A Bimodal Emotion Recognition Algorithm for Audio and Video Based on Emotion Modeling

Yang Liu
Department of Psychiatry, Fundamental and Clinical Research on Mental Disorders Key Laboratory of Luzhou, Affiliated Hospital of Southwest Medical University, China
LIUYANG030325@163.com

Shudan Feng
School of Humanities and Management Science Southwest Medical University China
Fengshudan2517@163.com
*Corresponding author

Kaiyong Li
School of Physics and Electronic Information Engineering Qinghai Minzu University China
likaiyong8558@163.com

Abstract: *In audio-video bimodal emotion recognition, audio features and video features come from different modalities and have different representations and semantic information. Traditional methods rely only on the information of a single modality, which makes the fused features unable to comprehensively represent the emotional state, resulting in poor recognition results and small correlation coefficients. For this reason, a bimodal emotion recognition algorithm based on emotion modeling is proposed for audio and video. Firstly, the emotional audio is sub-framed by Fourier transform to obtain the Meier Frequency Cepstrum Coefficient (MFCC) features of emotional audio, extract the frame-level speech time-domain signal input features, and mine the audio SoundNET coding features of emotion; Secondly, the above three features are spliced together to complete the mining of total emotional audio features of emotion; then, the Recurrent Neural Network (RNN) and the long and Short-Term Memory Network (LSTM) are used to capture the emotional video features in depth; Finally, cross-modal learning and attention mechanism are used to integrate the extracted emotion features, and the emotion type is analyzed by the decision-level fusion network to complete the audio and video bimodal emotion recognition, which effectively avoids the problem of poor single-modal recognition results and improves the recognition accuracy and reliability. The results show that the proposed algorithm is effective in recognizing bimodal emotions in audio and video, and the correlation coefficient of the recognition results is large.*

Keywords: *Audio, video, fourier transform, soundnet neural network, recurrent neural network, long and short-term memory network.*

Received March 6, 2025; accepted September 21, 2025
https://doi.org/10.34028/iajit/23/2/3

1. Introduction

Nowadays, when mental health is getting more and more attention, accurately identifying and analyzing individual's emotional state is of vital importance for preventing and treating mental diseases and improving mental health. Traditional emotion recognition methods, such as observing facial expressions, analyzing language content or detecting physiological indicators, although to a certain extent capable of revealing an individual's emotional state, generally suffer from the problems of strong subjectivity, lack of accuracy and high cost. Therefore, emotion recognition algorithms have emerged, which mathematically describe and model human emotions to enable machines to understand and recognize individual emotional states. However, in practice, the accuracy of emotion recognition algorithms is often constrained by a variety of factors. The complexity and diversity of individual facial expressions, voice characteristics, language content and body postures increase the difficulty of the algorithms in recognizing emotions accurately. In addition, the differences in emotional expressions in different cultures and backgrounds may also lead to errors in the recognition process [1, 7]. In order to better apply emotion recognition algorithms to mental health, many scholars have conducted in-depth research on them.

Zhang *et al*. [18] utilized a Long Short-Term Memory Network (LSTM) to capture the features of each modality in emotional speech and video, and these features were memorized and processed in the LSTM network through a gating mechanism. These features are then analyzed in depth using a Multilayer Feedforward Neural Network (MLFN) to identify and classify emotional states. This method shows strong feature mining ability in emotion recognition task, and can accurately capture speech and video features that change over time, thus significantly improving the accuracy of emotion recognition. However, this method adopts LSTM network, which leads to relatively high computational complexity, long training time and more computational resources when dealing with long sequence data or large datasets. Zong *et al*. [21] used hypergraph to establish multivariate relationships of emotion multimodality, and mined the information of each time step in the sequence by capsule network and graph convolution. After analysis, the current mental state is identified and the emotion recognition task is completed. This method replaces the traditional graph structure based on binary relations by introducing

hypergraph to establish multimodal multivariate relations, and realizes more adequate and efficient multimodal feature fusion. However, the method needs to train the model with historical data, and the performance and accuracy of the model decreases in some new scenarios. Zhang *et al*. [16] constructed a multimodal emotion recognition model using Deep Emotional Arousal Network (DEAN), which accurately recognizes the emotions conveyed by an individual's facial expression and body language by using the cross-modal transformer module and Bidirectional Long Short-Term Memory (BiLSTM) operations. By using the cross-modal transformer module, the DEAN model can effectively fuse information from different modalities (e.g., facial expression, body language, etc.,) which significantly improves the accuracy and robustness of emotion recognition. However, this method requires high hardware and is limited in the use of scenarios. Liu *et al*. [10] used a convolutional neural network to extract spatial features, captured the temporal information through a temporal transformer, and introduced an attention mechanism to represent the extracted spatial features; then, the features were fused in a graphical convolutional network and trained to generate the final recognition results. This method has a powerful feature extraction capability and can generate richer and more comprehensive feature representations, thus significantly improving the accuracy of the recognition results. However, the method applies multiple neural network models and involves a large number of parameters, and the recognition process may require a long computation time, resulting in a long waiting time for the users.

Aiming at the above problems, the paper proposes a bimodal emotion recognition algorithm for audio and video based on emotion modeling. Firstly, Fourier transform and Recurrent Neural Network (RNN) are used to deeply excavate the features of each modal information of emotion, and then the modal features are interacted and fused by Cross-Modal hierarchical Attention (CMA) to generate new features to reduce the difficulty of the subsequent feature judgments, and then relying on the decision-making level fusion network to comprehensively consider the multimodal fused feature information and accurately determine the category of the emotion belonging to it, which can effectively avoid the problem of poor recognition results of the single-modal information to realize the high accuracy and high efficiency of emotion recognition and detect the potential signs of psychological problems in time, so as to prevent and treat the mental health problems.

2. Audio and Video Bimodal Feature Mining Based on Sentiment Modeling

Emotional expression is a complex process involving multiple information channels. Audio-video bimodal feature extraction technology can comprehensively reflect an individual's emotional state by capturing the bimodal information of sound (e.g., intonation, speech rate, volume, etc.,) and vision (e.g., facial expression, body movement, eye contact, etc.). In the process of emotional expression, audio and video have their own unique focus: sometimes the audio information is more significant, such as the roar of anger, and sometimes the video information is more critical, such as the expression of smile or frown. The bimodal feature extraction technique can effectively make up for the shortcomings of single modality in emotion recognition by extracting the information of these two modalities. Therefore, a bimodal feature mining model of audio and video for emotion modeling is constructed to avoid the degradation of recognition accuracy and reliability caused by single-modal information, in order to capture the audio and video modal features of emotional expressions.

