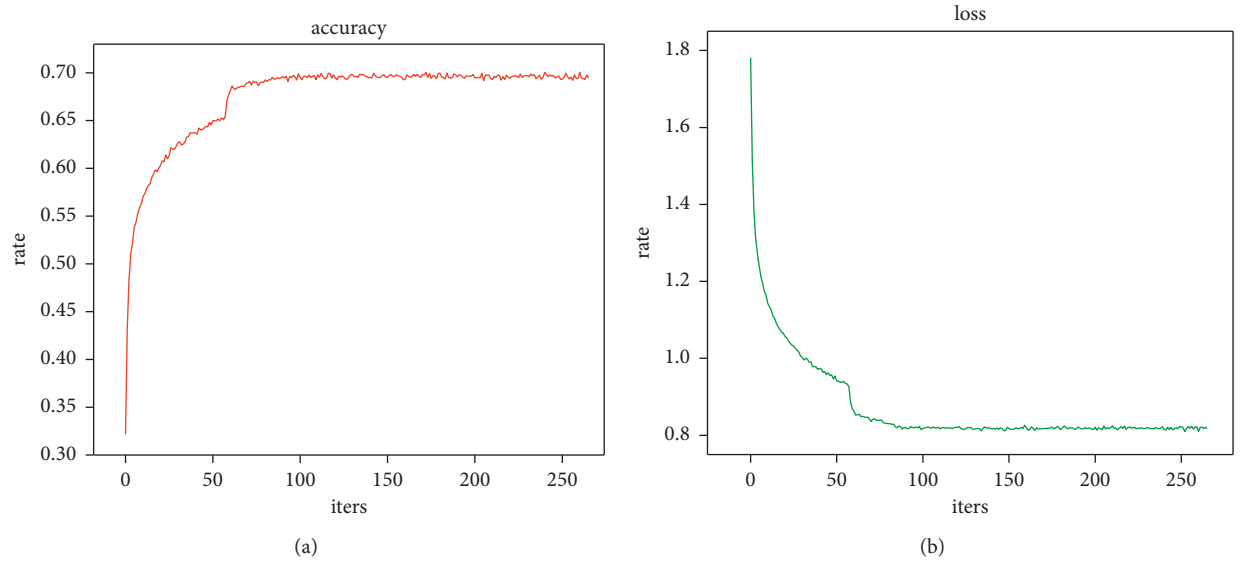


Figure 14: (a) Accuracy of CNN + LSTM under multimodal fusion database and (b) loss function of the training model.

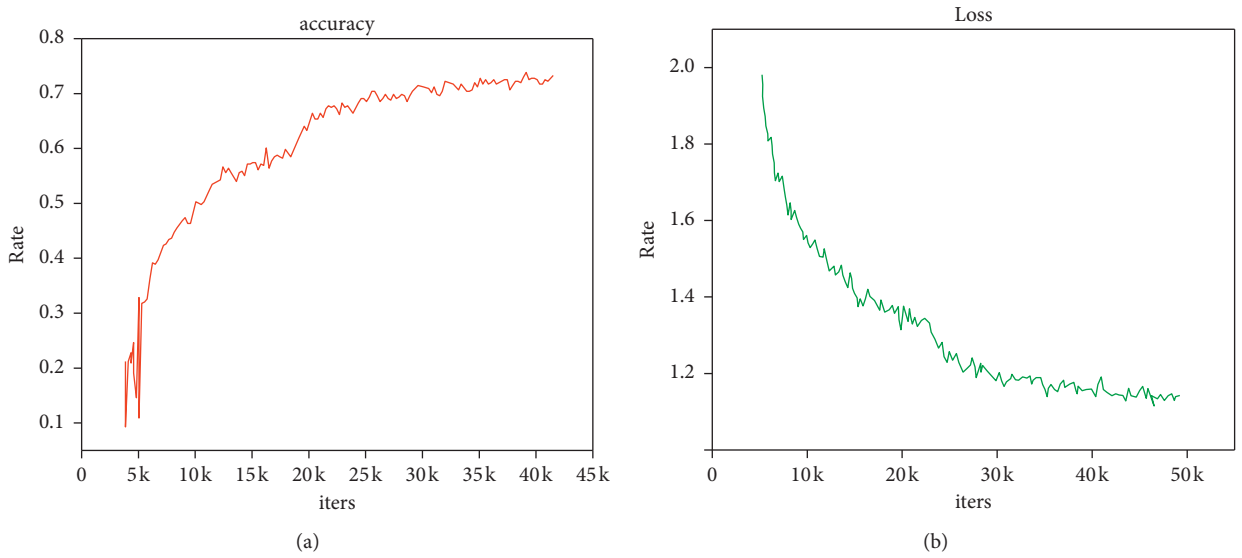


Figure 15: Accuracy curve (a) loss function (b) for training with CNN neural network.

