An Emotion Recognition Method Based on Eye Movement and Audiovisual Features in MOOC Learning Environment

Jindi Bao, Xiaomei Tao, and Yinghui Zhou

***Abstract*— In recent years, more and more people have begun to use massive online open course (MOOC) platforms for distance learning. However, due to the space–time isolation between teachers and students, the negative emotional state of students in MOOC learning cannot be identified timely. Therefore, students cannot receive immediate feedback about their emotional states. In order to identify and classify learners' emotions in video learning scenarios, we propose a multimodal emotion recognition method based on eye movement signals, audio signals, and video images. In this method, two novel features are proposed: feature of coordinate difference of eyemovement (FCDE) and pixel change rate sequence (PCRS). FCDE is extracted by combining eye movement coordinate trajectory and video optical flow trajectory, which can represent the learner's attention degree. PCRS is extracted from the video image, which can represent the speed of image switching. A feature extraction network based on convolutional neural network (CNN) (FE-CNN) is designed to extract the deep features of the three modals. The extracted deep features are inputted into the emotion classification CNN (EC-CNN) to classify the emotions, including interest, happiness, confusion, and boredom. In single modal identification, the recognition accuracies corresponding to the three modals are 64.32%, 74.67%, and 71.88%. The three modals are fused by feature-level fusion, decision-level fusion, and model-level fusion methods, and the evaluation experiment results show that the method of decision-level fusion achieved the highest score of 81.90% of emotion recognition. Finally, the effectiveness of FCDE, FE-CNN, and EC-CNN modules is verified by ablation experiments.**

***Index Terms*— Emotion recognition, feature extraction, massive online open course (MOOC), multimodal analysis.**

I. Introduction

EMOTIONS play a very important role in our life. People feel emotions in daily life, interpersonal communication, decision-making, learning, or cognitive activities. Emotion recognition is an interdisciplinary issue, including computer science, psychology, and cognitive science. An important application of emotion recognition is online education based on massive online open course (MOOC). Since the first year of MOOC in 2012 [1], MOOC has developed rapidly. The three major platforms, Coursera, Udacity, and edX, have cooperated with universities around the world, making MOOC very popular. Online MOOC learning platforms have become even more important and necessary since the COVID-19 outbreak in 2020, as learners rely more than ever on distance learning as a source of their education. Compared with traditional teaching methods, the core advantage of MOOC is that they are not constrained by time and space, making learning more flexible and thus promoting the sharing of high-quality educational resources. However, MOOC is also facing a number of problems, for example, one of the biggest concerns is the high dropout rate. According to the University of Pennsylvania, the average dropout rate of MOOC is as high as 90% [2]. Most of the current studies have focused on the impact of course quality, course evaluation, and the influence of students' messages on students' grades. A few studies have focused on the impact of students' emotional states on learning outcomes, especially in online video learning [3]. In fact, students' learning outcomes are affected by their emotional state. Generally speaking, a positive emotional state will promote students' learning outcomes, and similarly, negative states reduce the efficiency of the learners. Therefore, emotion plays an important role in MOOC learning [4].

At present, the common modals used for emotion recognition are physiological signals, facial expressions, voice signals, texts, and so on. Most physiological signals, such as EEG signals, are collected by wearing sensors, which may cause the participant to have unnatural reactions. For students in learning, wearing sensors will distract them and produce interference to their learning. Therefore, it is hard to distinguish whether the emotions are caused due to the electrode cap or the learning materials. Eye tracker can record the characteristics of eye movements, while the person is processing visual information. As a nonintrusive device, screen eye tracker will not interfere with the learners' learning experience. Thus, it is suitable for MOOC learning environments. The generation of emotion is closely related to the context of the stimulating material. In the MOOC learning environment, students' emotional state is mainly stimulated by video learning materials. If the speaker

Manuscript received 31 March 2022; revised 16 June 2022, 8 August 2022, 14 September 2022, and 25 October 2022; accepted 1 November 2022. This work was supported in part by the National Science Foundation of China under Grant 61906051 and Grant 62267001, in part by the Guangxi Natural Science Foundation Program under Grant 2018GXNSFBA050029, and in part by the Foundation for Doctoral Research of Guilin University of Technology under Grant GUTQDJ2005015. *(Corresponding author: Xiaomei Tao.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of Guangxi Normal University.