2.1. Audio Feature Mining for Sentiment Modeling

In the feature mining of emotional audio signals, different types of features play important roles. The Meier Frequency Cepstrum Coefficient (MFCC) has the unique advantage of simulating the sensitivity of the human ear to different frequencies, which makes it excellent in the field of emotion recognition and provides strong support for subsequent analysis. IS09, IS11 and IS13 frame-level features focus on the audio signal in different time periods, and are able to present the detailed emotional information contained in each time period. In particular, this detailed sentiment information can help to realize the precise alignment between acoustic features and visual signals in cross-modal analysis, thus effectively improving the accuracy and robustness of the sentiment analysis. In addition, SoundNet encoded features are obtained from video data through transfer learning, based on which the learned features have obvious advantages for cross-modal sentiment analysis, and can contribute unique perspectives to the comprehensive analysis of different modal data [10, 11]. To this end, the MFCC, IS09, IS11, IS13 frame-level features, and SoundNET coding features of emotional audio signals were mined, with the aim of fully exploiting the emotional information embedded in these features, providing a solid foundation for more accurate sentiment analysis, and improving the accuracy of mental health analysis.

2.1.1. Audio MFCC Feature Mining for Emotions

- *Step* 1: Fourier transform. The audio of the emotion is processed into frames in order to obtain a sequence composed of short emotional speech fragments, $A=\{a_1^t, a_2^t, \ldots, a_M^t\}$, using the fast Fourier transform to transform each speech frame a_m from a time domain signal to a frequency domain signal [4]. The specific operations are:

$$a_k^f = \sum_{m=0}^{M} a_m^t e^{-j\frac{2\pi}{M}mk}, k = 0,1,2,\dots,M-1 \quad (1)$$

Among them, M denotes the total number of frames in the sequence, f denotes the frequency domain, t denotes the time domain. j denotes imaginary units. k denotes the sample point index variable within the speech frame; the e denotes a natural constant. m indicates the voice frame sequence number.

- *Step* 2: Take the logarithm. The frequency domain speech signal obtained by Fourier transform is input into the Meier filter bank for filtering operation, then the frequency response of each Meier filter is $H_l(k)$:

$$H_l(k) = \begin{cases} 0, k \le f(l-1) or\ k \ge f(l+1) \\ \frac{2(k-f(l-1)}{(f(l+1)-f(l)-f(l-1)'} \\ f(l-1) \le k \le f(l) \\ \frac{2(f(l=1)-k)}{(f(l+1)-f(l-1))(f(l)-f(l-1)'} \\ f(l) \le k \le f(l+1) \end{cases} \quad (2)$$

Among them, $f(l)$ represents the center frequency of the filter, l indicates the l-th filter. Accordingly, the filtered output of all frequency domain signals $A^f = \{a_1^f, a_2^f, \dots, a_M^f\}$ can be expressed as:

$$Log(l) = 1n(\sum_{k=0}^{M-1} \left|a_k^f\right|^2 H_l(k)), 0 \le m \le M \quad (3)$$

- *Step* 3: Inverse transform. The output of each filter bank is used as the input of the discrete cosine transform, in order to obtain the final MFCC features of the emotional audio, which is computed as follows:

$$C(a_m^t) = \sum_{l=0}^{M-1} Log(l) cos\frac{m\pi(l-0.5)}{L}, m = 1,2,\dots,L$$
$$A_{MFCC} = \sum_{m=0}^{M} C(a_m^t) \quad (4)$$

Among them, L indicates the number of Mel filter banks.

2.1.2. IS09, IS11, and IS13 Frame-Level Feature Mining for Emotional Audio

Since the open-source audiovisual signal processing tool openSMILE can provide various standard feature sets for emotion recognition, openSMILE is chosen to extract frame-level features of speech signals. Take IS11 for example, the feature set contains a total of 4,368 statistical features calculated from 108 low-level descriptors of the audio signal. Among them, the low-level descriptors cover four energy levels, fifty spectral levels and fifty-four delta features, and the statistical functions cover 39 basic functions such as maximum value, mean, variance, quadratic mean, duration of pitch change, etc., [3, 14].

In emotion audio recognition, facing many standard feature sets, IS09, IS11 and IS13 feature sets, which are highly adaptable to the emotion recognition task, are selected with the help of speech emotion filtering technology to extract features from the frame-level speech time-domain signal input. The specific feature extraction process is as follows:

$$A_{openSMILE} = openSMILE(A, IS*)$$
$$IS* \in \{IS09, IS11, IS13\} \quad (5)$$

2.1.3. Audio SoundNET Coded Feature Mining for Emotions

SoundNET is an acoustic representation pre-training neural network, the training process of the network is shown in Figure 1. According to Figure 1, SoundNET first decomposes the unlabeled video data into RGB image frame sequence and audio waveform sequence; for each RGB image frame, SoundNET inputs it into two pre-trained image convolutional neural networks, the first of which is ImageNET for reasoning about the specific class of objects appearing in the image, and the second is Places, for reasoning about the natural scene to which the image belongs; For each audio signal, SoundNET is trained using the 1-dimensional full convolutional network ConvNET, and the final inference layer has two channels, corresponding to the object category information and the scene information obtained by the image frame inference [2].

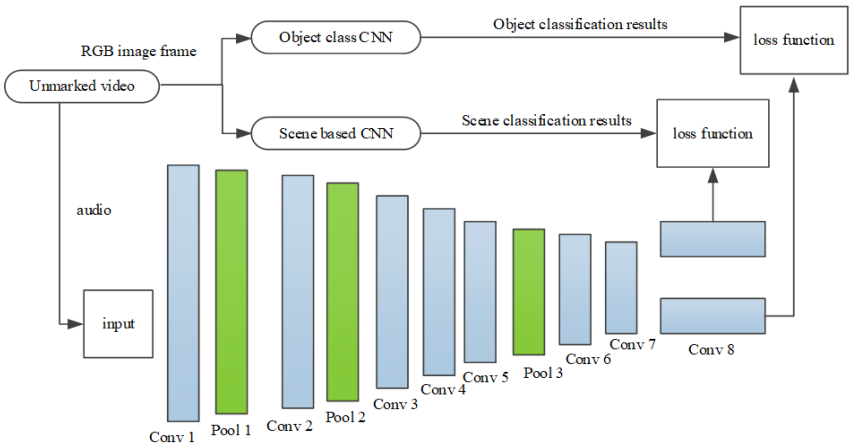

Figure 1. Architecture of the SoundNET network.