After modifying the structural parameters of the neural network, the second training in the multimodal feature fusion database was started. For this experiment, we used a migration learning approach to first pretrain microexpressions on the multimodal fused dataset using CNN. The accuracy of the model during training is shown in Figure 15(a), and the loss function is shown in Figure 15(b).

From the figure, we can see that the model accuracy reaches about 80% and the loss function drops to around 1.2. Then, the model was allowed to continue training in the network structure of CNN + LSTM using the data from the multimodal fusion database, and the accuracy curve reached about 75.0% as shown in Figure 16(a), and the loss function curve is shown in Figure 16(b). The average accuracy of the final model on the test set is 75.1%, as shown in Figure 17.

Finally, we use CNN + LSTM trained separately in the microexpression images in Chapter 3, and the average accuracy obtained on the test set is shown in Figure 18. The average recognition rate obtained by training on the self-built multimodal fusion database is plotted against the average recognition rate obtained by training the microexpression images alone, as shown in Figure 19.

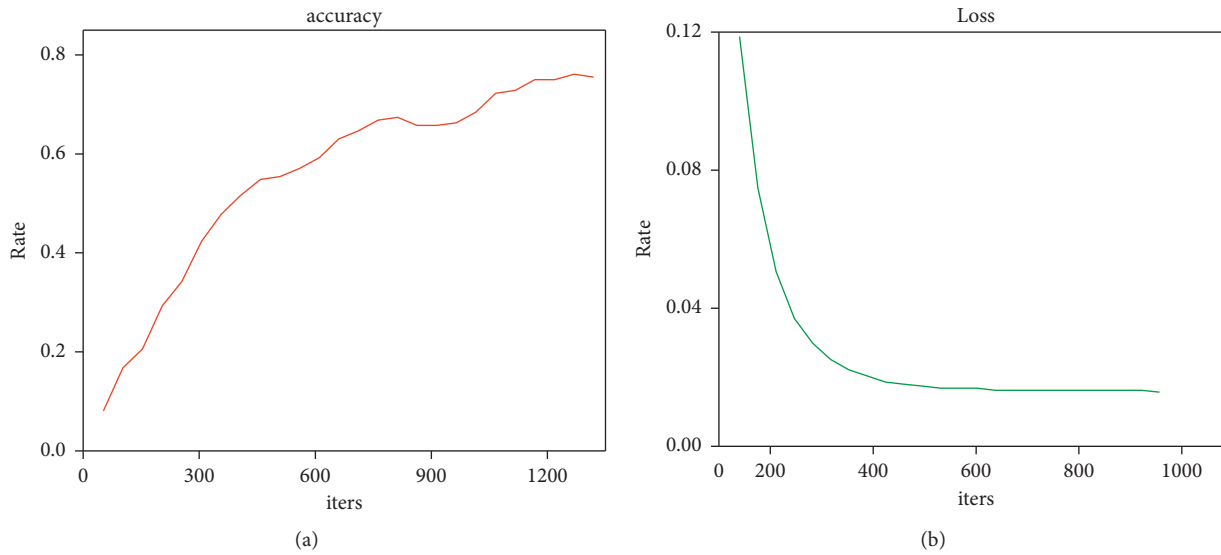


Figure 16: Accuracy curves in CNN + LSTM training (a) and loss function (b).

```
23840/25000 [=======================>..] - ETA: 0s
24512/25000 [=======================>.] - ETA: 0s
25000/25000 [==========================] - 2s 94us/step
test_loss: 0.745502, accuracy: 0.751280
```

Figure 17: Average accuracy of models on multimodal fusion database.

```
23296/25000[=======================>...] - ETA: 0s
23936/25000[=======================>..] - ETA: 0s
24544/25000[=======================>.] - ETA: 0s
25000/25000[==========================] - 4s 145us/step
test_loss: 0.539321, accuracy: 0.637120
```

Figure 18: Average accuracy of models trained on microexpression images alone.

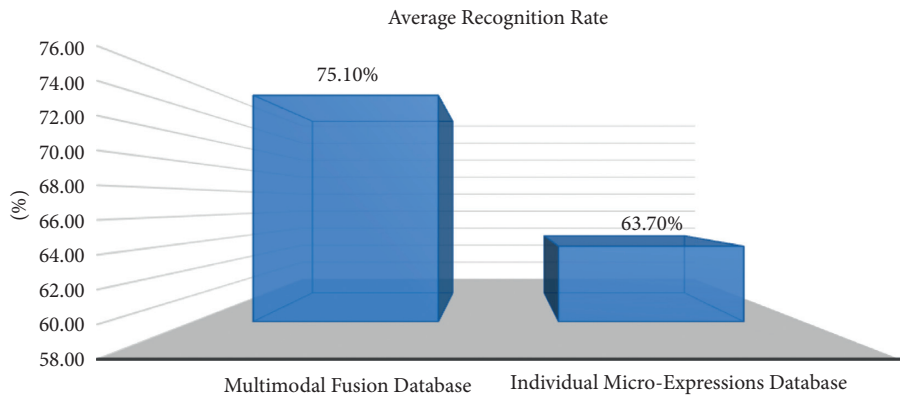


Figure 19: Average recognition rate of CNN + LSTM model in multimodal fusion database vs. average recognition rate of microexpression images trained alone.