Jindi Bao and Yinghui Zhou are with the School of Information Science and Engineering, Guilin University of Technology, Guilin 541004, China (e-mail: 94335913@qq.com; zhouyh9794@163.com).

Xiaomei Tao is with the School of Computer Science and Engineering and the School of Software, Guangxi Normal University, Guilin 541000, China (e-mail: txm800505@163.com).

Digital Object Identifier 10.1109/TCSS.2022.3221128

has a flat tone, the students may feel uninterested, while a tone that appropriately fluctuates will attract the students' attention and enhance their interest. Monotonous video images will make students feel bored, while fast-changing images will evoke students' interest in the material matter. Therefore, the audio and visual features in the instructional video, as part of the MOOC learning scenarios, are just as important as the eye movement features to infer the emotional state of students. However, there are few studies to analyze students' emotions by combining eye movement signals and audiovisual features.

This article proposes an emotion recognition method based on the fusion of eye movement signals and audiovisual features in a video learning scenario. This method focuses on the audio and video images of learning videos and the eye movement signals generated by learners and integrates them through a variety of methods. The main contributions of this article are given as follows.

1) This study proposes that not only learners' physiological signals, such as eye movement features, but also learning scenario features, such as instructional video features and audio features, should be considered for learner emotion classification in MOOC learning scenarios. The combination of physiological features and scenario features is helpful to improve the classification accuracy. The experimental results show that learning scenario features are closely related to the learners' emotional states and can improve the recognition accuracy of the model.
2) In order to better represent learners' eye movement features and scenario features in video learning scenes, the feature of coordinate difference of eyemovement (FCDE) and pixel change rate sequence (PCRS) are proposed. FCDE is extracted by combining eye movement coordinate trajectory and video optical flow trajectory, which can represent the learner's attention degree. In addition, PCRS is extracted from the video image, which can represent the switching speed of the image.
3) In order to better extract the deep features and the complementary relationship between features from one single modal, a feature extraction network feature extraction network based on convolutional neural network (CNN) (FE-CNN) is designed. The network can extract the deep features without losing the original shallow features and integrate the shallow features with the deep features. Through a series of experiments, an effective and optimal model of multimodal emotion classification is determined, which integrates the new feature proposed in 2), the deep and shallow feature fusion network, and the multimodal decision fusion module.

II. Related Work

A. Eye Movement-Based Approaches

The research of emotion recognition using eye movement signals has attracted more and more attention in recent years. Li et al. [5] proposed a method that combined eye movement signals with other physiological signals to identify depression and achieved good classification accuracy. The method proves that eye behavior is one of the main features of depression, which has an important reference value for the automatic diagnostic system for establishing clinical applications. Tarnowski et al. [6] evoked people's emotions by presenting 21 dynamic film video clips. Emotion categories are high arousal and low valence; low arousal and moderate valence; and high arousal and high valence. The highest average classification accuracy is 80% obtained by eye movement features. In [7], a method of combining electrooculogram (EOG) signal and eye movement signal is proposed. According to the eye movement track, the invalid EOG signal can be removed. The quality of EOG signal is improved and the accuracy of classification recognition is improved. Eye movement is used in many fields, but at present, it is mostly used to supplement EEG and EOG signals [8]. In [9], EEG signals and eye movement features were fused by deep canonical correlation analysis (DCCA) to achieve a good recognition effect. Also, it was found that EEG signals and eye movement signals are complementary to each other in distinguishing positive and negative emotions. In [10], [11], and [12], the stability of EEG and eye movement signals over time was studied. The two types of signals were fused at different levels, and the results showed that the fusion can provide more supplementary information for identifying emotions. For the identification of neutral emotions and fear, eye movement signals have obvious advantages. In addition, in order to obtain high arousal emotion, most experiments use movie clips with strong stimulation [6], [9], [10], [11]. However, the academic emotion and its intensity induced in MOOC learning scenarios are different, which needs further research.