By pre-training SoundNET on a large number of unlabeled video signals, it is possible to have a rich and comprehensive acoustic representation of the audio signals corresponding to the convolutional network learning. After the model pre-training is completed SoundNET to obtain some of the audio features. The process can be simplified as Equation (6):

$$A_{soundNET} = soundNET(A) \tag{6}$$

The above three features are spliced to obtain the total emotional audio features, namely:

$$X^a = [A_{MFCC}, A_{openSMILE}, A_{soundNET}] \tag{7}$$

2.2. Video Feature Mining Sentiment Modeling

Emotional video sequences have temporal dynamic properties, i.e., the relationship between data points changes over time, and these sequences also contain rich emotional expressions (e.g., facial expressions, actions, etc.). RNNs can efficiently handle such temporal dynamic properties by virtue of their internal recurrent connections and hidden state update mechanism. When mining emotional video feature vectors for sequential emotion modeling using RNN, the hidden states of RNN can accurately capture the temporal dependencies between video frames, which significantly enhances the accuracy of mining features in video, and accordingly improves the accuracy of emotion recognition [5, 6]. To this end, RNN is chosen to construct the emotional feature mining model for emotional videos, and the model structure is shown in Figure 2.

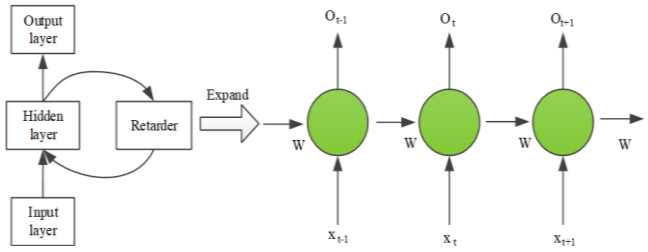

Figure 2. Affective feature mining model for emotional videos.

Figure 2 shows that the hidden layer of the model is a recursive structure, and the input of each hidden layer is jointly determined by the output of the hidden layer in the previous time step and the output of the input layer in the current time step. At the same time, the weights of each hidden layer are shared. The specific formulas for emotional video feature mining under RNN are as follows:

$$\begin{aligned} O_t &= g(V \cdot s_t) \\ s_t &= f(U \cdot x_t + W \cdot s_{t-1}) \end{aligned} \tag{8}$$

Among them, x_t denotes that in the time series of emotion video parsing, the output value of the input layer at the moment t, which carries visual characteristics data such as picture color, character expression contour, movement amplitude, etc., extracted and preprocessed from the current video frame. S_t represents the value of the hidden layer at the moment t, as a "memory" that constantly accumulates and updates the emotional features of the video, fuses the features of the past video clips with the new input information of the present time. U represents the weights between the input layer and the hidden layer, which is responsible for controlling the weight distribution and mapping of the original visual feature information passed from the input layer to the hidden layer. W denotes the weight of the hidden layer connections, which controls the influence of the previous state of the hidden layer on the current state of the hidden layer in the process of time continuation, and the feature transfer rules. V represents the weight between the hidden layer and the output layer, which determines the quantitative relationship between the stage-by-stage results of emotional video feature mining contained in the hidden layer and the output layer, so as to realize the efficient and orderly mining of emotional video features.

When using RNN to carry out emotional video feature mining, RNN has a good performance in conventional emotional video sequence processing, especially in the face of relatively short sequences, limited information span of emotional video segment parsing scenarios, it can effectively capture the correlation features between neighboring video frames, and help emotional cue mining. However, when the length of the emotional video sequence is long, due to the instability of the gradient back propagation mechanism in the long sequence transmission process, RNN will be trapped in the gradient explosion or gradient disappearance, it is difficult to accurately capture the deep dependency association between the beginning and the end of the emotional video and the video frames far away from each other, and some key features hidden in the depths of the long video sequences, which are related to the emotional ups and downs, are easily missed, severely limiting the ability to capture the emotional cues. Some key features hidden deep in the long video sequences, which are related to the changes of emotions, are easily missed, which seriously limits the effectiveness of comprehensive and in-depth feature mining of emotional videos [8, 17]. For this reason, LSTM are introduced, and its network structure is shown in Figure 3.

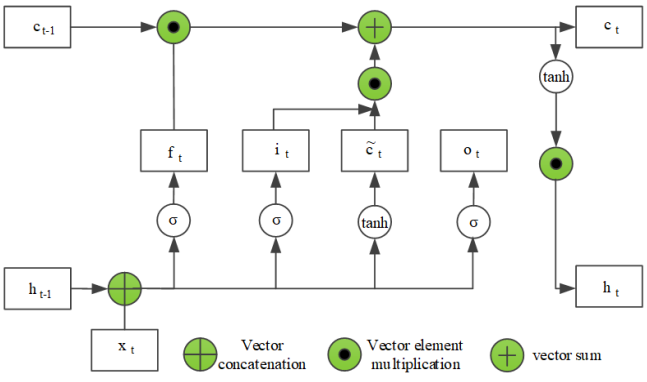

Figure 3. LSTM cell structure.

According to Figure 3, it can be seen that the LSTM takes the internal state c_{t-1} accumulated in the previous moment (the core emotional feature information embedded in past video frames and retained through filtering) and the external state of the previous moment (a slightly shorter-term, instantaneous state that flexibly reflects the emotional fluctuations of the approaching video frames) are taken as inputs, and three gates are introduced, i.e., the forgetting gates f_t, input gate i_t, output gates o_t, under the synergistic operation of these three gates, it is decided which emotional video information should be retained, which should be discarded, and which should be outputted and delivered. At the same time, with the help of tanh function and Sigmoid function as activation function, vector multiplication, vector sum, vector splicing and other operations are introduced, and finally, the current internal state c_t and the current external state h_t as the output and fed to the next loop cell. The main idea of LSTM is that by forgetting the gate f_t and input gates i_t updating the hidden state, i.e., the external state h_t (i.e., short-term memory), and provide output gates o_t to pass the transmits information of the internal state c_t (the retaining information is longer than short-term memory h_t but much less so than long-term memory, also known as long- and short-term memory), to the external states h_t.

Of which, the forgetting gate f_t maps the feature information into the [0, 1] interval through the Sigmoid function to achieve a selective "forgetting" of the information in the internal state c_{t-1} of the previous moment; input gates i_t and the output gate o_t take the same Sigmoid function to realize the candidate states $\tilde{c}_t$ for the current moment, preservation or output of information in the internal state of the current moment c_t [17]. The specific realization of the process for forgetting gate f_t, input gate i_t and output gates o_t are shown in Equation (9).

$$\begin{aligned} f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\ i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \end{aligned} \tag{9}$$

Second, the candidate state at the current moment $\tilde{c}_t$ is generated by the hidden state of the previous moment h_{t-1} and the input at the current moment x_t through the tanh function. Then we have:

$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{10}$$

Finally, the hidden state h_t at the current moment is similarly generated by the internal state c_t at the current moment through the tanh activation function and used for the input of the LSTM unit in the next layer. The computational procedure is as follows:

$$h_t = o_t * \tanh(c_t) \tag{11}$$

Through the above process, the emotional video features are mined, and the total features are represented in X^b. In this regard, the cepstrum of Meier's frequency extracted to the emotional features is shown in Figure 4.

