J Inf Process Syst, Vol.17, No.4, pp.754~771, August 2021
https://doi.org/10.3745/JIPS.02.0161

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Audio and Video Bimodal Emotion Recognition in Social Networks Based on Improved AlexNet Network and Attention Mechanism

Min Liu* and Jun Tang*

Abstract

In the task of continuous dimension emotion recognition, the parts that highlight the emotional expression are not the same in each mode, and the influences of different modes on the emotional state is also different. Therefore, this paper studies the fusion of the two most important modes in emotional recognition (voice and visual expression), and proposes a two-mode dual-modal emotion recognition method combined with the attention mechanism of the improved AlexNet network. After a simple preprocessing of the audio signal and the video signal, respectively, the first step is to use the prior knowledge to realize the extraction of audio characteristics. Then, facial expression features are extracted by the improved AlexNet network. Finally, the multimodal attention mechanism is used to fuse facial expression features and audio features, and the improved loss function is used to optimize the modal missing problem, so as to improve the robustness of the model and the performance of emotion recognition. The experimental results show that the concordance coefficient of the proposed model in the two dimensions of arousal and valence (concordance correlation coefficient) were 0.729 and 0.718, respectively, which are superior to several comparative algorithms.

Keywords

AlexNet Networks, Attention Mechanism, Concordance Correlation Coefficient, Deep Learning, Feature Layer Fusion, Multimodal Emotion Recognition, Social Networks

1. Introduction

With the development of social network, people tend to use video instead of text to express their opinions on products or services [1,2]. The voice data in video expresses the speaker's mood, while visual data conveys facial expressions, all of which help us to understand the user's emotional state. In recent years, artificial intelligence researchers have sought to empower machines with the ability to perceive and recognize emotions in order to identify and express them, which is called sentiment analysis [3,4]. Sentiment analysis has become a new trend of social media, which can effectively understand opinions expressed by users on social network platforms.

Affective computing is a new interdisciplinary research field, which brings together researchers and practitioners from different fields such as artificial intelligence, natural language processing, cognitive science and social science [5,6]. With the proliferation of online videos about product reviews, movie reviews, political views, etc., the number of multimodal online content is exponentially growing. The

Manuscript received July 27, 2020, first revision September 15, 2020; accepted October 4, 2020.
Corresponding Author: Jun Tang (cstangjun2019@126.com)
* Software School, Hunan Vocational College of Science and Technology, Hunan, China (toliumin@163.com, cstangjun2019@126.com)

research on affective computing has evolved from the traditional single-mode analysis to the more complex multimodal analysis. The multi-pattern recognition is an important branch in the field of artificial intelligence, which enables machines to observe surrounding environment and to learn how to make corresponding judgments on different patterns [7,8].

Multimodal emotion recognition is a challenging topic in affective computing. It can be regarded as the fusion of information of different forms [9,10]. The multimodal fusion is a process in which data collected from various modes are combined to analyze tasks. It can extract discriminative features from multimodal data to identify subtle gaps in human emotions. In a video, voice data expresses speaker's mood, while visual data conveys facial expression, both of which help us to understand the user's emotional state. In order to better integrate multiple modal data to improve the performance of emotion recognition, it is necessary to discover a general multimodal sentiment analysis framework.

2. Related Research

Many methods have been proposed to solve the problem of multimodal emotion recognition. Earlier methods of continuous dimension emotion recognition mainly use manual features combined with traditional machine learning algorithms [11]. In [12], a multimodal emotion recognition method based on emotional audio and facial expressions is proposed. By extracting acoustic features and facial features related to human emotional expressions, product rules are used in the decision-level fusion, and the support vector machine (SVM) is used for classification, which improves the model's accuracy. Dobrisek et al. [13] proposed a multimodal emotion recognition system using audio and video information. By processing two information sources respectively to generate corresponding matching scores, and then combining the calculated matching scores with weights and weighted rules to obtain classification decisions, the accuracy of the model was successfully improved. Wang [14] proposed a multimodal emotion recognition method based on the edge network emotional element compensation and the data fusion. In [15], an emotion recognition algorithm based on decision rules for audio and video bimodal decision-level fusion is proposed. First of all, it performs single mode recognition of audio and facial expressions and tests, and then sets decision rules to fuse audio-visual information in decision-making level to identify emotions. Although the above methods have achieved satisfactory results, the high accuracy of existing models is mainly a result of the manual features provided by the database and high computational costs in model training and testing. Therefore, how to develop a more reasonable method for emotion recognition of multimodal continuous dimension has become a current challenge.

With the development of deep learning, convolutional neural network (CNN) has been applied to the field of dimension emotion recognition. In [16], a classifier fusion emotion recognition method based on Kalman filter is proposed. By treating the video as a time dynamic sequence, Kalman filter is used to combine a variable number of measurements to deal with the sensor failure in the same framework, which improves the accuracy of the model. In [17], an end-to-end multi-mode emotion recognition method using CNN is proposed. CNN is used to extract features from audios, and a 50-layer deep residual network is used to extract features from videos. Long and short-term memory (LSTM) networks are used to model the context; thus, the models' accuracy is enhanced. Huang et al. [18] proposed a method of multimodal emotion recognition based on the deep neural network transfer learning. By inputting the audio and visual features of extracted datasets into LSTM network to train the model, the ensemble learning method is

implemented in the decision-making level fusion to improve the accuracy of the model. In [19], a hybrid model was proposed. Firstly, CNN and 3D-CNN were used to generate audio-visual segment features, and then audio-visual segment features were fused into deep belief networks (DBNs). The method is divided into two stages. In the first stage, in the emotion recognition task, the CNN and 3D-CNN models trained in large-scale image and video classification tasks are finetuned to learn the features of audio and video clips, respectively. In the second stage, the output of CNN and 3D-CNN models are combined into a fusion network constructed by the DBN model. The training fusion network is used to learn the feature representation of different audio-visual segments. The segment features learned by DBN are averagely merged to form global video features of fixed length, and then the linear SVM is used to classify video emotions. In [20], a hierarchy modular neural network (HMNN) is proposed and applied to the multimodal emotion recognition. A HMNN is constructed to simulate the hierarchical modular structure displayed in human brain. Each module contains several sub-modules to process features from different patterns.

