IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 39, No. 5, October, 2022, pp. 1475-1481
Journal homepage: http://iieta.org/journals/ts

Emotion Recognition of College Students Based on Audio and Video Image

Chenjie Zhu[1], Ting Ding[2], Xue Min[3*]

[1] School of Tourism and Culinary Arts, Zhejiang Business College, Hangzhou 310053, China
[2] School of Accounting and Finance, Zhejiang Business College, Hangzhou 310053, China
[3] School of E-commerce, Zhejiang Business College, Hangzhou 310053, China

Corresponding Author Email: 00861@zjbc.edu.cn

https://doi.org/10.18280/ts.390503

Received: 29 May 2022
Accepted: 18 September 2022

Keywords:
audio emotion recognition, video image emotion recognition, deep learning, decision-making layer feature fusion, multimodal emotion recognition

ABSTRACT

Emotional problems are common among contemporary college students. To improve their mental health, it is urgent to quickly identify college students' negative emotions, and guide them to improve their emotional development. Students' emotions are expressed through multiple modalities, such as audio, facial expressions, and gestures. Using the complementarity between multi-modal emotional information can improve the accuracy of emotion recognition. This paper proposes a multi-modal emotion recognition method for voice and video images based on deep learning: (1) For voice modal recognition, the voice is firstly preprocessed to extract voice emotional features, and then the attention-based long-short-term memory network (LSTM) is adopted for emotion recognition; (2) For video image modal recognition, the extended local binary pattern (LBP) operator is used to calculate the image features, LBP block weighting and multi-scale partitioning are combined for feature extraction, principal component analysis (PCA) is adopted to reduce the dimensionality of eigenvectors, and the VGG-16 network model is constructed with the transfer learning training strategy to realize emotion recognition. (3) The voice and video image emotions recognized by single-modal recognitions are weighed and subjected to feature fusion at the decision-making layer, and used to classify multi-modal emotions. Experimental results show that on the test set of the cheavd2.0 Chinese emotion database, the recognition accuracy of our multi-modal fusion recognition algorithm is better than the single-modal recognition methods.

1. INTRODUCTION

With the prolonged and recurrent epidemics, rising international political risks, and numerous internal and external shocks like the unpredictability of the economic environment, social competition has grown more intense in recent years, which has put a lot of psychological pressure on today's college students. Emotional and negative emotional issues among college students are more and more prevalent on campus. Negative emotions have a significant impact on students' lives and careers, as they lead to depression, anxiety, or other mental diseases or traumas. According to statistics, more than 60% of college students who experience psychological difficulties or hurdles face emotional issues and contradictions. It is clear that emotional issues are now a fairly widespread and important issue among today's college students. To address this issue and enhance college students' mental health, it is crucial to figure out how to encourage the growth of pleasant emotions in today's college students.

Behavior and mental health are impacted by emotion. According to physiological psychology, emotion is a complex condition of an organism that manifests itself in audio, voice, expression, body language, and other ways [1]. Currently, the following techniques are most commonly used in mental health counseling offices for college students to identify emotions: (1) Emotional and psychological assessments are conducted using various psychological test scales or psychological assessments in conjunction with scoring standards; (2) Detecting students' physiological signals such as breathing , heart rate, body temperature, EEG, etc.; (3) Detecting emotional behavior such as facial expression recognition, voice emotion recognition and gesture recognition [2].

The respondent may conceal the true circumstances and provide false responses using the first technique, which will have an impact on the recognition results. The data collection for the second method is challenging and requires the use of specialized medical equipment. In real-world situations, it cannot be broadly applied. The third technique requires some effort to gather audio and facial expressions. It has a high recognition accuracy and can evaluate and model audio and video data gathered by the school's access control facial recognition system and psychological census equipment.

College campus human-computer interface areas including online education, mental health diagnosis, and safety monitoring can all benefit from automatic emotion recognition for students. Facial expression and audio data are used to analyze students' emotional states, determine their current learning level, and then deliver more individualized and scholarly educational advice to both teachers and pupils. Likewise, automatic emotion recognition technology can assist in the diagnosis of mental health and identify students' emotions to promote healthy development by analyzing the voice and expression data collected by the university access control system and psychological survey equipment from the students [3].

A facial emotion identification technique for college students was proposed by Liu et al. [4] in the area of security surveillance. For the purpose of achieving intelligent surveillance of the region where college students are located, this method makes use of a number of image preprocessing technologies, including moving target identification, target classification, and target tracking. When a potentially harmful target is discovered, it can automatically sound an alarm.