5. Conclusion

This study introduces the data preprocessing part of the current microexpression recognition task in Chapter 3. Both traditional and deep learning methods require preprocessing of data. Reasonable and effective preprocessing methods tend to make the results better and better. In this study, we completed the partitioning of face regions and face feature point detection in the preprocessing part and extended the microexpression cycle using TIM algorithm to prepare for the feature extraction later. Subsequently, we compared the recognition rates of LBP, LSTM, and CNN + LSTM algorithms in the public microexpression database. The experimental results show that the recognition rate increases with the number of image region chunks in the LBP algorithm and reaches the highest recognition rate of 62.8% when the chunks of regions are 7×7. The feature extraction using LBP is concise and efficient, but there is still much room for improvement. The recognition rate based on LSTM algorithm reaches 52.6% in CASME dataset, 55.1% in CASME II, and 50.8% in SAMM. The recognition rate based on CNN + LSTM deep learning algorithm reaches 63.2% on the CASME database. It reached 67.1% in CASME II and 64.3% in SAMM. It can be seen that the CNN + LSTM deep learning algorithm is superior in recognizing microexpression images. Then, we proposed a diversified information coupling method combining temperature data and microexpression images and designed an algorithmic framework to conduct an experimental study of multimodal fusion for microexpression recognition. The temperature data of the head, chest, and shoulders of volunteers were collected simultaneously by a thermometer, and the mean temperature values of each category corresponding to the three parts were calculated. Finally, use the CNN + LSTM network to train in the self-built multimodal fusion database, and use the k-nearest neighbor classification to get the experimental results. The experimental results show that a microexpression recognition method based on multimodal fusion is accurate and effective in improving the recognition rate of microexpressions, and the recognition rate reaches 75.1% on the multimodal feature fusion database, which proves that the method is effective and feasible.

Data Availability

The data used to support the results of this study are available on request from the corresponding authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to thank their colleagues for their guidance and the help of their classmates. This paper was sponsored by the "Qing Lan Project" of Jiangsu Universities, the General Natural Science Research Project of Jiangsu Universities (19KJD510005), Guangxi Key Laboratory of Automatic Testing Technology and Instruments will be open to fund in 2021 (YQ21207), and the industry-university cooperative education project of the Ministry of Education (201902168015).

References

[1] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[2] M. Bai, "Detection of micro-expression recognition based on spatio-temporal modelling and spatial attention," in *Proceedings of the ICMI'20: International Conference on Multimodal Interaction*, Utrecht, Netherlands, October 2020.

[3] J. Wen, W. Yang, L. Wang, W. Wei, S. Tan, and Y. Wu, "Cross-database micro expression recognition based on apex frame optical flow and multi-head self-attention," in *Proceedings of the International Symposium on Parallel Architectures, Algorithms and Programming*, pp. 128–139, Springer, Singapore, December 2020.

[4] S. Li, "Research on facial micro-expression recognition method based on deep learning," Master thesis, China University of Mining and Technology, Xuzhou, China, 2020.

[5] S. Yuting and H. Xu, "Micro-expression recognition algorithm based on multiple motive feature fusion," *Laser & Optoelectronics Progress*, vol. 57, no. 14, Article ID 141504, 2020.

[6] H. Q. Khor, J. See, R. C. W. Phan, and W. Lin, "Enriched long-term recurrent convolutional network for facial micro-expression recognition," in *Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 667–674, IEEE, Xian, China, May 2018.

[7] Y. Liu, H. Du, L. Zheng, and T. Gedeon, "A neural micro-expression recognizer," in *Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pp. 1–4, IEEE, Lille, France, May 2019.

[8] D. Matsumoto and H. S. Hwang, *Reading Facial Expressions of Emotion*, University of Oxford, England, UK, 2011.

[9] M. Shreve, S. Godavarthy, V. Manohar, D. Goldgof, and S. Sarkar, "Towards macro-and micro-expression spotting in video using strain patterns," in *Proceedings of the 2009 Workshop on Applications of Computer Vision (WACV)*, pp. 1–6, IEEE, Snowbird, UT, USA, December 2009.

[10] S. Porter, L. Ten Brinke, and B. Wallace, "Secrets and lies: involuntary leakage in deceptive facial expressions as a function of emotional intensity," *Journal of Nonverbal Behavior*, vol. 36, no. 1, pp. 23–37, 2012.