Although above methods have achieved expected results, the influence of different modes on the emotional state varies. For example, when the speaker is depressed, the low intonation is more likely to represent the current emotional state than the expressionless, which has not been paid attention to in above methods. Therefore, this paper studies the feature-level fusion of two most important modes (voice and facial expression) in emotion recognition, and proposes an improved AlexNet network combined with the attention mechanism for audio and video bimodal emotion recognition. The main contributions of the proposed method are summarized as follows:

1) The improved AlexNet network is used to extract facial expression features instead of the traditional AlexNet network. The full connection layer is realized by CNN. By using the feature of convolution layer weight sharing, the number of parameters needed to be trained in the full connection layer is reduced and the network training pressure is reduced.
2) Multimodal attention mechanism is used to calculate the contribution of two modes to the emotion recognition, and to fuse the two modes. It solves the problem of incomplete expression from the single mode information and improves the recognition accuracy.

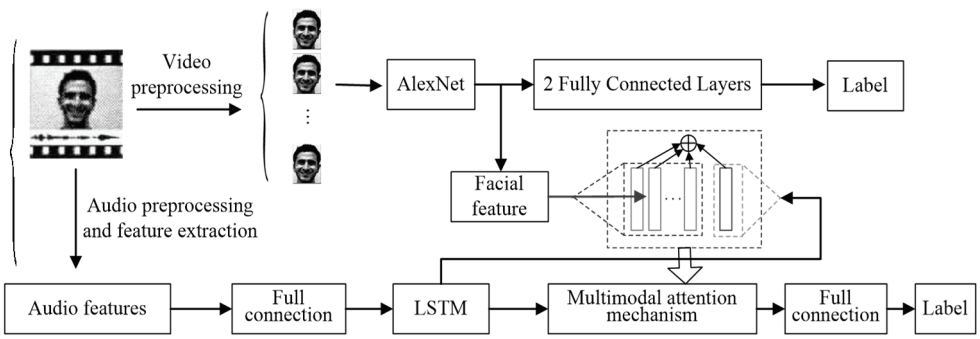


Fig. 1. Model structure of audio and video bimodal emotion recognition algorithm.

3. Overall Framework

The structure of audio and video bimodal emotion recognition model is shown in Fig. 1. Firstly, the audio signal and video signal are preliminarily preprocessed. Then, in the aspect of audio feature

extraction, the method of feature extraction based on the prior knowledge is adopted. In terms of the feature extraction of facial expressions, the improved AlexNet network is used to extract facial expression features. Then, facial expression features and audio features are fused based on the multimodal attention mechanism. Finally, in the model test stage, test videos are input into the fully trained multimodal attention mechanism to predict the final emotion. This paper will first introduce the audio and video preprocessing and feature extraction process.

4. Audio and Video Preprocessing and Feature Extraction

4.1 Audio Signal Preprocessing and Feature Extraction

First, the audio signal captured by the device is an analog signal; therefore, it needs to be converted into a digital signal before the feature can be extracted in order to proceed with the subsequent processing. Secondly, the non-related noise in the collected audio signal caused by the mixed environment and the recording equipment will reduce the accuracy of emotion recognition, and in order to improve the accuracy of emotion recognition and to enhance the ability of model generalization, before the extraction of emotional characteristics, the audio signal is pre-processed and the feature extraction is carried out afterwards.

Step 1. Tune the audio sample down to 16 kHz and quantify the number 16 bits. This not only reduces the noise disturbance but also retains the effective emotional information.

Step 2. The task of suppressing the low-frequency signal to raise the high-frequency signal is achieved by using the first-order digital filter in the high-pass filter to pre-emphasize the process of the audio ingress. The transfer functions and filter functions are shown in formula (1) and formula (2):

$$H(z) = 1 - a \times z^{-1} \tag{1}$$

$$Y(n) = X(n) - a \times X(n-1) \tag{2}$$

where α is the pre-weighting coefficient ($0.9 \le \alpha \le 1$). This paper sets α to 0.94. $X(n)$AA is the audio sampling value at the time n, and $Y(n)$ is the processed signal.

Step 3. In order to ensure the smoothness of the signal, the audio signal is divided by the window function, and the segment is the audio frame. In order to ensure a smooth transition of audio signals, there are a few overlaps between adjacent audio frames. In this paper, we adopt the standard in the AVEC2014 competition [21], in which the frame length is 3 seconds, and the frame shift is 1 second.

Audio signals are framed by the weighting of sliding windows and window functions. Commonly used sliding window functions are rectangular and Hamming windows, and the expressions of rectangular windows and Hamming windows are in formula (3) and formula (4):

$$\omega(n) = \begin{cases} 1, & 0 \le n \le N-1 \\ 0, & others \end{cases} \tag{3}$$

$$\omega(n) = \begin{cases} 0.54 - 0.46\cos\left[2\pi n/(N-1), 0 \le n \le N-1\right] \\ 0, \ others \end{cases} \tag{4}$$

In the expression of window function, n represents the n sampling point, and N is the number of sampling points in the audio frame. By comparing the two window functions, it is found that the main flap width of the Hamming window is wider than that of the rectangular window. As the main flap is wider and the slope of the beginning and the end of the audio signal is slower, the signal in the window will be more stable. Therefore, the Hamming window function is selected for the frame operation of the audio signal.

Step 4. To avoid the growth of training time due to the influence of silent fragments in the audio sample, the start and end points of the audio signal are judged using the two-gate endpoint detection method to strip the silent fragments of the audio sample. The short-term energy-based endpoint detection is used to determine the start and end points of the audio, and its expression are in (5):

$$E_n = \sum_{m=n-N+1}^{n} \left[x(m)\omega(n-m) \right]^2 = X^2(n) \times h(n) \tag{5}$$

Among them, $X(n)$ is the audio signal, $h(n)= \omega^2(n)$ and $\omega(n)$ are window functions.

The end point detection method based on zero crossing rate [22] is used to determine the dividing point between clear consonant and silent signal, and the expression is as follows:

$$Z_n = \frac{1}{2} \sum_{m=-\infty}^{\infty} \left| \text{sgn}\left[x(m) \right] - \text{sgn}\left[x(m-1) \right] \right| \omega(n-m) \tag{6}$$

where, sgn[ ] is the symbolic function.

Step 5. By extracting and analyzing the characteristic parameters in the pre-processed audio signal, the characteristics of the audio signal are obtained. In the feature extraction, the feature extraction method based on prior knowledge is used to extract features, and it is necessary to select acoustic emotional feature sets including prosody, spectrum and sound quality with the help of the prior knowledge from experts [23,24]. In this paper, the low-level features such as rhythm, spectrum and sound quality are extracted by using the characteristic extraction method based on prior knowledge, and then the full feature extraction is carried out with the help of Open SMILE.