Numerous neural network classifiers based on deep learning have been demonstrated to be efficient in improving the emotion identification rate and classification efficiency in the field of automatic emotion recognition. Many academics have conducted studies on the use of deep learning technologies to recognize emotions in college students. For instance, Chen et al. [5] developed an interactive convolutional neural network (CNN)-based emotion identification and psychological intervention system for college students and experimentally tested its efficacy. The findings demonstrate that deep learning CNN outperforms decision tree (DT) and backpropagation neural network (BPNN) algorithms at detecting the emotions of students.

Li and Zhou [6] proposed a psychological emotion recognition algorithm based on multi-source data, using a one-dimensional convolutional neural network (1D-CNN) to mine online patterns of students from online behavior sequences. It provides a systematic methodological and theoretical tool for identifying problematic students and provide interventions. Liu et al. [7] put forward a self-adaptive emotion identification model for college students using upgraded multi-feature deep neural network technology. The model employs a deep neural network (DNN) to classify the emotional EEG data that has been acquired, and it then determines the emotions of college students based on the classification results. Deep neural network-based automatic emotional identification of students and its application in the detection of depression were proposed by Ding et al. [8]. A DNN network model framework based on contextual emotional information was also developed, and automatic auxiliary emotional classification was achieved.

According to psychologist Mehrabian, just 7% of information sent by people in daily conversation is communicated through text, 38% is communicated through audio, and 55% is communicated through facial expressions. It is evident that, audio and facial expressions are crucial in the study of affective computing. Although the voice modality and the facial expression modality have different emotional expression modes, both modalities can simultaneously convey an individual's emotional state. Generally speaking, changes in audio are synchronized with changes in facial expression [9].

Some researchers have discovered through experimental research that the emotion recognition algorithm based on the fusion of bimodal features (audio and facial expressions) performs better than single-modal emotion recognition because bimodal emotional features include both audio and facial expressions. They create complementing advantages when compared to single mode features [10].

In this paper, the cheavd2.0 Chinese dataset established by the Institute of Automation, Chinese Academy of Sciences is adopted. Considering the synchronization of the two modalities of facial expression and voice, the data of these two modalities is extracted from the video. The collected data has the advantages of obvious features and high precision. By collecting these two modal signals and extracting features, a multi-modal emotion recognition classification model is constructed based on deep learning. The robustness and real-time performance of the model is further improved by combining the fusion of decision-making layer and transfer learning.

2. FEATURE EXTRACTION

2.1 Extracting audio emotion features

Audio is crucial in the dialogue and communication process because it allows for the expression of emotion. The majority of information regarding the speaker's emotional state can be found in the tone quality, prosody, and spectral aspects of the audio signal. Audio emotion feature extraction includes three steps: preprocessing, feature extraction, and standardization. It primarily targets time-series audio signal input to get pertinent features.

(1) Preprocessing of voice data

First, separate the audio data from the video. The process for achieving video vocal filtering uses Python to invoke the moviepy package of video editing tools along with ffmpeg and spleeter. The spleeter software simply needs to select the audio stream for the separated audio stream because the ffmpeg software can extract the audio stream and video stream in the audio and video files. The human voice can be distinguished from the background noise if the track separation method is the two-stems model.

Second, carry out audio framing and windowing. Video data is not the same as voice data. The segmented audio must be organized into the frame operation data structure for programming in order to be sent and stored. The frame length of this paper is 25ms, and the frame shift is 10ms since research has demonstrated that the audio signal is microscopically stable and exhibits short-term stationarity [11] (the audio signal can be considered roughly unchanged within 10-30ms). Framing, however, will cause the issue of spectrum leakage. This study employs the Hamming window function to frame the solution to this issue. The new audio signal can be expressed as:

$$S(n) = s(n) * w(n) \tag{1}$$

where, $w(n)$ is the Hamming window function; $s(n)$ is the original audio signal; $s(n)$ is the new windowed audio signal after framing. The Hamming window function can be expressed as:

$$w(n) = \begin{cases} 0.54 - 0.46\cos[2\pi n/(N-1)], & 0 \le n \le N-1 \\ 0, & n = else \end{cases} \tag{2}$$