[11] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[12] Q. Wu, X. Shen, and X. Fu, "The machine knows what you are hiding: an automatic micro-expression recognition system," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, pp. 152–162, Springer, Berlin, Germany, October 2011.

[13] S. Luo, J. Zhang, Q. Zhang, and X. Yuan, "Multi-operator image retargeting with automatic integration of direct and indirect seam carving," *Image and Vision Computing*, vol. 30, no. 9, pp. 655–667, 2012.

[14] B. Jiang, L. Xie, X. Liu, J. Han, and Z. L. Wang, "Micro-expression spotting using optical flow magnitude estimation," *Journal of ZheJiang University (Engineering Science)*, vol. 51, no. 3, pp. 577–583, 2017.

[15] Y. Zhang, B. Lu, X. Hong, G. Zhao, and W. Zhang, "Micro-expression recognition based on local region method," *Journal of Computer Applications*, vol. 39, no. 5, pp. 1282–1287, 2019.

[16] M. A. Takalkar and M. Xu, "Image based facial micro-expression recognition using deep learning on small datasets," in *Proceedings of the 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–7, IEEE, Sydney, Australia, November 2017.

[17] R. Mohanty, S. L. Happy, and A. Routray, "Spatial—spectral regu-larized local scaling cut for dimensionality reduction in hyper-s pectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 6, pp. 932–936, 2019.

[18] F. Xu, J. Zhang, and J. Z. Wang, "Micro-expression identification and categorization using a facial dynamics map," *IEEE Transactions on Affective Computing*, vol. 8, no. 2, pp. 254–267, 2017.

[19] W. Su, Y. Wang, F. Su, and Z. Zhao, "Micro-expression recognition based on the spatio-temporal feature," in *Proceedings of the 2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1–6, IEEE, San Diego, CA, USA, July 2018.

[20] S. T. Liong and K. Wong, "Micro-expression recognition using apex frame with phase information," in *Proceedings of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 534–537, IEEE, Kuala Lumpur, Malaysia, December 2017.

[21] J. Pei and P. Shan, "A micro-expression recognition algorithm for students in classroom learning based on convolutional neural network," *Traitement du Signal*, vol. 36, no. 6, 2019.

[22] Y. Li, X. Huang, and G. Zhao, "Can micro-expression be recognized based on single apex frame," in *Proceedings of the International Conference Onlmage*, pp. 3094–3098, 1EEE, Athens, Greece, October 2018.

[23] Z. Zhou, G. Zhao, and M. Pietikäinen, "Towards a practical lipreading system," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR) 2011*, pp. 137–144, IEEE, Colorado Springs, CO, USA, June 2011.

[24] L. Nummenmaa, E. Glerean, R. Hari, and J. K. Hietanen, "Bodily maps of emotions," *Proceedings of the National Academy of Sciences*, vol. 111, no. 2, pp. 646–651, 2014.

[25] M. X. Jiang, C. Deng, M. M. Zhang, J. S. Shan, and H. Zhang, "Multimodal deep feature fusion (MMDFF) for RGB-D tracking," *Complexity*, vol. 2018, Article ID 5676095, 8 pages, 2018.

[26] H. Yang, P. Qian, and C. Fan, "An indirect multimodal image registration and completion method guided by image synthesis," *Computational and Mathematical Methods in Medicine*, vol. 2020, Article ID 2684851, 10 pages, 2020.

[27] A. Abdalbari, X. Huang, and J. Ren, "Endoscopy-MR image fusion for image guided procedures," *International Journal of Biomedical Imaging*, vol. 2013, Article ID 472971, 10 pages, 2013.

[28] D. Wu, J. Chen, W. Deng, Y. Wei, H. Luo, and Y. Wei, "The recognition of teacher behavior based on multimodal information fusion," *Mathematical Problems in Engineering*, vol. 2020, Article ID 8269683, 8 pages, 2020.

[29] Y. Huang, J. Yang, P. Liao, and J. Pan, "Fusion of facial expressions and EEG for multimodal emotion recognition," *Computational Intelligence and Neuroscience*, vol. 2017, Article ID 2107451, 8 pages, 2017.

[30] Y. Wang, X. Liu, and C. Yu, "Assisted diagnosis of alzheimer's disease based on deep learning and multimodal feature fusion," *Complexity*, vol. 2021, Article ID 6626728, 10 pages, 2021.

[31] J. B. Li, J. Wang, M. Xu, and C. Wang, "Twitter sentiment analysis based on hopping LSTM-CNN model," *Computer Simulation*, no. 08, pp. 478–481+496, 2021.

[32] S. Hochreiter and J. Schmidhuber, "LSTM can solve hard long time lag problems," *Advances in Neural Information Processing Systems*, pp. 473–479, 1997.