A brief description of the characteristics is presented below:

Mel frequency cepstral coefficients (MFCC): Based on the short time Fourier transform (STFT), the logarithmic amplitude of the amplitude spectrum is firstly adopted, and then the fast Fourier transform is grouped and smoothed according to the Mel frequency scaling of the perceptual excitation.

Spectral centroid: The spectral centroid is the gravity center of STFT amplitude spectrum. $M_i[n]$ represents the amplitude of the Fourier transform at frequency n and frame i. The centroid is used to measure spectral shapes. The higher the centroid value, the brighter the texture and the higher the frequency. The spectral centroid calculation is shown in formula (7):

$$c_i = \frac{\sum_{i=0}^{n} n M_i[n]}{\sum_{i=0}^{n} M_i[n]} \tag{7}$$

Spectral flux: the spectral flux is defined as the square difference between the normalized amplitudes of continuous windows, and its calculation is shown in formula (8):

$$F_i = \sum_{n-1}^{n} \left(N_t[n] - N_{t-1}[n] \right)^2 \tag{8}$$

$N_t[n]$ and $N_{t-1}[n]$ are the normalization amplitude of the Fourier transformation at the current frame t and the previous frame $t-1$, and the spectral flux represents the amount of local spectral variation.

Beat histogram: This is a histogram, showing the relative intensity of different rhythm periods in a signal, and its calculation method is the root mean square autocorrelation [25].

Pause duration: The pause time is the percentage of time the speaker is muted in the audio segment.

After the feature extraction of the audio frame sequence, the audio emotion characteristics of different dimensions are obtained. Since the training of emotion model spends much time when the value range of different feature parameters is different, the feature normalization based on zero mean normalization is carried out, and the expression is as follows:

$$\overline{X} = \frac{\boldsymbol{X} - \boldsymbol{\mu}}{\sigma} \tag{9}$$

where $\boldsymbol{X}$ is the characteristic matrix of the sample, $\boldsymbol{\mu}$ is the mean vector, and σ is the mean square error.

4.2 Visual Signal Preprocessing and Feature Extraction

Considering that there are only seven basic categories of human emotions (anger, disgust, fear, happiness, sadness, surprise, and neutrality), after the research on AlexNet, the traditional AlexNet network structure is improved, and the improved AlexNet network shown in Fig. 2 is used to extract the features of visual signals.

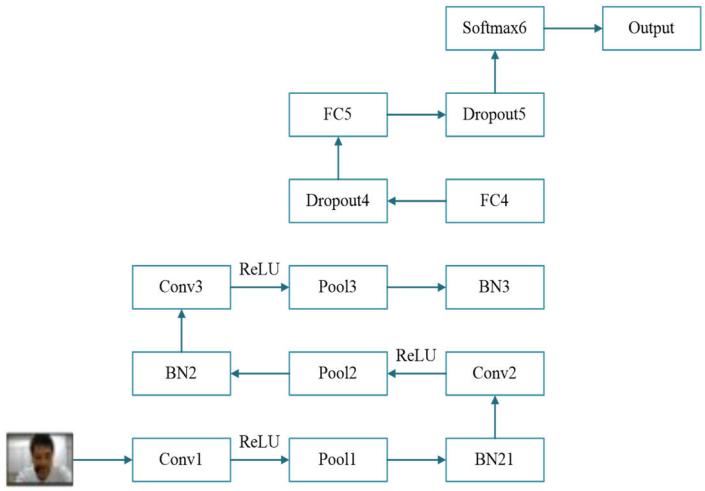


Fig. 2. Convolutional neural network structure for feature extraction of visual signals.

The improved AlexNet network consists of six layers, including three convolution layers, two full-connection layers, and one softmax layer. Each convolution layer is of the same size with 3×3 convolution

kernels. At the same time, a nonlinear excitation layer, a pooled layer, and a batch normalization layer are added behind each convolution layer, and the Dropout is added behind each full-connection layer to reduce overfitting. The layers in the CNN feature are as follows.

4.2.1 Convolution layer

Convolution layer is the core of CNN, and most of the heavy work is done in convolution layer. As shown in Section 2, the main features of convolution layer are local connection, sparse interaction and weight sharing. The convolution layer mainly depends on a set of filters. The output of the convolution layer can be regarded as a set of characteristic graphs extracted by the network. The output size is determined by the super parameters of convolution depth (*K*), convolution kernel size (*F*), convolution step size (*S*), and zero filling size (*P*). The convolution depth corresponds to the number of filters, each of which convolutes the inputs and learns the characteristic representation of the inputs. The convolution step size can be regarded as the unit interval of filter sliding. When the image is convoluted with a step size of 1, the filter will move one pixel position after each convolution. The larger the convolutional step, the smaller the output of the convolution, and the sparser the resulting features. In order to control the size of the feature plot during convolution, the edges of the input can usually be filled with zeros. $W_1 \times H_1 \times D_1$ assumes that the size of the input data is A and the size of the convolution output is $W_2 \times H_2 \times D_2$, which satisfies:

$$W_2 = \left(W_1 - F + 2P\right)/S + 1 \tag{10}$$

$$H_2 = \left(H_1 - F + 2P\right)/S + 1 \tag{11}$$

$$D_2 = K \tag{12}$$

Taking the first layer of CNN in Fig. 2 as an example, the input picture size is 224×224. Assuming that the convolutional core is 3×3, the convolution step size is 2, the size of the zero fill is 1, the convolution depth is 96, and the size of a single output feature plot is $(224 - 3 + 2 \times 1)/1 + 1 = 224$, the total output size should be 224×224×96. In addition, the example shows that the size of input data can remain unchanged by simple zero padding.

4.2.2 Pool layer

In the architecture of CNNs, the pooling operation is usually performed after the convolution. In addition to gradually reducing the spatial size of feature representation and decreasing the amount of parameters and calculations in the network, the pooling operation can also control the occurrence of overfitting to a certain extent. Pooling acts separately on the input depth slices, which aggregates the excitation convolutional layer features. The common methods include maximum pooling and average pooling, and the maximum pooling is selected in our research. The output of the pooling layer is determined by the size and the step size of the filter. Suppose the size of input data is $W_1 \times H_1 \times D_1$, the output size is $W_2 \times H_2 \times D_2$, which satisfies the following requirements:

$$W_2 = \left(W_1 - F + 2P\right)/S + 1 \tag{13}$$

$$H_2 = \left(H_1 - F + 2P\right)/S + 1 \tag{14}$$

$$D_2 = D_1 \tag{15}$$

The most common pooling layer has a selection of a 2×2 filter. In each depth slice of the input along with the width and the height of the Step 2 down-sampling operation, discarding 75% of the characteristic values, the input depth remains unchanged. Fig. 3 shows the output diagram of two different inputs in the action of this pooling layer. The left input size is 224×224×96, and the output is 112×112×96; the right input size is 4×4, and the output is 2×2.