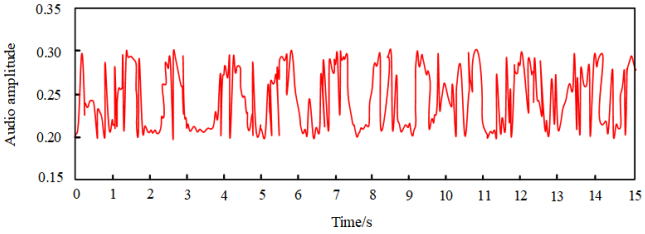

Figure 4. Mel frequency cepstrum of emotional traits.

3. Audio and Video Bimodal Emotion Recognition

3.1. Audio and Video Modal Feature Fusion

Cross-modal learning is an advanced technique that is capable of processing and fusing feature data from different modalities (e.g., video, audio, etc.). In the emotion recognition task, audio and video, as two crucial modalities, provide both acoustic and visual information, respectively. Through cross-modal learning, the complementary features of these two modalities can be fully utilized to improve the accuracy of emotion recognition. As an effective feature selection and weighting method, the attention mechanism can dynamically assign weights to features of different modalities [12, 16]. Introducing the attention mechanism in cross-modal learning can determine the importance of different modal feature information, thus fusing multi-source data more effectively. Especially in the process of fusion of emotional audio and video features, the attention mechanism can focus on emotion-related features such as tone and volume in audio, and facial expressions and movements in video, so as to realize more accurate emotion recognition. Therefore, a CMA approach is chosen to fuse audio and video features to further improve the accuracy of emotion recognition.

The process of fusing audio and video modal features based on CMA is as follows:

First, X^a and X^b are projected onto the common vector space, the dimensions of all these features are d_u, the projective operations is:

$$\begin{aligned} \tilde{X}^a &= M_{X^a} X^a + a_{X^a} \\ \tilde{X}^b &= M_{X^b} X^b + a_{X^b} \end{aligned} \tag{12}$$

Among them, M_{X^a} and M_{X^b} are projection matrices for the audio and video modes, respectively. a_{X^a} and a_{X^b} corresponding to the bias term X^a and X^b respectively, which assists in adjusting the positional offset after projection, ensures that the projected features can fit the unified space while retaining the core affective ideographic information. Based on Equation (12), the mathematical expression of CMA mechanism can be obtained as follows:

$$X_{x,y} = CMA(x,y) = \tilde{X}x softmax(\frac{\tilde{X}_x^i(\tilde{X}_Y^i)}{\sqrt{d}}) \tag{13}$$

Where, $X_{x,y}$ represents a new cross-modal feature generated by modal x and modal y through the fusion of cross-modal attention, and $\tilde{X}_x^i$ represents the i-th linear projection under dimension d. d is the dimension of each projection space. The cross-modal layered attention aims to achieve modal interaction in order to better mine the interrelated information in emotional audio and video modalities [19]. The results of each modal interaction are shown in Equation (14).

$$\begin{aligned} X_{X^a,X^b} &= CMA(X^a, X^b) \\ &= \tilde{X}_{X^a}^i softmax(\frac{\tilde{X}_{X^a}^i (\tilde{X}_{X^b}^i)^T}{\sqrt{d}}) \\ X_{X^b,X^a} &= CMA(X^b, X^a) \\ &= \tilde{X}_{X^b}^i softmax(\frac{\tilde{X}_{X^b}^i (\tilde{X}_{X^a}^i)^T}{\sqrt{d}}) \end{aligned} \tag{14}$$

In order to fuse these interactions, the information is spliced with:

$$X_m = X_{X^a,X^b} \oplus X_{X^b,X^a} \tag{15}$$

After the operation specified in Equation (13), the fusion of audio and video features of emotions can be realized. In order to further capture richer information between modalities and thus strengthen the fusion results, take and construct a modal treatment opposite to X_m. Specifically, the neural network G_Φ with the parameters Φ is applied to achieve this goal. G_Φ is used to interpret the similarity measure between the input X_m and the original mode X^a, X^b. Based on the results of this similarity measure, unimodal features after modal reconstruction are generated. In order to effectively evaluate the accuracy of the prediction and to measure the degree of agreement between the reconstructed unimodal features and the original unimodal features, the Mean Square Error (MSE) is selected as the loss function. Its mathematical expression is shown as follows:

$$\begin{aligned} G_\Phi^\delta(X_m) &= \sigma(\frac{1}{N^\delta}\sum_{i=1}^{N^\delta} D(X_m, \Phi_\Phi^\delta) + a_\Phi) \\ r_{X^a} &= G_\Phi^{X^a}(X_m) \\ r_{X^b} &= G_\Phi^{X^b}(X_m) \end{aligned} \tag{16}$$

Among them $D(\cdot)$ is the similarity measure function, σ is the activation function, a_Φ is the learned bias term G_Φ. This neural network generates the predictive features from X_m.

$$\begin{aligned} r_{X^a} &= G_\Phi^{X^a}(X_m) \\ r_{X^b} &= G_\Phi^{X^b}(X_m) \end{aligned} \tag{17}$$

The loss function for backward prediction is the MSE as follows:

$$\begin{aligned} L_{REC} &= L_{REC}(X^a) + L_{REC}(X^b) \\ &= (X^a - X^b)^2 + (X^b - X^a)^2 \end{aligned} \tag{18}$$

According to the result of Equation (18), the parameters and operation logic of the neural network are continuously optimized, so that the reconstructed unimodal features can be closer to the original unimodal features, and then achieve the purpose of improving the overall modal fusion effect and exploring more intermodal correlation information.

3.2. Decision-Level Fusion Recognition

Decision-level fusion has the ability to fuse features of different natures in machine learning problems, and has significant advantages such as flexibility and efficiency. In view of this, the technique is used to determine the current emotional state of an individual, and the emotional categories involved cover nine states: joy, trust, anxiety, fear, surprise, sadness, anticipation, anger and disgust. In the specific implementation process, multimodal sentiment analysis was carried out through decision-level fusion, setting β, χ as hyperparameters measuring the importance of audio and video modal prediction results, respectively, their fused sentiment analysis results are expressed as follows:

$$\hat{y}_m = \frac{1}{1+\beta+\chi}(y_m + \beta y_{X^a} + \chi y_{X^b}) \tag{19}$$

In order to ensure the accuracy of sentiment analysis and to continuously optimize model performance, introduce the $L1$ loss function as the basis of classification and evaluation, then there are:

$$L_c = \frac{1}{N}\sum_i^N (|\hat{y}_m^i - \hat{y}_m^{i*}| + M_S^i \cdot |y_S^i - y_S^{i*}|) \tag{20}$$

Among them, N is the number of training samples, $M_S^i \cdot |y_S^i - y_S^{i*}|$ is the weight of the i-th unimodal sample in the modal reconstruction task. $\hat{y}_m$ and y_S are the predicted result. $\hat{y}_m^*$ and $\hat{y}_S^*$ are the true result. The total loss in emotion recognition consists of both categorization loss and reconstruction loss, namely:

$$L = L_c + L_{REc} \tag{21}$$

Continuously optimize the judgment results based on Equation (21) to achieve high-quality emotion recognition and understand the individual's current mental health.