B. Audio-Based Approaches

Speech is a type of complex signal, which contains a variety of information, such as the message to be conveyed, the speaker's language, gender, and emotion. Speech emotion recognition is very important for natural human–computer interaction. In [13], the acoustic features are extracted from the speech signal, and the mel frequency cepstrum coefficient (MFCC) coefficients are extracted to recognize the speaker's emotion. The average recognition accuracy of this method for happiness, sadness, and anger is about 80%. In [14], a speech emotion recognition model based on an improved deep belief network (DBN) is proposed to enhance the representation ability of speech signals and the recognition accuracy of speech emotion recognition. The model uses RELU instead of the traditional DBN activation function to extract the short-time energy, short-time zero crossing rate, the fundamental frequency, formants, and MFCC coefficients of the speech signal as the basic features. Using these features as the input of the model, the model can automatically recognize six emotions: anger, fear, joy, calmness, sadness, and surprise, with an average recognition accuracy of about 60%. In [15], a new speech emotion recognition technology based on the combination of deep and shallow neural networks is proposed. The parallel training sample set is established, and the DBN is used to automatically extract and recognize the speech emotional features. The shallow neural network is used to obtain the final recognition results. The five emotions of sadness, surprise, anger, happiness, and neutral emotions were classified, and the average recognition accuracy was 89.8%. At present, audio modal is widely used in emotion recognition, but it is rarely used as the stimulus signal and combined with subjects' physiological.

C. Video-Based Approaches

At present, video is the most common stimulus material used to evoke subjects' emotions. Mao et al. [16] proposed a multimodal local–global attention network (MMLGAN) for affective video content analysis, which extends the attention mechanism to multilevel fusion and includes a multimodal fusion unit to obtain a global representation of affective video. The effectiveness of the method was proven by public datasets. In [17], a variety of machine learning algorithms and neural network models are used to fuse video features and EEG signals. The result shows that the video emotion classification accuracy achieves 96.79% for valence (positive/negative) and 97.79% for arousal (high/low). In [18], audiovisual features were obtained by 3-D CNN and then fused into DBNs. This method performs well on three common datasets. From the above research, it can be seen that movie clips are currently selected in most studies, but learning videos in MOOC are seldom studied as stimulating materials.

The rest of this article is arranged as follows. Section III introduces the data collection experiment in the MOOC learning scenarios. Section IV introduces the process of data preprocessing and feature extraction, including FCDE and PCRS. In Section V, the single modal emotion classification experiment and multimodal emotion classification experiment are introduced and the experimental results are analyzed. Section VI draws the conclusion.

III. Data Collection

This section is part of the data collection experiment related to this article. In the experiment, we selected four instructional videos of different types as stimulating materials. The eye tracking device is TobiiTX300, which includes an eye movement module and a 23-in display module, and the sampling frequency is 60 Hz/s. An HP desktop computer connected to the eye tracking device is used to control the experimental procedure and record the data. The 68 subjects are all college students aged 20–23, with a male-to-female ratio of 1:1.

The experiment was carried out in a laboratory environment with constant brightness. The eye tracker was calibrated based on the individual subject. Then, the subjects were asked to stare at a cross in the center of the screen for 30 s to obtain the baseline value of the pupil diameter. In the process of learning by video, when the subjects felt bored, interested, happy, or confused, they can press the corresponding button on the keyboard to mark it. After learning, the subjects need to review the study videos and videos of their facial expressions during the study and expand the marked point into an emotional event. Each emotion event has a pair of start and end points, and an emotional intensity value, from strong (A5) to weak (A1) across five scales.