2.2 Extracting audio features

Prosodic features, sound quality features, and spectral features are examples of traditional features used in audio emotion recognition. Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Cepstral Coefficients Coefficient (LPCC), etc. are often used spectral feature parameters [12]. Prior to deep learning, which is constrained by algorithms, MFCC had a high degree of discrimination, making it the standard method for automatic voice recognition. The Fbank feature, however, has been demonstrated to be in line with the nature of the sound signal and fit the reception characteristics of the human ear with the gradual development of deep learning and neural networks. On the basis of Fbank, MFCC

is essentially applying a discrete cosine transform (DCT). The audio signal will lose some of its original, highly nonlinear components due to the linear nature of DCT. After actual verification, it is clear that MFCC's performance in the neural network falls short of Fbank's [13]. DNN/CNN, for instance, can minimize loss by better utilizing the correlation of Fbank features. In this paper, three different types of audio features are extracted, including audio MFCC features, MFCC first-order difference, and MFCC second-order difference; audio Fbank features, Fbank first difference, and Fbank second difference; and spectrogram features, in order to compensate for the shortcomings of MFCC in neural network classification and improve the effectiveness of network classification. Figure 1 below depicts the unique feature extraction procedure.

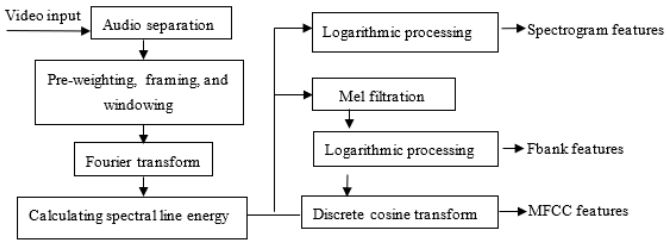

Figure 1. Flow of audio feature extraction

2.3 Normalizing audio data

Data normalization reduces the discrepancy between the test set and the training set throughout the deep learning process. The standardization operation is useful to the stability of the value in the process of network optimization in deep learning if the distribution difference between the training set and the test set is relatively big and the error of the test set may be lower than the error of the training set. Thus, after the training set and test set have been divided, the voice data standardization process in this paper performs the data standardization operation of removing the mean value on the training set and the test set, respectively, in order to equalize the frequency spectrum and improve the signal-to-noise-ratio (SNR). The approach is to subtract the mean value from the data, and then divide with variance:

$$X=X-\text{mean\_value} \tag{3}$$

$$X=X/\text{std\_value} \tag{4}$$

The normalization of audio data makes the neural network easier and faster to converge, and helps to improve the class recognition rate.

2.4 Emotion recognition

After the frame division operation, there is a specific association between each frame because audio has temporal features. The long short-term memory network (LSTM) can recognize audio emotions because to the frame-level features of this connection. Nevertheless, Qiu et al. [14] showed that the addition of the attention mechanism in machine learning can, on the one hand, reduce the amount of high-dimensional input feature calculation required and, on the other hand, can direct the network model's search for data information to input features that are significantly associated with the current output, enhancing the output quality. The loss of information at the top of the network and degradation problems were successfully resolved by the attention-based dense LSTM audio emotion recognition method proposed by Xie et al. [15]. This method uses the attention mechanism to add weight coefficients across links between multi-layer LSTM networks. For the purpose of identifying audio emotions, Li et al. [16] employs a bidirectional long short-term memory (BLSTM-DSA) network with directional self-attention. To appropriately identify audio frames in the temporal network that contain emotional information, this system may automatically label the weight of audio frames. Experimental results suggest how the algorithm effectively raises the recognition accuracy for audio emotions.

The typical LSTM uses a traditional encoder-decoder structure, and no matter how long the input data sequence is, it is encoded into a fixed-length vector representation. Although the LSTM's memory function can store long-term states, it is unable to effectively handle large multidimensional and multivariable datasets in practice, and the model may fail to take into account some crucial timing information during training. The model's performance declines, which has an impact on the precision of the predictions. This study introduces the attention mechanism on the foundation of LSTM with a focus on the flaws of LSTM itself. The goal is to maintain the intermediate state of the LSTM encoder, train the model to selective learning on these intermediate states, and overcome the restriction of the conventional encoder-decoder utilizing fixed-length vectors in the encoding process. In order to train the network, the previously extracted voice feature is first used as the input. Next, the attention mechanism is merged with LSTM, and the output of each LSTM node is recorded as the input of attention. The conventional LSTM network employs the self-attention technique. The forget gate and input gate are changed into attention gates, the features of each frame are combined by a weight, and the attention is weighted on the cell state at various intervals to construct a depth attention gate, enhancing algorithm performance. Figure 2 below depicts the organization of the enhanced LSTM network model based on the attention mechanism (Attention-LSTM model):