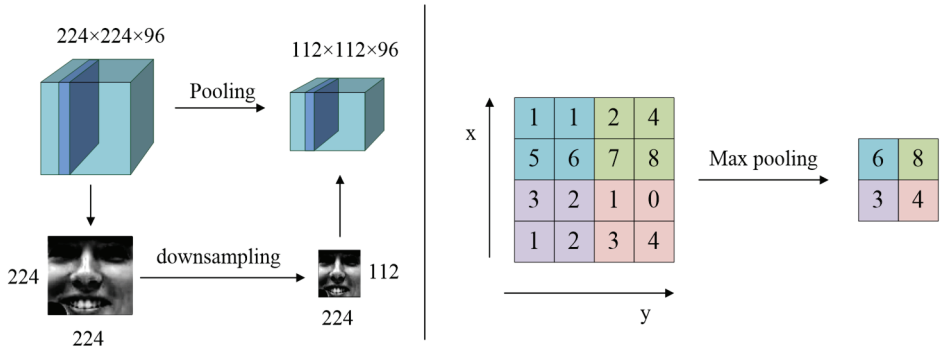


Fig. 3. Output diagram of pooling layer.

4.2.3 Batch normalization layer

The complexity of the deep network is that as training progresses, the parameters of the first few layers change, which affects the distribution of all subsequent network layer inputs. In order to mitigate this effect in actual training, it is often suggested to set a smaller learning rate and smaller initialization parameters, but this also greatly reduces the speed of training and increases the difficulty of obtaining saturated nonlinear models [22,26]. In this paper, the optimization method of batch normalization (BN) is used to improve the training of deep network.

The best method of data preprocessing is the whitening preprocessing, but its also has a large amount of calculations and cannot be differentiable everywhere. Therefore, different from the whitening preprocessing, the input features are selected to be separately preprocessed so that the mean value is 0 and the variance is 1. Suppose that a layer of d-dimensional input in the network is $X = (x^{(1)} \ldots x^{(d)})$ the normalization of each dimension of data is as formula (16):

$$\hat{x}^{(k)} = \frac{x^{(k)} - E\left[x^{(k)}\right]}{\sqrt{Var\left[x^{(k)}\right]}} \tag{16}$$

Considering that normalizing the input of each layer in the above formula will only change the feature distribution extracted from the previous layer to a certain extent, the transformation and the reconstruction are carried out on the basis of Eq. (16). The learnable parameters γ and β are introduced, and the expression is changed into:

$$y^{(k)} = \gamma^{(k)}\hat{x}^{(k)} + \beta^{(k)} \tag{17}$$

The advantage of this method is that when $\gamma^{(k)} = \sqrt{Var\left[x^{(k)}\right]}$ and $\beta^{(k)} = E\left[x^{(k)}\right]$, the original feature distribution of the network layer can be restored. The reason why it is called batch normalization is that

the data input during training is in the unit of batch, and the random gradient descent algorithm used in training is also calculated in the unit of batch. Assuming that the input of batch normalization layer is $B=\{X_{1\ldots m}\}$ and the output is $\{y_i=BN_{\gamma,\beta}(x_i)\}$, the forward propagation process can be expressed as Eqs. (18)–(20), where D is the number of samples in the current batch.

$$\mu_B \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \tag{18}$$

$$\sigma_B^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i-\mu_B)^2 \tag{19}$$

$$\hat{x}_l \leftarrow \frac{x_i-\mu_B}{\sqrt{\sigma_B^2+e}} \tag{20}$$

$$y_i \leftarrow \gamma\hat{x}_l+\beta \tag{21}$$

This method relaxes the requirements of training parameters and simplifies the training of deep network. At the same time, it can also be used as a regulator of network parameters to replace the dropout.

2.2.4 Full connection layer

The neurons in the full connection layer are fully connected to the output neurons in the previous layer, and their output can also be expressed as a form of weighted sum of inputs. The traditional full-connection layer uses a large number of parameters, generally accounting for 90% of the parameters required by the entire network; thus, its existence greatly reduces the efficiency of CNN training. Comparing the full connection layer with the convolution layer, the only difference between them is that the neurons in the convolution layer and the input are locally connected, and the parameters are shared between the neurons in the same layer. However, the calculation of the convolution layer is also the weighted sum of the input and the weight, which is essentially the same. Therefore, the full connection can be realized in the form of a convolution layer. For example, the convolution kernel with a size of F=7 can be set when the number of neurons is 4096, and the output size of the previous layer is 7×7×512, the fill parameter P=0 and the convolution step size S=1. The convolution depth K=4096. At this time, the output of the convolution layer is 1×1×4096, which is equivalent to the full connection layer. Compared with the two implementation methods, the number of parameters needed to be trained in the full connection layer is $7\times7\times512\times4096\approx10\text{M}$, while the number of parameters needed to be trained in the convolution layer is $7\times7\times4096\approx200\text{K}$ due to the sharing of parameters. The number of parameters to be trained in the latter is several orders of magnitude lower than that of the former, which can greatly improve the training of the whole network.

2.2.5 Softmax layer

When CNN is used in multi-classification tasks, softmax regression is often employed in the last layer of classification, which is called the softmax layer. As a matter of fact, softmax regression is an extension of logistic regression in multi-classification.

Logistic regression is a classical binary classification method in statistical learning, which is a supervised classification method. The training set consists of m labeled sample pair $\{(x^{(1)},y^{(1)}),\ldots,(x^{(m)},y^{(m)})\}$,

where the input feature $x^{(i)} \in \mathbb{R}^{n+1}$, and $x_0^{(i)} = 1$. The output sample belonging to the class is marked $y^{(i)} \in \{0,1\}$. The hypothesis function of logistic regression is:

$$h_\theta(x) = \frac{1}{1+\exp(-\theta^T x)} \tag{22}$$

The definition of loss function is shown in Eq. (23), where θ is a trainable parameter.