Excluding the subjects whose pupil calibration accuracy is less than 80% and also excluding the subjects whose data loss rate is more than 25%, the eye movement data of 59 subjects are finally selected. The subjects' eye movement data and synchronized audio and video image data of the instructional videos are extracted to build the datasets for this study.

IV. Data Preprocessing and Feature Extraction

A. Data Preprocessing

There might be a deviation at the start time and the end time of the emotional event labeled by the subjects during the review, so the first 30 frames and the last 30 frames of the interval are removed. The characteristics of the data with weak emotional intensity are not obvious enough, so we extracted the data with emotional intensity between A3 and A5 to build the dataset for this experiment.

The most commonly used indicators of eye movement signals are pupil diameter and eye movement coordinates (also called gaze points, the pixel coordinates of the eye on the screen are calculated by the Pupil-CR technology [19]). The common eye states are eye blinking, fixation, saccade, and so on. During the data acquisition process, data of the pupil diameter and eye moving coordinates might be lost due to eye blinking and excessive head movements. The eye movement data might be lost due to the shake of the head and the low calibration accuracy. Next, the methods of data completion for missing eye state, pupil diameter, and gaze points are introduced.

Pupil diameter is a kind of physiological information, which conforms to the law of continuous changes in physiological signals over time. In order to avoid the change of data structure and reduce the data standard difference, the linear interpolation method is used to complete the lost pupil diameter data, as in the following equation:

$$y = y_0 + \frac{(x - x_0)y_1 - (x - x_0)y_0}{x_1 - x_0} \tag{1}$$

where (x_0, x_1) is a known pupil diameter pair, (y_0, y_1) is a known time pair corresponding to (x_0, x_1), x is the lost pupil diameter data, and y is the time corresponding to the lost pupil data. To find the missing value x, Formula (1) is transformed to get the following equation:

$$x = x_0 + \frac{(y - y_0)(x_1 - x_0)}{y_1 - y_0}. \tag{2}$$

Gaze points are not physiological information, and the coordinate changes are all controlled by the subjects. The change of gaze points has no obvious rule, and the distribution of gaze points is not continuous. Thus, linear interpolation is not suitable for the missing gaze points. Therefore, the average interpolation method is used to supplement the data of total gaze points, as in the following equations:

$$z = \frac{\sqrt{(gp_{y1} - gp_{y0})^2 + (gp_{x1} - gp_{x0})^2}}{(\mathrm{n} + 1)} \tag{3}$$

$$(gp_x, gp_y) = (gp_{x0} + i * z, gp_{y0} + i * z) \tag{4}$$

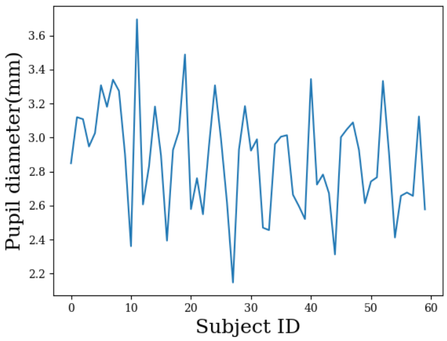


Fig. 1. Pupil diameter baseline value of each subject.

where n is the number of the missing values, i means the ith missing values ($i = 1, 2, \ldots, n$), z is the difference between a known value and an adjacent missing value, (gp_{x0}, gp_{y0}) is the eye movement coordinate before the first missing value, (gp_{x1}, gp_{y1}) is the eye movement coordinate after the last missing value, and (gp_x, gp_y) is the missing eye movement coordinate data. Eye movement coordinate means the position of the subject's gaze on the screen.

Eye state is a constant process for each state, for example: fixation, whereby the subject gazes upon a certain time interval, the eye state remains unchanging. The lost eye state data can only be copied from the previous frame.