4. Results and Analysis of Simulation Experiments

4.1. Experimental Environment and Parameter Settings

The subjects in a mental health counseling room were selected as the subjects of this experiment, and the layout of the counseling room is shown in Figure 5. In the experiment, a Logitech C920e video capture device and a Sennheiser MK4 audio capture device was used to capture the subjects' voice signals, facial expressions, and motion images, and the parameters of the devices are shown in Table 1. The operating system used was Ubuntu 18.04, and the programming language was Python 3.7. A Dell PowerEdge R750 server was used to process and store the audio/video data, a Myriad PM-

9000Express cardiac monitor was used to obtain the physiological signals of the subjects, and a Sony BDP-S6700 Blu-ray player was used to play the video clips to induce the emotional changes of the subjects.

In the device environment, the specific data collection and annotation settings are shown in Table 2 below:

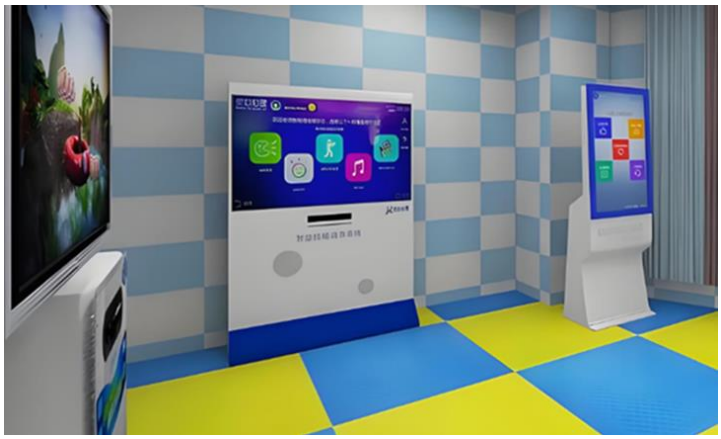

Figure 5. Emotion recognition site map.

Table 1. Details of equipment parameters.

Equipment name	The relevant parameters	Details/Description
Logitech C920e video capture device	diagonal field of view	78°
	resolution and frame rate	1080p/30fps and 720p/30fps
	autofocus with automatic light correction	Built-in HD autofocus ensures clear picture quality during video calls; and with RightLight 2 automatic light correction, it delivers clear images in a wide range of lighting environments, whether it's in low light or direct sunlight.
	privacy	Features a removable privacy lens cover that can be flipped up and down to conveniently cover or expose the lens for physical privacy.
	connectivity	Plug and play via USB-A interface, cable length 1.5 meters.
Sennheiser MK4 audio capture device	frequency response range	20Hz~20kHz
	sensitivity	25mV/Pa
	product sound pressure	140 dB max
	shape	Diameter 57mm (max), length 160mm
	weight	485g

Table 2. Data collection and annotation settings.

Attribute	Details	In detail
participant	85 people (aged 18-65)	Gender distribution (42 males/43 females), clinical grouping (60 healthy individuals/15 depressed individuals/10 anxious individuals)
Collection environment	Standardized psychological counseling room	Light control (500-700 lux), background noise (<35 dB)
Stimulative materials	IAPS images+self-selected videos	Stimulus library accounts for 70%
Labeling protocol	Three-layer annotation mechanism	1. Russell's circular model annotation by expert psychologists (valence/arousal) 2. Participants' Self-Assessment Manikin (SAM) scale 3. Cross validation of physiological signals Electrodermal Activity (EDA)/Heart Rate Variability (HRV)
data distribution	Statistics of 9 types of emotional samples	Including trust, joy, anxiety, fear, surprise, sadness, expectation, anger, and disgust

4.2. Analysis of the Effectiveness of Audio and Video Bimodal Emotion Recognition

In order to verify the effect of the proposed algorithm in audio-video bimodal emotion recognition, the nine emotions of trust, joy, anxiety, fear, surprise, sadness, anticipation, anger, and disgust in the emotional audio are divided into nine grades, and each grade corresponds to a numerical value range, which are as follows: 0-0.1 for trust, 0.1-0.2 for joy, 0.2-0.3 for anxiety, 0.3-0.4 for fear, 0.4-0.5 for surprise, 0.5-0.6 for sadness, 0.6-0.7 for anticipation, 0.7-0.8 for anger,0.8-1.0 disgust.

Audio feature extraction parameter settings: the number of Mel filter banks is 40, the frame length is 25ms, and the frame shift is 10ms; video feature extraction parameter settings: the number of RNN hidden layer units is 128, and the number of LSTM units is 256; Cross modal fusion parameter settings: attention mechanism dimension is 64, projection matrix dimension is 128; Decision level fusion: hyperparameters β and χ (hyperparameters that measure the importance of audio and video modal prediction results) are set to 0.6 and 0.4, respectively. The dataset is divided into training set, validation set, and testing set in an 8:1:1 ratio. Using random shuffling for segmentation to ensure the randomness and representativeness of the data.

Using a 5-fold cross validation method, the model is trained and tested multiple times to evaluate its stability and generalization ability. The model was trained on Dell PowerEdge R750 servers using NVIDIA Tesla V100 GPU acceleration. The average training time is

about 2 hours per round. The model architecture provides a detailed description of the network structure for audio feature extraction, video feature extraction, cross modal fusion, and decision level fusion, including the input-output dimensions and activation functions of each layer.

In the experimental process, assuming that the actual emotion of the subject is the anxiety state, the proposed algorithm is used to recognize the emotion of the audio and video clips. The recognition results are shown in Figure 6.

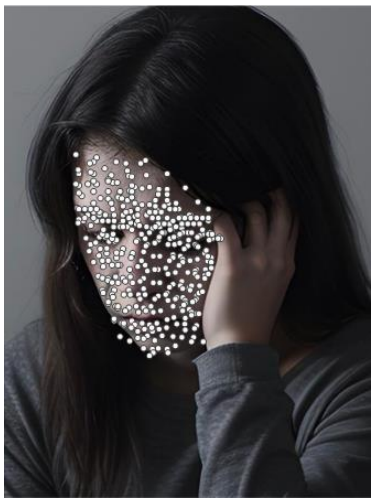

Figure 6. Emotion video recognition analysis.