In Figure 2, $x^{<t>}$ is the input series; $a^{<t>}$ is the features learned for the input series; $w^{<k,t>}$ is the attention weight of each feature, i.e., the effect of the t-th feature on result $y^{<k>}$, or the attention that should be paid by y to $a^{<t>}$at time t; c is the memory unit; $y^{<k>}$ is the classification result of the output emotion. The attention weight $w^{<k,t>}$ can be calculated by:

$$w^{<k,t>}=\frac{\exp(e^{<k,t>})}{\sum_{t=1}^{T}\exp(e^{<k,t>})} \tag{5}$$

In Figure 2, the first LSTM network is used to process the input sequence in order to achieve high-level feature learning. The output vector from the LSTM layer is then used as the input of the Attention layer, and the memory unit is then solved by judiciously allocating attention weights. The second LSTM network realizes the specific classification output of audio emotion. When the LSTM power prediction model and the Attention mechanism work together, the model's training efficiency is increased and it can determine the significance of information at each point in the input process.

Without increasing the calculation and storage cost of the model, the attention mechanism assigns varied weights to the input features of LSTM, highlights the major influencing aspects, and aids LSTM in rendering accurate judgments.

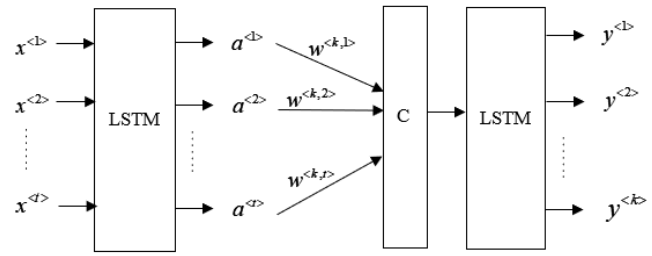

Figure 2. Structure of Attention-LSTM

3. MODAL RECOGNITION

Face image acquisition, expression feature extraction, and emotion classification are the three basic phases in facial emotion recognition using video images. Figure 3 shows the flow of our algorithm.

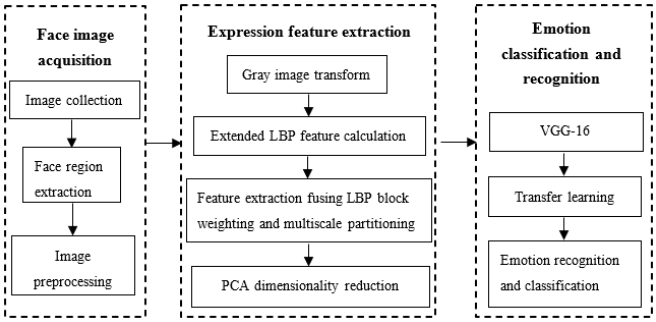

Figure 3. Flow of video image modal recognition

3.1 Extraction of facial expression features

Active appearance model (AAM), Gabor wavelet transform (GWT), local binary pattern (LBP) are the key feature extraction techniques for facial expressions based on changes in facial texture or geometric deformation features. AAM is a mixed-feature feature extraction technique. To provide a parametric description of the face, it primarily incorporates shape and texture data. Multi-scale and multi-directional texture alterations can be found using the Gabor wavelet transform approach. An operator called Local Binary Pattern (LBP) is used to characterize the local aspects of images. Significant benefits of LBP characteristics include rotation invariance and grayscale invariance. The benefit of these algorithms over the first two is that they are more resistant to illumination changes. The calculation is straightforward and practical for real-time analysis. In order to compare the classification accuracy of local binary patterns (LBP) and histograms of oriented gradients for facial emotion recognition, Subudhiray et al. [17] compared the histogram of oriented gradients and local binary pattern coefficients (HOG). The comparison demonstrates that LBP performs better than HOG. Convolutional neural networks and local binary patterns were combined to create Zhu et al. [18] proposed a face emotion recognition technique, and coupled CNN with an optimized central local binary pattern (CLBP) algorithm. In this way, they built a CNN-CLBP system to detect facial expressions of emotion.