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)} log h_\theta\left(x^{(i)}\right) + \left(1-y^{(i)}\right) log h_\theta\left(1-x^{(i)}\right)\right] \tag{23}$$

As an extension of logical regression, the training set of softmax regression is also composed of *m* labeled samples to $\left\{\left(x^{(1)}, y^{(1)}\right),\ldots,\left(x^{(m)}, y^{(m)}\right)\right\}$, where the main difference is the sample label $y^{(i)} \in \{1,2,\ldots,k\}$, corresponding to the output of *k* different classifications. The corresponding hypothesis function satisfies Eq. (24).

$$h_\theta\left(x^{(i)}\right) = \begin{bmatrix} p\left(y^{(i)}=1 \middle| x^{(i)};\theta\right) \\ p\left(y^{(i)}=2 \middle| x^{(i)};\theta\right) \\ \vdots \\ p\left(y^{(i)}=k \middle| x^{(i)};\theta\right) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}} \begin{bmatrix} e^{\theta_1^T x^{(i)}} \\ e^{\theta_2^T x^{(i)}} \\ \vdots \\ e^{\theta_k^T x^{(i)}} \end{bmatrix} \tag{24}$$

where $\theta_1, \theta_2, \ldots, \theta_k \in \mathbb{R}^{n+1}$ is a trained parameter. The output of the function is a *k*-dimensional column vector, and each element in the vector corresponds to the estimated probability, where input *x* belongs to the current class, and the sum of all estimated probabilities is 1. In order to be consistent with the logistic regression, the same symbol θ is used to represent all parameters in the regression, and $\theta \in \mathbb{R}^{k\times(n+1)}$ is stacked by $\theta_1, \theta_2, \ldots, \theta_k$ according to the row, i.e.:

$$\theta = \begin{bmatrix} \theta_1^T \\ \theta_2^T \\ \vdots \\ \theta_K^T \end{bmatrix} \tag{25}$$

The loss function of softmax regression can be defined as:

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\left\{y^{(i)}=j\right\} log \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_l^T x^{(i)}}}\right] + \frac{\lambda}{2}\sum_{i=1}^{k}\sum_{j=0}^{n}\theta_{ij}^2 \tag{26}$$

where 1{ } is an indicator function. The function value is 1; otherwise 0, when the expression in parentheses is true. Different from the logistic regression, the weight attenuation term $\frac{\lambda}{2}\sum_{i=1}^{k}\sum_{j=0}^{n}\theta_{ij}^2$ is added after the loss function, where $\lambda > 0$. The purpose of adding this term is to solve the problem of parameter redundancy in softmax regression, making the cost function become a strict convex function, so as to ensure the unique optimal solution.

5. Emotion Feature Fusion based on Multimodal Attention Mechanism

5.1 Multimodal Attention Mechanism

Audio features are input into LSTM to learn the temporal context information through the full connection layer. Suppose that the audio characteristic at time t is x_t, the output of LSTM hidden layer at the previous moment is h_{t-1}, LSTM gating function is $f()$. Then the output of the hidden layer at time t is defined as：

$$h_t = f\left(h_{t-1}, x_t\right) \tag{27}$$

Suppose that the facial expression feature at time t is F, taking VGG19 as an example, the extracted feature is the output of the fifth convolution module, and the feature size is 196×512. Therefore, the feature dimension of F is 196 and the depth of each feature is 512. Then, the calculation process of each dimension feature's attention weight is as follows:

$$a_t = \text{soft}\max\left(\boldsymbol{w}_h^T \tanh\left(\boldsymbol{W}_\upsilon F + \boldsymbol{W}_g h_t\right)\right) \tag{28}$$

where, $\boldsymbol{w}_h^T$, $\boldsymbol{W}_\upsilon$ and $\boldsymbol{W}_g$ are the weight matrix of attention model input, video feature and hidden layer output, respectively. The facial expression features processed by the attention model are as follows:

$$c_t' = \sum_{i=1}^{k} a_{ti}\upsilon_{ti} \tag{29}$$

The audio context information and facial expression features in LSTM are integrated, and the calculation process is as follows:

$$c_t = \varsigma_t h_t + \left(1-\varsigma_t\right)c_t' \tag{30}$$

where, ς_t is the weight of audio feature at time t and $1-\varsigma_t$ is the weight of video feature at time t.

5.2 Modal Scaling Optimization Function

In the experiment, it is found that when the speaker is speaking, there is no face in camera. At this time, there exists only the environmental background, such as the computer, the desk, etc.; thus, the multimodal attention model cannot accurately judge whether there is a face in the video. It is only able to rely on the emotional label to propagate back the information according to dimensions in the feature, and to assign corresponding attention weights. Therefore, it is possible to assign larger contribution values to facial expression features when there is no face. Similarly, when a face appears in the camera, the speaker does not speak. It may be the audio of a remote video recorder. However, it is still possible that a large attention weight are assigned to the audio features, which may mislead the emotion recognition.

In the process of deep learning gradient descent, L2 regularization function [27] will be added to the original loss function to prevent overfitting. Its principle is to limit the influence of irrelevant parameters in the original loss function by adding auxiliary functions, and to guide the direction of reverse derivation of the total loss function. Therefore, inspired by L2 regularization, this paper uses the method of adding

auxiliary tag and auxiliary loss function to guide the gradient descent direction of the total loss function in the process of backpropagation. In extreme cases (with face but without audio, with voice but without face), the range of multimodal fusion ratio is limited.

For the audio mode, the short-term energy measures the strength of voice energy at a certain time. Due to the distance between the remote recorder and the speaker, the energy intensity varies greatly. By setting the energy threshold, the energy intensity below the threshold is set to 0, so as to avoid the interference of remote recorder. Therefore, the audio's short-term energy is extracted and normalized to $[0,1]$ as an audio auxiliary tag. For the video mode, the face image is detected by using the mature face detection database in Open CV. If a face is detected, the auxiliary tag is set to 1, and if not, it is set to 0. The auxiliary loss function is constructed as follows:

$$L_1 = \frac{1}{2}\left(\left(m-\varsigma_t\right)^2 + \left(n-\left(1-\varsigma_t\right)\right)^2\right) \tag{31}$$

where, m is the short-term energy and is a real number between $0 \sim 1$, and n is the set $\{0,1\}$, indicating whether the face is detected. The gradient of ς to L_1 at time t is:

$$\frac{\partial L_1}{\partial \varsigma_t} = 2\varsigma_t + n - m - 1 \tag{32}$$