In Figure 6, the character's expression is tense, his eyes are flickering, his body language reveals uneasiness, and his overall emotion is anxiety. After the proposed algorithm, the emotional video recognition result of anxiety state is the same as the actual situation, which is because it is based on RNN to realize the video feature mining. RNN can share the weights between different time steps, so as to efficiently process each element in the video sequence, and use the previous information to influence the current output. This capability allows the RNN to learn the long-term dependencies in the video sequence and thus understand the emotional expressions in the video.

4.3. Performance Analysis of Audio and Video Bimodal Emotion Recognition Algorithms

4.3.1. Weighted Accuracy (WA)

Weighted Accuracy (WA) is an important index to measure the overall performance of audio-video bimodal emotion recognition algorithms, which takes into account the ratio of the number of samples correctly predicted by the model for each emotion category to the total number of samples. In the audio/video bimodal emotion recognition task, which usually involves multiple emotion categories (e.g., happy, sad, angry, calm, etc.,) the WA can reflect the accuracy of the algorithm in categorizing these emotion categories as a whole.

Experiments through the device were collected nine types of emotions each 100 audio and video samples, a total of 9000 audio and video samples, these samples randomly disrupted randomly generated 15 groups, respectively, using the proposed algorithm with LSTM, DEAN on these audio and video samples for emotion recognition, comparison and analysis of the WA of the recognition results of each method, the results are shown in Figure 7.

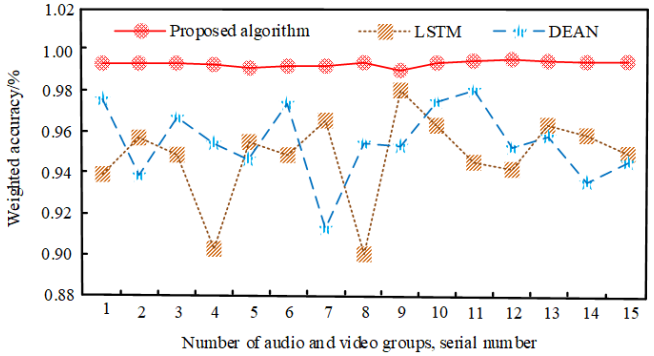

Figure 7. WA analysis of emotion recognition results for each method.

According to the results shown in Figure 7, the proposed algorithm consistently yields higher WA values of emotion recognition results than LSTM and DEAN methods in the emotion recognition task for audio and video samples. This excellent performance is mainly attributed to the CMA fusion strategy adopted by the proposed algorithm, which can effectively integrate the complementary information from two different modalities, i.e., audio and video. In this way, the algorithm's ability to analyze complex emotional signals is significantly improved, and its adaptability to emotional changes in different contexts is enhanced. As a result, the proposed algorithm not only exhibits higher accuracy but also greater robustness than LSTM and DEAN methods in the emotion recognition task.

4.3.2. Unweighted Accuracy

Unweighted Accuracy (UA), also known as the average recall of sentiment categories, does not take into account the differences in the sample size of each sentiment category, but simply averages the recall of each sentiment category. This can avoid the situation that the overall assessment results are biased in favor of certain categories due to the excessive number of samples in these categories. Therefore, the UA metric is chosen to evaluate the performance of the audio-video bimodal emotion recognition algorithm, which can more comprehensively and accurately reflect the performance of the algorithm on different emotion categories.

The experimental environment was unchanged, and the UA values of emotion recognition of each group under each method were compared and analyzed, and the results are shown in Figure 8.

According to Figure 8, the proposed algorithm's UA in the audio-video emotion recognition task is particularly outstanding. This is mainly due to its decision-level fusion strategy. This strategy makes full use of the complementary nature of different modal information, and fuses the emotional features of both video and audio modalities after they have been extracted and classified separately. The advantage of this approach is that it avoids the problem of high feature dimensionality caused by feature splicing, and thus reduces the computational complexity. At the same time, since the fusion occurs after the classification, the information of different modalities can be integrated more effectively, which significantly improves the UA of emotion recognition.

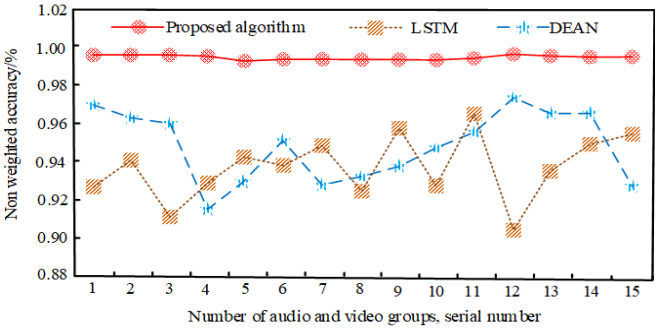

Figure 8. UA analysis of the emotion recognition results of each method.

4.3.3. Correlation Coefficient

Correlation coefficient is an important index to measure the performance of audio-video bimodal emotion recognition methods, which is usually used to assess the degree of linear correlation between predicted and actual emotions, thus reflecting the accuracy and reliability of emotion recognition algorithms. The correlation coefficients of the 15 samples under each method were calculated, and the results are shown in Figure 9.

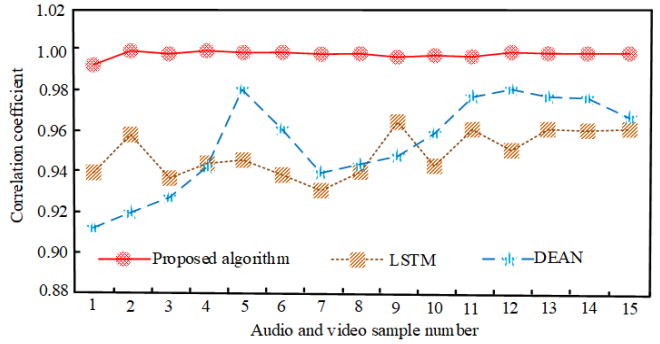

Figure 9. Analysis of correlation coefficients of the results of emotion recognition by each method.

In Figure 9, the correlation coefficients of this method for each type of emotion recognition results are above 0.98, while the correlation coefficients of the LSTM method for each type of emotion recognition results do not exceed 0.96, and the correlation coefficients of the DEAN method for each type of emotion recognition results do not exceed 0.98, which indicates that the correlation coefficients of this method for the recognition results of the various samples of emotions with the actual emotions are the largest, and the reliability of the recognition is effectively improved. This is because the method in this paper is able to correct the emotion recognition results through the loss function, which effectively improves the recognition accuracy.

4.3.4. Emotion Recognition Latency

This study used a self-collected dataset and compared it with the publicly available Interactive Emotional dyadic Motion CAPture (IEMOCAP) dataset to further validate the practicality of the design method. The IEMOCAP dataset is a multimodal emotion recognition dataset collected by the Speech Analysis and Interpretation Laboratory (SAIL) at the University of Southern California (USC). This dataset contains approximately 12 hours of audiovisual data, involving video, speech, facial motion capture, and text transcription, recorded by 10 actors (5 males and 5 females) in scripted and improvised scenes, aimed at eliciting specific emotional expressions. The test results are shown in Table 3 below:

Table 3. Emotion recognition delay/ms.