The research mentioned above demonstrate how well LBP and neural networks work together to categorize and recognize expressions. However, because the original LBP feature employs gray values in a fixed neighborhood, the encoding of the LBP feature will be incorrect when the image's scale changes. The feature will not accurately reflect the pixel's neighboring texture information. Frequency texture features of various sizes will manifest during facial expression recognition. This paper improves the original LBP operator into the extended LBP operator [19] for feature calculation: extend the 3x3 neighborhood to any neighborhood, and replace the square neighborhood of the original LBP operator with a circular neighborhood. This allows the algorithm to adapt to texture features of different scales and meet the requirements of grayscale and rotation invariance. Using each pixel as the center, this paper determines how the surrounding pixels' gray values relate to it and execute a binary transformation to get the LBP coded version of the entire image; Then, the 64 sub-regions of the LBP coded image are divided into individual regions. Using the multi-scale partition histogram statistical feature extraction approach, the histograms of various size partitions are combined, and the weighted histogram of the partitions is merged into the entire image, producing eigenvectors of the entire image. Then, the principal component analysis (PCA) [20] is employed to reduce the dimensionality of the eigenvectors. PCA aims to reduce the size of source data while retaining the most relevant information, and finally inputs these eigenvectors into neural networks for classification and recognition. Thus, the steps of the statistical feature extraction approach for multi-scale partition histograms are as follows:

(1) Divide the input image into 64 sub-regions, and use a fixed value weighting method to give different weights to different regions.

(2) After extracting the LBP gray histogram by block, weight the histogram of each sub-block according to the weight set in step (1).

(3) Concatenate the histogram features of all sub-blocks to generate the feature vector of the whole image.

(4) Partition the entire image at different scales, gradually expanding from small-scale partitions to large-scale partitions.

(5) Concatenate the gray histogram features of each level to form the feature vector of the entire image.

(6) The eigenvectors generated by cascading steps (3) and (5) are used as the eigenvectors to be identified in the next step after PCA dimensionality reduction.

The specific algorithm flow is shown in Figure 4 below:

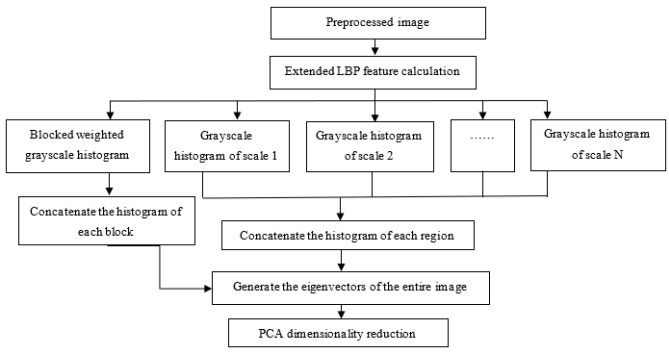

Figure 4. Flow of facial feature extraction

3.2 Improved VGG-16

At this stage, a classifier must be designed or chosen in order to train the input training samples and defined classes, retrieve the classifier's parameter information, and then

classify the test samples. The classifier based on the convolutional neural network model (CNN) in the deep learning process has a high accuracy in recognizing facial expressions. If the depth of the neural network model is increased, the classification effect of the system can be improved, presuming that the dataset is sizable enough [21]. But it might also take longer and be more prone to over-fitting, which would lead to subpar network performance.

This research chooses the VGG-16 model of the CNN network as the core model to address this issue and builds a better VGG-16 model based on the transfer learning strategy, primarily using the transfer machine learning approach. [22] In this way, the authors pre-trained the VGG-16 neural network (the model structure is displayed in Figure 5 below), and then loaded the pre-trained network parameters to initialize the network signal, producing a better network model. To avoid poorer classification accuracy brought on by the dearth of video expression datasets, the transfer learning of "pre-training model + fine-tuning mechanism" is employed to pre-train the network during model parameter training. The fer2013 image data is used initially. In the target dataset, the cheavd2.0 video dataset [23], the Softmax layer in the original model is modified, the parameters of the fully connected layer are adjusted, and the transfer learning training process is carried out, according to the 7 classifications of expression data. The VGG-16 network model is trained on the training set, and the obtained model parameters are used as the initial parameters of the facial expression classification model (parameter transfer). Transfer learning increases the network's capacity for learning from tiny datasets while simultaneously accelerating the network's convergence rate. The following Figure 6 illustrates the network training procedure.