When the short-term energy is 1, no face is detected, i.e., $n=0$, here $\frac{\partial L_1}{\partial \varsigma_t} = 2\varsigma_t - 2 < 0$, adjusting to the minimum; When a face is detected, i.e., $n=1$, the short-term energy m is 0, here $\frac{\partial L_1}{\partial \varsigma_t} = 2\varsigma_t > 0$, adjusting to the minimum; When there is both face and voice, $m=n=1$; When there is neither face nor voice, $m=n=0$, $\frac{\partial L_1}{\partial \varsigma_t} = 2\varsigma_t - 1$, close to 0. The effect on the main loss function is very low. Therefore, the total loss function is defined as follows:

$$L = \frac{1}{2T}\sum_{n=1}^{T}\left(y_t' - y_t\right)^2 + L_1 \tag{33}$$

6. Experiment and Discussion

6.1 Experimental Dataset

In order to verify the recognition effect of the model, this paper selects the database provided by AVEC2016 (International Audio/Visual Emission Challenge and Workshop) challenge to conduct experiments. AVEC2016 database is a subset of RECOLA (Remote Collaboration and Affective Interaction) database. The database provides natural data recorded from people who participate in a video conference. The database provides 27 videos of training sets, validation sets and test sets of 5 min in total, which were labeled by six French researchers on arousal and valence in two emotional dimensions. Every 40 ms in the range of $-1 \sim 1$ is marked. Among them, the horizontal axis validity represents the valence degree, indicating the positive and the negative degree of an emotion, and the vertical axis refers to the arousal degree, indicating the intensity and the depression of an emotion. The length of each video is 7,500 frames. Finally, six researchers label each frame, and the average value is taken. The official database emphasizes the workload of database construction and encourages researchers who use the

database to adopt more reasonable methods to extract features.

6.2 Experimental Setup

In the stage of video feature learning, 14 groups of videos are selected as the training set and 7 groups of videos are selected as the test set. For the facial expression feature learning, VGG19, traditional AlexNet, improved AlexNet and InceptionV3 were used for comparative experiments. As the above four kinds of depth CNN require a size of 224×224 or 299×299, the number of batch training is set to 150. In the frequency attention mechanism model, the CNN structure of three-layer convolution and three-layer pooling is used as the audio context feature learning model. The size of audio spectrum's each frame is 24×120, and the input window is 5 frames of audio information; thus, the size of input spectrum is 120×120. The size of convolution kernel in the first layer is 2×2, and the number of convolution kernels is 4. The size of convolution kernel in the second layer is 3×3 and the number of convolution kernels is 16. The size of convolution kernel in the third layer is 3×3 and the number of convolution kernels is 32. In order to ensure the consistency between the output sizes of whole frequency learning and frequency local learning feature map, the full 0 filling is used in the convolution process. The pooling size is set to the maximum pooling size of 3×3 with step size of 1.

In the cases of multiple previous experiments, 1000 epochs were set in order to ensure adequate training. The gradient descent optimization algorithm is selected from the stochastic gradient descent (SGD) and Adam. The initial learning rate is set to 0.0005. In order to more intuitively compare the difference between training and test results, each epoch is trained and tested once on the corresponding data set. The experimental hardware environment is PC, and the operating system is Ubuntu14.04.5. Linux kernel is installed and GPU is used to accelerate deep learning. The main software and version information used in the experiment are shown in Table 1.

Table 1. Experimental software information

Number	Software name	Version
1	Keras	2.1.2
2	TensorFlow	1.2.1
3	Caffe	2.0.1
4	Python	2.7.13
5	Numpy	1.12.1
6	Scipy	0.19.0
7	Matlab	2012a

6.3 Evaluation Index

In the stage of video feature learning, this paper uses R^2 coefficient as the evaluation index of feature learning, which represents the fitting degree of prediction value and label value in the regression task by calculating the change of data. The larger the R^2 coefficient is, the higher the fitting degree is and the better the feature extraction effect is. The R^2 coefficient function is as follows:

$$R^2 = 1 - \frac{\sum\left(Y_{\_actual} - Y_{\_predict}\right)^2}{\sum\left(Y_{\_actual} - Y_{\_mean}\right)^2} \tag{34}$$

where, Y_{actual} is an emotional reality label sequence, $Y\_{predict}$ is a sequence of emotional predictors, and Y_{mean} is the average value of emotional reality label sequence.

In the training and testing stage, this paper uses the concordance correlation coefficient (CCC) provided by the official database as the evaluation index of emotion recognition. The calculation formula is as follows:

$$\mathrm{CCC}(x,y)=\frac{2\rho\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2+(\mu_x-\mu_y)^2} \tag{35}$$

where, μ_x and μ_y are the average values of emotional predictive value sequence and emotional reality label sequence, respectively; σ_x and σ_y are the standard deviations of emotional predictive value sequence and emotional reality label sequence, respectively; ρ is the Pearson correlation coefficient between the two sequences. The formula is as follows:

$$\rho=\frac{\frac{1}{n}\sum_i^n(x_i-\bar{x})(y_i-\bar{y})}{\sigma_x\sigma_y}=\frac{C_{\mathrm{CCC}}}{\sigma_x\sigma_y} \tag{36}$$

In the whole experiment, root mean square error (RMSE) is used as the loss function. It is defined as follows:

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_i^n(x_i-y_i)^2} \tag{37}$$

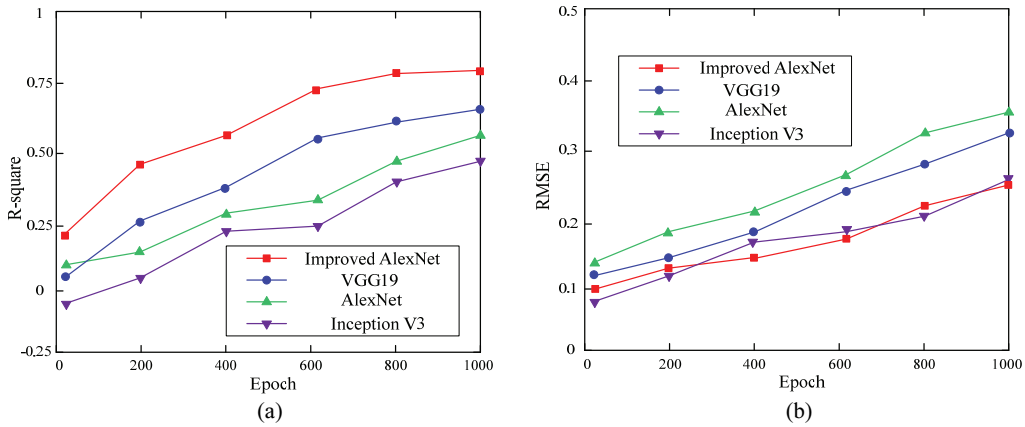
where, x_i represents the emotional prediction value of the *i*-th frame, and y_i represents the emotional reality label of the *i*-th frame.