Data volume	Self-built dataset			IEMOCAP dataset		
	Proposed algorithm	LSTM	DEAN	Proposed algorithm	LSTM	DEAN
100	17	39	40	16	40	41
200	18	40	44	17	42	42
300	18	42	45	18	43	46
400	19	45	46	18	45	48
500	21	48	46	19	48	49

On self-built datasets, the proposed algorithm has lower latency than LSTM and DEAN algorithms at different data volumes. With the increase of data volume, the latency of all three algorithms has increased, but the latency growth of the proposed algorithm is relatively small. On the IEMOCAP dataset, the proposed algorithm also exhibits lower latency, and the latency at different data volumes is lower than that of LSTM and DEAN algorithms. The results of this experiment show that the proposed algorithm has significant advantages in emotion recognition and exhibits stable low latency performance on different datasets and data volumes.

4.3.5. Confusion Matrix

In order to further analyze the error classification of the test method, confusion matrix analysis was carried out. The results are shown in Figure 10 below:

According to the table above, the average accuracy is 93.4% (mean of the main diagonal), with the three most easily confused emotions being anxiety→fear (4%), disgust→anger (4%), and trust→expectation (4%). Trust is 93% accurate, with a main misjudgment of 4%

expectation and 3% joy, because all three emotions are expressed as positivity. Anxiety is 91% accurate, with the main misjudgments being 4% fear and 2% disgust, due to the overlapping key features of the frown muscle activity in these three emotions. Although there is a certain degree of error, the accuracy is still higher than 90%. This is because RNN and LSTM can capture temporal dependencies in video sequences, especially LSTM, which can effectively process long sequence data and capture subtle expressions and motion changes in videos through its internal forget gate, input gate, and output gate mechanisms.

	1	2	3	4	5	6	7	8	9	
1	0.93	0.03	0.02	0	0	0	0.04	0	0	1.trust
2	0.02	0.94	0	0	0	0	0.04	0	0	2.joy
3	0	0	0.91	0.04	0.02	0.01	0	0	0.02	3.anxiety
4	0	0	0.04	0.94	0	0	0	0.01	0.01	4.fear
5	0.03	0.01	0	0	0.94	0	0.02	0	0	5.amazed
6	0	0	0.02	0.01	0	0.95	0	0	0.02	6.sad
7	0.02	0.03	0	0	0	0	0.95	0	0	7.expectation
8	0	0	0.01	0	0	0	0	0.96	0.03	8.anger
9	0	0	0.02	0.01	0	0	0	0.04	0.92	9.hate

Figure 10. Confusion matrix.

5. Conclusions

Given that the accurate capture and analysis of emotional state is crucial for mental health monitoring, early warning and intervention, and the current traditional emotion recognition methods are limited by unimodal information, it is difficult to comprehensively and efficiently understand the whole picture of individual emotions, and in practical applications, it has limited help for personalized psychological assessment and diagnosis. To this end, a bimodal emotion recognition algorithm based on emotion modeling is proposed. The main research contents are as follows:

1. In this paper, in response to the difficulty of the traditional method of relying too much on a single modal information resulting in poor recognition results, the emotion modeling is divided into two parts. The first part utilizes Fourier transform, openSMILE tool and SoundNET neural network to dig into the key features of audio modality, i.e., to excavate the features that can reflect the change of the rhythm of the sound and the emotional implication of the timbre texture; the second part uses RNN to capture the emotional representations of the video modality with the change of the time series in depth. In the second part, we use RNN to capture the emotion in the video modality with the time sequence. We fully extract the unique and valuable feature information of the emotion in both audio and video modalities.
2. Inter-modal interaction enhancement and feature depth fusion, with the help of CMA mechanism, according to the importance of different modal information, differentiated allocation of weights, so that the audio and video inter-modal interaction can be strengthened to achieve complementary fusion of information; and then use the decision-level fusion network to summarize and weigh the multimodal fusion and reconstructed comprehensive information, and to comprehensively determine the emotional category, so as to enhance the performance of the algorithm in all aspects and the robustness to deal with complex scenarios.
3. After rigorous experimental verification, the proposed algorithm shows good emotion recognition ability, and its WA is excellent, capable of accurately recognizing the emotional state, and providing strong technical support for the efficient development of mental health-related work.

Acknowledgment

University-level Scientific Research, project of Southwest Medical University: Value Exploration and Utilization of Psychological Education Resources of Excellent Traditional Chinese Culture from the Perspective of School-Family-Society Collaborative Education 2024SKYB008. Project of Sichuan Hospital Management and Development Research Center, An inquiry into the construction of humanistic care in psychiatry-A case study of Psychiatry in Affiliated Hospital of Southwest Medical University, (SCYG2023-22).

References

[1] Antonino V., Chiara B., and Giovanna M., "The Effect of Emotion Intensity on Time Perception: A Study with Transcranial Random Noise Stimulation," *Experimental Brain Research*, vol. 241, no. 8, pp. 2179-2190, 2023. DOI:10.1007/s00221-023-06668-9

[2] Demiris G., Oliver D., Washington K., Chadwick C., and et al., "Examining Spoken Words and Acoustic Features of Therapy Sessions to Understand Family Caregivers' Anxiety and Quality of Life," *International Journal of Medical Informatics*, vol. 160, pp. 104716, 2022. DOI: 10.1016/j.ijmedinf.2022.104716

[3] Dhelim S., Chen L., Ning H., and Nugent C., "Artificial Intelligence for Suicide Assessment using Audiovisual Cues: A Review," *Artificial Intelligence Review*, vol. 56, no. 6, pp. 5591-5618, 2022. DOI: 10.1007/s10462-022-10290-6

[4] Fu Y., Huang B., Wen Y., and Zhang P., "FDR-MSA: Enhancing Multimodal Sentiment Analysis Through Feature Disentanglement and Reconstruction," *Knowledge-Based Systems*, vol. 297, no. 3, pp. 1-12, 2024. DOI:

10.1016/j.knosys.2024.111965
[5] Garg R., Gao R., and Grauman K., "Visually-Guided Audio Spatialization in Video with Geometry-Aware Multi-Task Learning," *International Journal of Computer Vision*, vol. 131, no. 10, pp. 2723-2737, 2023. DOI: 10.1007/s11263-023-01816-8
[6] Hussain S., Chalicham N., Garine L., Chunduru S., and et al., "Low-Light Image Restoration Using a Convolutional Neural Network," *Journal of Electronic Materials*, vol. 53, no. 7, pp. 3582-3593, 2024. DOI: 10.1007/s11664-024-11079-9
[7] Lahoti G., Ranjan C., Chen J., Yan H., and Zhang C., "Convolutional Neural Network-Assisted Adaptive Sampling for Sparse Feature Detection in Image and Video Data," *IEEE Intelligent Systems*, vol. 38, no. 1, pp. 45-57, 2023. DOI: 10.1109/MIS.2022.3215779
[8] Lee C., Ortiz J., Glenn C., Kleiman E., and Liu R., "An Evaluation of Emotion Recognition, Emotion Reactivity, and Emotion Dysregulation as Prospective Predictors of 12-Month Trajectories of Non-Suicidal Self-Injury in an Adolescent Psychiatric Inpatient Sample," *Journal of Affective Disorders*, vol. 358, no. 1, pp. 302-308, 2024. DOI: 10.1016/j.jad.2024.02.086
[9] Lin M., Wu J., Meng J., Wang W., and Wu J., "State of Health Estimation with Attentional Long Short-Term Memory Network for Lithium-Ion Batteries," *Energy*, vol. 268, no. 1, pp. 126706, 2023. DOI: 10.1016/j.energy.2023.126706
[10] Liu J., Wang Z., Nie W., Zeng J., and et al., "Multimodal Emotion Recognition for Children with Autism Spectrum Disorder in Social Interaction," *International Journal of Human-Computer Interaction*, vol. 40, no. 5/8, pp. 1921-1930, 2024. DOI: 10.1080/10447318.2023.2232194
[11] Middya A., Nag B., and Roy S., "Deep Learning Based Multimodal Emotion Recognition Using Model-Level Fusion of Audio-Visual Modalities," *Knowledge-Based Systems*, vol. 244, no. 23, pp. 108580, 2022. DOI: 10.1016/j.knosys.2022.108580
[12] Sayed H., Eldeeb H., and Taie S., "Bimodal Variational Autoencoder for Audiovisual Speech Recognition," *Machine Learning*, vol. 112, no. 4, pp. 1201-1226, 2023. DOI: 10.1007/s10994-021-06112-5
[13] Tian M., Dong H., Cao X., and Yu K., "Temporal Convolution Network with a Dual Attention Mechanism for φ-OTDR Event Classification," *Applied Optics*, vol. 61, no. 20, pp. 5951-5956, 2022. DOI: 10.1364/AO.458736
[14] Wang Z. and Zuo R., "Mineral Prospectivity Mapping Using a Joint Singularity-Based Weighting Method and Long Short-Term Memory Network," *Computers and Geosciences*, vol. 158, pp. 104974, 2022. DOI10.1016/j.cageo.2021.104974
[15] Wu X., Zhang X., Feng X., Lopez M., and Liu L., "Audio-Visual Kinship Verification: A New Dataset and a Unified Adaptive Adversarial Multimodal Learning Approach," *IEEE Transactions on Cybernetics*, vol. 54, no. 3, pp. 1523-1536, 2024. DOI: 10.1109/TCYB.2022.3220040
[16] Zhang F., Li X., Lim C., Hua Q., and et al., "Deep Emotional Arousal Network for Multimodal Sentiment Analysis and Emotion Recognition," *Information Fusion*, vol. 88, no. 12, pp. 296-304, 2022. DOI: 10.1016/j.inffus.2022.07.006
[17] Zhang J., Jiang Y., Wu S., Li X., and et al., "Prediction of Remaining Useful Life Based on Bidirectional Gated Recurrent Unit with Temporal Self-Attention Mechanism," *Reliability Engineering and System Safety*, vol. 221, pp. 108297, 2022. DOI: 10.1016/j.ress.2021.108297
[18] Zhang R., Qin B., Zhao J., Zhu Y., and et al., "Locating X-Ray Coronary Angiogram Keyframes via Long Short-Term Spatiotemporal Attention with Image-to-Patch Contrastive Learning," *IEEE Transactions on Medical Imaging*, vol. 43, no. 1, pp. 51-63, 2024. DOI: 10.1109/TMI.2023.3286859
[19] Zhang Y., Wu L., Wang J., and Li S., "Multi-Modal Emotion Recognition Based on Multi-LSTMs Fusion," *Journal of Chinese Information Processing*, vol. 36, no. 5, pp. 145-152, 2022. DOI: 10.3969/j.issn.1003-0077.2022.05.015
[20] Zhu Q. and Peng Y., "Semi-Supervised Kernel Discriminative Low-Rank Ridge Regression for Data Classification," *The International Arab Journal of Information Technology*, vol. 21, no. 5, pp. 800-814, 2024. DOI:10.34028/iajit/21/5/3
[21] Zong L., Zhou J., Xie Q., Zhang X., and Xu B., "Multi-modal Emotion Recognition Based on Hypergraph," *Journal of Computer Science*, vol. 46, no. 12, pp. 2520-2534, 2023. DOI: 10.11897/SP.J.1016.2023.02520

Yang Liu received Bachelor's degree in Medical Imaging from Chongqing Medical University in 2010 and Master's degree in Psychiatry from Chongqing Medical University in 2015, majoring in Psychiatry and Clinical Psychology. Work experience: July 2010-October 2012, Physician, Changshou District Hospital of Traditional Chinese Medicine, Chongqing, July 2015-present, Physician, Department of Psychosomatic Medicine, the Affiliated Hospital, Southwest Medical University. Academic situation: Published 3 papers, 2 academic works and teaching materials, and presided over 3 topics.

Shudan Feng received a Science degree in Psychology from Zhoukou Normal University in 2010 and a Master's degree in Development and Educational Psychology from Southwest University in 2013. Research interests: Clinical Psychology, Mental Health. Work experience: July 2013-present, Teacher and Lecturer, School of Humanities and Management Science, Southwest Medical University. Academic situation: Published 10 papers, 4 academic works and teaching materials, and presided over 8 topics.

Kaiyong Li obtained Bachelor's degree in Electronic Physics from Qinghai Nationalities University in 1993, Master's degree in Computer Technology Application from Qinghai Normal University in 2010, research direction: Computer Technology Application, Image Processing. Work experience: From 1993 to 2015, worked as a lecturer in the Department of Telecommunications at Qinghai University for Nationalities. From 2016 to 2020, worked as an Associate Professor at the School of Physics and Electronic Information Engineering at Qinghai University for Nationalities. From 2021 to present, worked as a Professor at the School of Intelligent Science and Engineering at Qinghai University for Nationalities. Academic situation: Published more than 30 academic papers in domestic and foreign journals, authorized 15 patents, 2 computer software copyrights, led and completed key research and transformation projects in Qinghai Province, participated in 4 provincial and ministerial level projects, and achieved international advanced level research results. Selected as the leader of natural science and engineering technology discipline in Qinghai Province's "Kunlun Talents Science and Technology Leading Talents" in 2022.