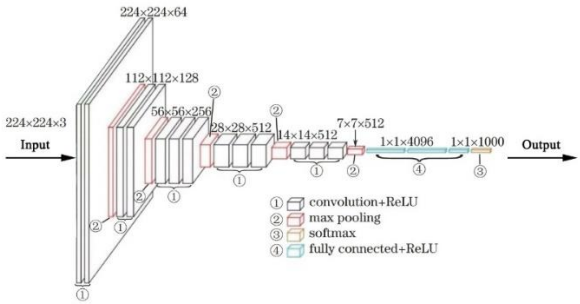

Figure 5. Structure of VGG-16

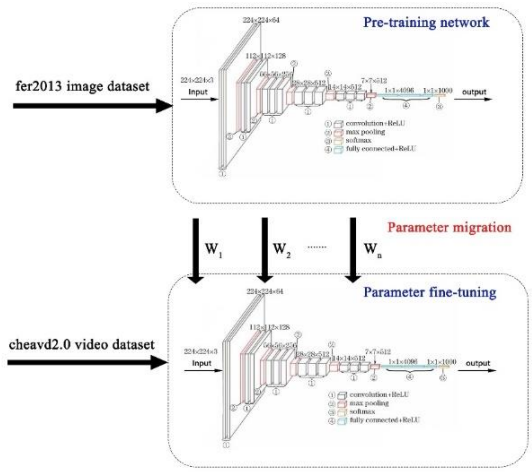

Figure 6. Training of transfer learning network

4. EXPERIMENTS AND RESULTS ANALYSIS

4.1 Dataset processing

This paper employs the cheavd2.0 audio dataset created by the Institute of Automation, Chinese Academy of Sciences, for training and testing at the voice mode recognition stage. The dataset includes more than 7,000 emotional video clips from movies, TV shows, and variety shows. The emotional tags in the dataset include angry, disgust, sad, happy, anxious, worried, surprise, and natural.

The public fer2013 image dataset and cheavd2.0 video dataset were utilized as the training and testing datasets in this paper for the stage of video mode recognition. The 35,886 facial expression images in the fer2013 data collection include 28,708 training sets, 3,589 validation sets, and 3,589 test sets. They are from many cultures and age groups. Each image is 48*48 pixels in size, and there are seven different expressions to choose from: anger; disgust; sad; happy; scared; surprised; normal. The fer2013 data collection contains several samples and has undergone preprocessing. The quality of the images is better than those taken from the cheeavd2.0 video. It can strengthen the model when used as a training set.

In the recognition process of the above two stages, the two datasets are very similar in emotion classification. In the early data processing, the worried and anxious in the voice dataset are classified as worried, so that the two databases are consistent in emotion classification, facilitating the fusion at the decision-making layer. Thus, the final emotions to be classified are labeled in 7 classes: 0 anger; 1 disgust; 2 sad; 3 happy; 4 scared; 5 surprised; 6 normal.

4.2 Feature fusion

The feature layer fusion and decision-making layer fusion are the two fundamental components of the multi-modal emotion fusion approach. The feature layer fusion method creates a super-large-dimensional vector by extracting feature vectors from audio modal feature data and video image modal feature data. When dealing with the issue of too many features, there is data redundancy, which might even result in the "curse of dimensionality." The decision-making layer merges the findings of each separator in accordance with predetermined principles after first extracting the characteristics of various modalities and sending them to their corresponding classifiers. The ideal matching of classifier weights during decision-making has strong recognition ability, and this layer is resilient against interference [24]. As a result, this paper uses the decision-making layer fusion method, in which the features of the audio and video modalities are first extracted, sent to the appropriate LSTM and VGG-16 neural network classifiers, and the results of each separator are then fused in accordance with a decision regarding the weight criterion to produce the final recognition result. The weight criterion can be expressed as:

$$E = \max \sum_{i=1}^{n} \alpha \times P_v + \beta \times P_s \tag{6}$$

where, E is the class of facial emotions; P_v is the classification probability on the video image classifier; P_s is the classification probability on the audio classifier; α=0.6 and β=0.4 are the weights of the two classifiers, respectively.

4.3 Results analysis

The effect of fusion of the decision-making layer on multi-modal emotion detection is examined in this research using the test set of cheavd2.0. Table 1 below displays the comparing results of tests on single-modal recognition and multi-modal recognition. When the "MFCC feature + FBANK feature + Spectrogram features" are paired with the "voice mode," the recognition accuracy was improved by 8.3%. The recognition accuracy increased by 7.7% after the video image modality was combined with the LBP block weighting and multi-scale partitioning features and the migration learning approach. Higher recognition accuracy is obtained following the modal fusion, demonstrating the effectiveness of the multi-modal fusion recognition technique.