6.4 Training Results and Analysis

In the stage of facial expression feature learning, four kinds of deep convolution neural networks were used in the two dimensions of arousal and valence respectively. The results of feature learning are shown in Fig. 4 and Fig. 5, respectively.

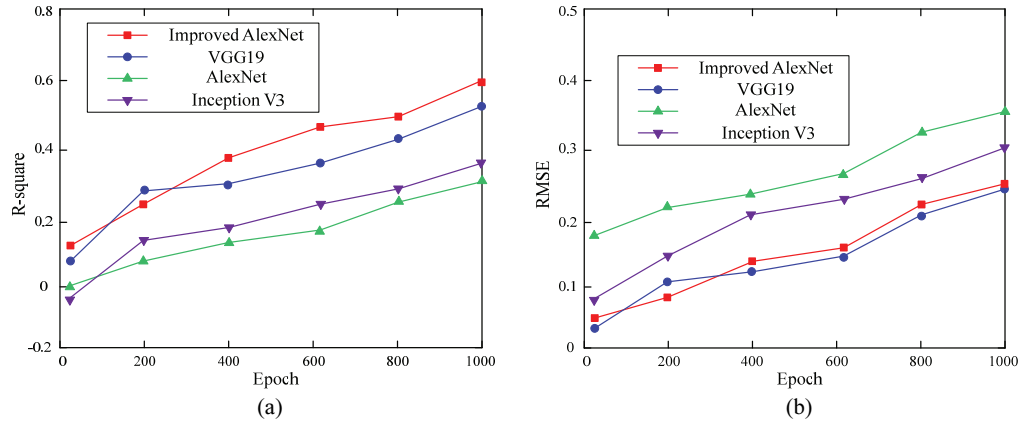
As shown in Fig. 4, in the arousal dimension, the R^2 coefficient of the improved AlexNet is very close to 0.75, which is higher than several contrastive neural networks. In terms of loss, InceptionV3's performance is almost equal to that of the improved AlexNet. However, in terms of training time, InceptionV3 takes longer time than the improved AlexNet, and the single epoch training time is as long as 48 seconds, while that of the improved AlexNet is only 45 seconds. This is because the sizes of convolution kernels adopted by the improved AlexNet are all 3×3, rather than larger convolution kernels such as 5×5, 7×7, etc. The purpose of choosing smaller convolution kernels is to minimize the number of parameters without affecting the size of receptive field. The test results show that the improved AlexNet is the right choice to learn facial expression features in the arousal dimension.

As shown in Fig. 5, in the valence dimension, the loss of the improved AlexNet is very close to the lowest of VGG19, and its R^2 coefficient can be as high as 0.61, which is much higher than that of the traditional AlexNet and InceptionV3. Therefore, it is the right choice to use the improved AlexNet to learn facial expression features in the valence dimension.



(a) (b)

Fig. 4. Learning results of affective salient features in arousal dimension video: (a) comparison of R^2 coefficient test results and (b) comparison of loss function test results.



(a) (b)

Fig. 5. Learning results of affective salient features in valence dimension video: (a) comparison of R^2 coefficient test results and (b) comparison of loss function test results.

6.5 Comparison of Dimension Emotion Recognition Results between Multimodal Attention Mechanism and Other Methods

In order to further verify the performance of proposed multimodal emotion classification model, it is compared with methods in other references based on AVEC2016 dataset. The comparison is shown in Fig. 6.

As shown in Fig. 6, compared with other methods, the multimodal attention mechanism is inferior to the method in [19] in terms of loss, but CCC can better reflect the fitting degree of emotion prediction value and emotional label value. The CCC of the multimodal attention mechanism is the highest. Similarly, in the valence dimension, compared with other methods, although the loss of multimodal attention mechanism is inferior to that of [18], the CCC is still higher than that of other comparison methods. As shown in Fig. 6, the multimodal attention mechanism has outperformed most methods in CCC. CCC has reached 0.729 and 0.718 in arousal and valence dimensions, respectively. The experimental results show that the multimodal attention mechanism can effectively extract emotional salient features from audio and video for fusion.

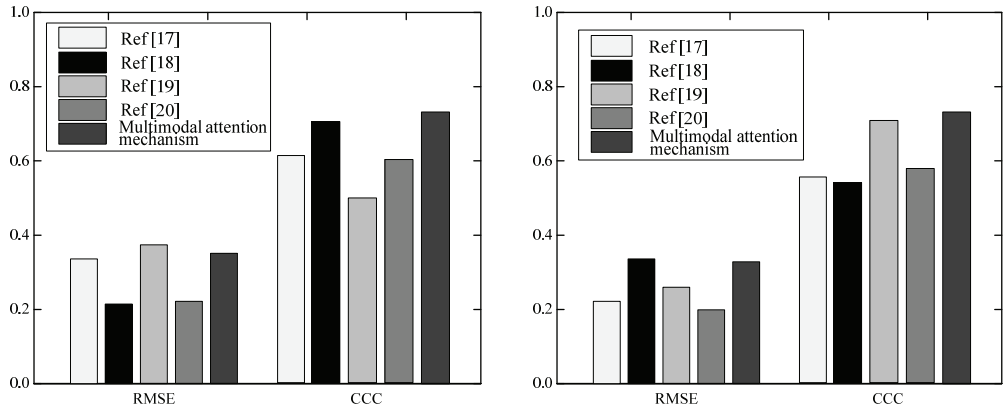


Fig. 6. Learning results of different methods in the arousal dimension (a) and the valence dimension (b).

7. Conclusion

Based on the continuous dimension emotion recognition, an improved AlexNet network combined with the attention mechanism is proposed for audio and video bimodal emotion recognition. Experimental results show that, compared with the current mainstream methods, our proposed method uses the attention mechanism to selectively enhance the features learned by using context information. It simplifies the process of feature preprocessing and reduces the interference of emotion independent factors. It has produced expected results in the task of dimension emotion recognition in continuous video and audio modes.