Eye movement signals belong to physiological signals, and baseline values of physiological signals in a state of calm are different for each individual [20], [21], [22]. Therefore, we calculated the pupil diameter baseline value of 59 subjects in a calm state, as shown in Fig. 1. As can be seen from the figure, the pupil diameter of different subjects varied greatly. The subjects with the largest pupil diameter baseline value are 1.72 times that of the subjects with the smallest. Therefore, in order to eliminate individual differences in pupil diameter [23], the individual pupil diameter baseline value is subtracted from pupil diameter for each subject, as in the following equation:

$$x = x_0 - \frac{\sum_{i=1}^{n} x_i}{n} \tag{5}$$

where x_0 is the original pupil diameter and n is the number of data frames in the process of watching the cross; thus, the average pupil diameter values $(\sum_{i=1}^{n} x_i/n)$ are calculated as the individual pupil diameter baseline value.

B. Feature Extraction of Eye Movement Signal

Wang et al. [24] studied intention recognition through pupil diameter, saccade amplitude, and fixation time, and found that the more eye movement indicators there were, the higher the recognition accuracy would be, and the position characteristics of fixation points had a great influence on the accuracy. Therefore, we extract statistical features from the four indexes of fixation, saccade, eye blink, and pupil diameter as a part of the index of eye movement signal. After correlation calculation, these characteristics are highly correlated with emotional state. In addition to the above features, this article proposed the FCDE. All the eye movement features used in this study are shown in Table I. The sampling window size is a fixed time interval, such as 2 s.

TABLE I

FEATURES EXTRACTED FROM EYE MOVEMENT SIGNALS

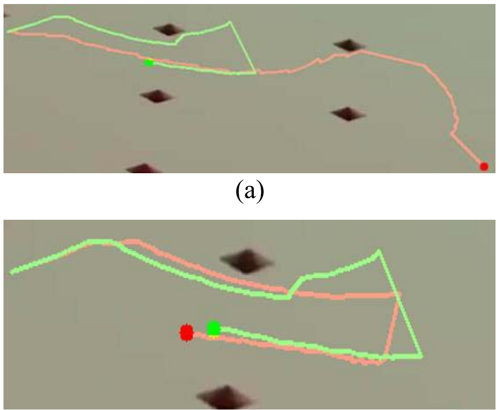
Index	Features of eye movement index
fixation	Fixation times, maximum, minimum and average of fixation time, average fixation speed.
saccade	The number of saccades, the maximum, minimum and average of saccade duration, and the average of saccade speed.
eye blink	Blink frequency, maximum, minimum and average of blink duration
pupil diameter	The maximum, minimum, average, standard deviation, variance of the pupil diameter of the left and right eyes. The crest factor pulse factor, margin factor, kurtosis factor, wave factor and skewness factor of the pupil diameter of the left and right eyes. The standard deviation, variance and statistical characteristics of the first-order difference of mean pupil diameter of left and right eyes.
eye movement coordinate	Feature of coordinate difference of eye movement.

Only individual eye movement coordinates cannot provide enough effective information, so the feature FCDE is extracted by combining these data with the corresponding video data. In terms of the different states of eye movement, FCDE is divided into two features FCDE_s and FCDE_f. FCDE_s means the FCDE in a saccade process, and FCDE_f means the FCDE in fixation. FCDE_s is calculated with Formula (6). In a saccade process, the eye movement coordinate of the first frame $(x_{p1,}y_{p1})$ is taken as the starting point of the eye movement trajectory. The coordinates of the same position as (x_{p1}, y_{p1}) in the video image are called feature points (also known as corner points), denoted by (x_{v1}, y_{v1})

$$\text{FCDE}_s = \frac{\sum_{i=1}^{n} \sqrt{((x_{pi} - x_{vi})^2 + (y_{pi} - y_{vi})^2)}}{n}. \tag{6}$$