Table 2 displays the multimodal emotion detection model's accuracy in identifying emotions. It can be seen that the recognition rate for normal, happy, angry, sad, and other emotions is more than 80%; Although the overall recognition accuracy is 78.5%, the recognition accuracy for the disgust and surprise emotion samples is poor, less than 70%, and the samples may actually have been mistakenly classified as another emotion type. The accuracy of recognition has increased as compared to the conventional single voice mode.

Table 1. Effects of single-modal recognition and multi-modal recognition

Modal type	Feature contents	Classification model	Recognition accuracy
Audio	MFCC features	LSTM	54.2%
	MFCC features +FBANK features+ spectrogram features	Attention-LSTM	62.5%
Video image	LBPH features	CNN	65.4%
	Features fusing LBP block weighting and multiscale partitioning	Transfer learning+VGG-16	73.1%
Audio + video image	Decision-making layer fused features	Fused network	78.5%

Table 2. Accuracy of emotion classification and recognition

Output number	Emotion class	Number of samples	Number of correctly recognized samples	Recognition accuracy
0	anger	252	203	80.6
1	disgust	42	29	69.0
2	sad	132	116	87.9
3	happy	236	208	88.1
4	scared	293	223	76.1
5	surprised	51	34	66.7
6	normal	400	324	81.0
	Total	1406	1137	78.5

5. CONCLUSIONS

The proposed approach for recognizing emotions in college students is based on voice and video images and can be used in a variety of settings, including psychological research on college students, auxiliary medical care, online learning, security monitoring, and other areas. The novel approach is to train the speech signal from the training set of the cheavd2.0 video emotion database to develop a model using an enhanced long-short-term memory artificial neural network (LSTM) based on the attention mechanism. The parameters are initialized on the fer2013 data set to train the enhanced convolutional neural network (LBPH+PCA+VGG). The network is optimized by developing the video image recognition model in conjunction with the transfer learning technique. The recognition results are combined at the level of decision-making, and the sentiment classification as well as potential for other sentiment classifications are output.

The experimental results show that, in comparison to the single-module emotion recognition method, the multi-modal fusion method can further improve recognition accuracy; however, due to the limited number of some emotion samples, the accuracy of the classification results is low. Thus, two research directions will be taken into consideration in future: First, expand the image data set during the image preprocessing stage, or create a new voice and video data set using the information gathered by the university campus face collection system and psychological census system to address the issue of insufficient samples or an uneven distribution of sample categories. Second, boost the network's ability to distinguish between features of similar expressions.

REFERENCES

[1] Thuseethan, S., Rajasegarar, S., Yearwood, J. (2022). EmoSeC: Emotion recognition from scene context. Neurocomputing, 492: 174-187. https://doi.org/10.1016/j.neucom.2022.04.019

[2] Yang, H., Fan, Y., Lv, G., Liu, S., Guo, Z. (2022). Exploiting emotional concepts for image emotion recognition. The Visual Computer, 1-14. https://doi.org/10.1007/s00371-022-02472-8

[3] Maximiano-Barreto, M.A., Bomfim, A.J.D.L., Borges, M.M., de Moura, A.B., Luchesi, B.M., Chagas, M.H.N. (2022). Recognition of facial expressions of emotion and depressive symptoms among caregivers with different levels of empathy. Clinical Gerontologist, 45(5): 1245-1252. https://doi.org/10.1080/07317115.2021.1937426

[4] Liu, N., Liu, H., Liu, H. (2021). Mental health diagnosis of college students based on facial recognition and neural network. Journal of Intelligent & Fuzzy Systems, 40(4): 7061-7072. https://doi.org/10.3233/JIFS-189536

[5] Chen, M.W., Liang, X.J., Xu, Y. (2022). Construction and analysis of emotion recognition and psychotherapy system of college students under convolutional neural network and interactive technology. Computational Intelligence & Neuroscience., 2022: 5993839. https://doi.org/10.1155/2022/5993839

[6] Li, Y.B., Zhou, Y.T. (2022). Research on psychological emotion recognition of college students based on deep learning. Scientific Programming, 2022: 6348681. https://doi.org/10.1155/2022/6348681

[7] Liu, L., Ji, Y.F., Gao, Y., Li, T., Xu, W. (2022). A data-driven adaptive emotion recognition model for college students using an improved multifeature deep neural network technology. Computational Intelligence and Neuroscience, 2022: 1343358. https://doi.org/10.1155/2022/1343358

[8] Ding, Y., Zhong, S., Hua, L. (2020). Automatic recognition of student emotions based on deep neural network and its application in depression detection. Journal of Medical Imaging and Health Informatics, 10(11): 2634-2641. https://doi.org/10.1166/jmihi.2020.3265

[9] Huddar, M.G., Sannakki, S.S., Rajpurohit, V.S. (2019). Multimodal emotion recognition using facial expressions, body gestures, speech, and text modalities. International Journal of Engineering and Advanced Technology, 8(5): 2453-2459.