In the proposed method, the deep CNN is used to learn facial expression features. It is not integrated with the attention mechanism, which will lead to the loss of model optimization and incomplete feature learning. Therefore, the next step will be to use a more reasonable network structure to learn facial expression features, combining feature learning with model prediction. In addition, audio manual features will be introduced to enrich audio information, so as to further improve the model's recognition accuracy.

Acknowledgement

This work was supported by the Research Foundation of Education Bureau of Hunan Province (No.18B564), Science and Technology Plans of Development and Reform Commission of Hunan Province, China (No. 2013-1199).

References

[1] R. Jamwal, J. Enticott, L. Farnworth, D. Winkler, and L. Callaway, "The use of electronic assistive technology for social networking by people with disability living in shared supported accommodation," *Disability and Rehabilitation: Assistive Technology*, vol. 15, no. 1, pp. 101-108, 2020.

[2] S. Sharma, G. Singh, and A. S. Aiyub, "Use of social networking sites by SMEs to engage with their customers: a developing country perspective," *Journal of Internet Commerce*, vol. 19, no. 1, pp. 62-81, 2020.

[3] D. Tiwari and N. Singh, "Ensemble approach for twitter sentiment analysis," *International Journal of Information Technology and Computer Science*, vol. 11, no. 8, pp. 20-26, 2019.

[4] R. Bhargava, S. Arora, and Y. Sharma, "Neural network-based architecture for sentiment analysis in Indian languages," *Journal of Intelligent Systems*, vol. 28, no. 3, pp. 361-375, 2019.

[5] J. McDonald, A. C. M. Moskal, A. Goodchild, S. Stein, and S. Terry, "Advancing text-analysis to tap into the student voice: a proof-of-concept study," *Assessment & Evaluation in Higher Education*, vol. 45, no. 1, pp. 154-164, 2020.

[6] F. R. Sullivan and P. K. Keith, "Exploring the potential of natural language processing to support microgenetic analysis of collaborative learning discussions," *British Journal of Educational Technology*, vol. 50, no. 6, pp. 3047-3063, 2019.

[7] A. L. Afzal and S. Asharaf, "Deep multiple multilayer kernel learning in core vector machines," *Expert Systems with Applications*, vol. 96, pp. 149-156, 2018.

[8] G. Manogaran, R. Varatharajan, and M. K. Priyan, "Hybrid recommendation system for heart disease diagnosis based on multiple kernel learning with adaptive neuro-fuzzy inference system," *Multimedia Tools and Applications*, vol. 77, no. 4, pp. 4379-4399, 2018.

[9] Z. Dong and B. Lin, "BMF-CNN: an object detection method based on multi-scale feature fusion in VHR remote sensing images," *Remote Sensing Letters*, vol. 11, no. 3, pp. 215-224, 2020.

[10] A. Moussavi-Khalkhali and M. Jamshidi, "Feature fusion models for deep autoencoders: application to traffic flow prediction," *Applied Artificial Intelligence*, vol. 33, no. 13, pp. 1179-1198, 2019.

[11] M. Wollmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, and L. P. Morency, "Youtube movie reviews: sentiment analysis in an audio-visual context," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 46-53, 2013.

[12] S. Zhang, X. Wang, G. Zhang, and X. Zhao, "Multimodal emotion recognition integrating affective speech with facial expression," *WSEAS Transactions on Signal Processing*, vol. 10, pp. 526-537, 2014.

[13] S. Dobrisek, R. Gajsek, F. Mihelic, N. Pavesic, and V. Struc, "Towards efficient multi-modal emotion recognition," *International Journal of Advanced Robotic Systems*, vol. 10, article no. 53, 2013. https://doi.org/10.5772/54002

[14] Y. Wang, "Multimodal emotion recognition algorithm based on edge network emotion element compensation and data fusion," *Personal and Ubiquitous Computing*, vol. 23, no. 3, pp. 383-392, 2019.

[15] S. Sahoo and A. Routray, "Emotion recognition from audio-visual data using rule based decision level fusion," in *Proceedings of 2016 IEEE Students' Technology Symposium (TechSym)*, Kharagpur, India, 2016, pp. 7-12.

[16] M. Glodek, S. Reuter, M. Schels, K. Dietmayer, and F. Schwenker, "Kalman filter based classifier fusion for affective state recognition," in *Multiple Classifier Systems*. Heidelberg, Germany: Springer, 2013, pp. 85-94.

[17] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301-1309, 2017.

[18] J. Huang, Y. Li, J. Tao, and J. Yi, "Multimodal emotion recognition with transfer learning of deep neural network," *ZTE Communications*, vol. 15, no. S2, pp. 23-29, 2017.

[19] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio-visual emotion recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 3030-3043, 2018.

[20] W. Li, M. Chu, and J. Qiao, "Design of a hierarchy modular neural network and its application in multimodal emotion recognition," *Soft Computing*, vol. 23, no. 22, pp. 11817-11828, 2019.

[21] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "AVEC 2014: 3D dimensional affect and depression recognition challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, Orlando, FL, 2014, pp. 3-10.

[22] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 597-607, 2012.

[23] S. Poria, E. Cambria, N. Howard, G. B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, pp. 50-59, 2016.

[24] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proceedings of the 18th ACM International Conference on Multimedia*, Firenze, Italy, 2010, pp. 251-260.

[25] J. Ma, Y. Sun, and X. Zhang, "Multimodal emotion recognition for the fusion of speech and EEG signals," *Xi'an Dianzi Keji Daxue Xuebao/Journal of Xidian University*, vol. 46, no. 1, pp. 143-150, 2019.

[26] B. Knyazev, R. Shvetsov, N. Efremova, and A. Kuharenko, "Leveraging large face recognition data for emotion classification," in *Proceedings of 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG2018)*, Xi'an, China, 2018, pp. 692-696.

[27] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1-22, 2010.

Min Liu **https://orcid.org/0000-0001-5100-742X**

She has got Master of Engineering, she is an associate professor. She graduated from the Kunming University of Science and Technology in 2004. She entered the work of Hunan Vocational College of Science and Technology in 2004. In 2014, she graduated from Hunan University with a master's degree in computer science and technology. Her research interests include data mining, intelligence algorithm, etc.

Jun Tang **https://orcid.org/0000-0002-1504-0326**

He has got Master of software engineering, he is a senior engineer. He graduated from Hunan University with a master's degree in 2013. He is working a senior engineer in Hunan Vocational College of Science and Technology. His research interests include software architecture, intelligence algorithms, and application base of big data.