[10] Wang, X.S., Chen, X., Cao, C.J. (2020). Human emotion recognition by optimally fusing facial expression and speech feature. Signal Processing: Image Communication, 84: 115831. https://doi.org/10.1016/j.image.2020.115831

[11] Xie, Y., Liang, R.Y., Liang, Z., Huang, C., Zou, C., Schuller, B. (2019). Speech emotion classification using attention-based LSTM. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(11): 1675-1685. https://doi.org/10.1109/TASLP.2019.2925934

[12] Moradi, S., Lidestam, B., Ng, E.H.N., Danielsson, H., Rönnberg, J. (2019). Perceptual doping: An audiovisual facilitation effect on auditory speech processing, from phonetic feature extraction to sentence identification in noise. Ear and hearing, 40(2): 312-327. https://doi.org/10.1097/AUD.0000000000000616

[13] Nahar, Khalid. M.O., Al-Hazaimeh, O.M., Abu-Ein, A.A.K.H., Gharaibeh, N.Y. (2020). Phonocardiogram classification based on machine learning with multiple sound features. Journal of Computer Science, 16(11): 1648-1656. https://doi.org/10.3844/JCSSP.2020.1648.1656

[14] Qiu, J., Liu, Y., Chai, Y.H. (2019) Dependency-based local attention approach to neural machine translation. CMC-Computers, Materials & Continua, 59(2): 547-562. https://doi.org/10.32604/cmc.2019.05892

[15] Xie, Y., Liang, R., Liang, Z., Zhao, L. (2019). Attention-based dense LSTM for speech emotion recognition. IEICE TRANSACTIONS on Information and Systems, 102(7): 1426-1429. https://doi.org/10.1587/transinf.2019EDL8019

[16] Li, D.D., Liu, J.L., Yang, Z., Sun, L., Wang, Z. (2021). Speech emotion recognition using recurrent neural networks with directional self-attention. Expert Systems with Applications, 173: 114683. https://doi.org/10.1016/j.eswa.2021.114683

[17] Subudhiray, S., Palo, H.K., Das, N., Mohanty, S.N. (2021). Comparative analysis of histograms of oriented gradients and local binary pattern coefficients for facial emotion recognition. In 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom), pp. 18-22.

[18] Zhu, D.M., Fu, Y.X., Zhao, X., Wang, X., Yi, H. (2022). Facial emotion recognition using a novel fusion of convolutional neural network and local binary pattern in crime investigation. Computational Intelligence & Neuroscience., 2022: 2249417. https://doi.org/10.1155/2022/2249417

[19] Liu, L., Fieguth, P., Zhao, G., Pietikäinen, M., Hu, D. (2016). Extended local binary patterns for face recognition. Information Sciences, 358: 56-72. https://doi.org/10.1016/j.ins.2016.04.021

[20] Huang, D.S., Jiang, F.W., Li, K.P., Tong, G.S., Zhou, G.F. (2022). Scaled PCA: A new approach to dimension reduction. Management Science, 68(3): 1678-1695. https://doi.org/10.1287/mnsc.2021.4020

[21] Saedi, S.I., Khosravi, H. (2020). A deep neural network approach towards real-time on-branch fruit recognition for precision horticulture. Expert Systems with Applications, 159: 113594. https://doi.org/10.1016/j.eswa.2020.113594

[22] Sariyanidi, E., Gunes, H., Cavallaro, A. (2014). Automatic analysis of facial affect: A survey of registration, representation, and recognition. IEEE transactions on pattern analysis and machine intelligence, 37(6): 1113-1133. https://doi.org/10.1109/TPAMI.2014.2366127

[23] Nagaral, M.U., Reddy, T.H. (2018). Hybrid approach for facial expression recognition using HJDLBP and LBP histogram in video sequences. International Journal of Image, Graphics and Signal Processing, 10(2): 1-9. https://doi.org/10.5815/ijigsp.2018.02.01

[24] Huddar, M.G., Sannakki, S.S., Rajpurohit, V.S. (2020). Multi-level feature optimization and multimodal contextual fusion for sentiment analysis and emotion classification. Computational Intelligence, 36(2): 861-881. https://doi.org/10.1111/coin.12274