The trajectory of the saccade is obtained by successively connecting the eye movement coordinate points in continuous frames. The trajectory of the feature point in the image is obtained by successively connecting the feature point. The new coordinate of the feature point in the next video frame (x_{vi}, y_{vi}) can be obtained by using the optical flow method [25]. From frame $i = 1$ to n, calculate the distances of every pair of corresponding coordinate points of the two trajectories, sum up the distances, and finally calculate their mean value. n indicates the number of frames sampled during saccade. There are two examples of feature point trajectory and eye movement coordinate trajectory, which is shown in Fig. 2. In Fig. 2(b), in a state of confusion, the two tracks are in the same direction, which means that the subject still focuses on a particular object on the screen. However, in Fig. 2(a), in a state of boredom, the two tracks start in the same direction, but then there is a large deviation, which indicates that the subject's attention has declined from the particular object.

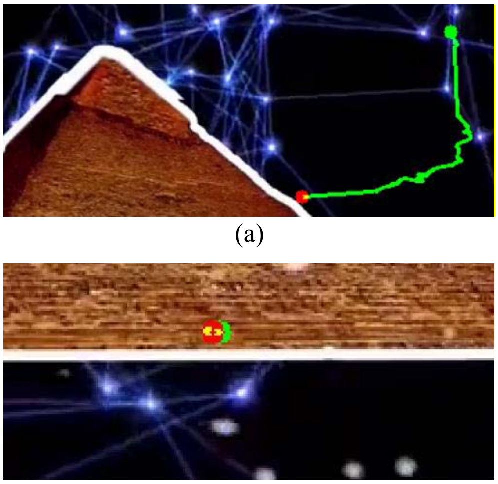
FCDE_f is calculated with Formula (7). During fixation, the eye movement coordinate (x, y) remains the same, and taking this coordinate as the feature point of the first frame in the video image, the new coordinate of the feature point in the



(a)

(b)

Fig. 2. Feature point trajectory and saccade trajectory. Green is the feature point track and red is the saccade track. (a) Track in a state of boredom. (b) Track in a state of confusion.



(a)

(b)

Fig. 3. Gaze points and feature points trajectory in fixation state. (a) Trajectory in boredom state. (b) Trajectory in interest state.

TABLE II

Correlation Coefficient Between Eye Movement Features and Emotional State

Eye movement features	$FCDE_s$	$FCDE_f$	The average coefficient of other features
Correlation coefficient	0.41	0.42	0.22

next video frame $(x_{vi},\ y_{vi})$ is obtained by the optical flow method as well such as in the saccade process, then summing the distance of the new coordinates and the gaze points, and finally calculating the average value

$$\mathrm{FCDE}_f = \frac{\sum_{i=1}^{n} \sqrt{((x - x_{vi})^2 + (y - y_{vi})^2)}}{n}. \tag{7}$$

There are two examples of gaze points' trajectory and feature points' trajectory under the fixation state, as shown in Fig. 3. Gaze points are red and feature points are green. In Fig. 3(a), in a state of boredom state, the trajectory of feature points has a large deviation from the trajectory of gaze points. In Fig. 3(b), in a state of interest, the feature trajectory is very close to the trajectory of gaze points, and the overlapping part of the two trajectories is yellow.

In Formulas (6) and (7), (x, y) is the original coordinate of eye movement, and (x_{pi}, y_{pi}) and (x_{vi}, y_{vi}) are the coordinates of eye movement in frame i and the feature point coordinates in frame i, respectively. The smaller FCDE_s or FCDE_f is, the more similar that the trajectory of gaze points and the trajectory of features points are. Also, this means that the subjects' attention is more concentrated.

If there are more than one fixation or saccade process in a sample window, the average value of each FCDE_s and FCDE_f is calculated as the value of FCDE.

The effectiveness of FCDE features is verified from two aspects: correlation and recognition accuracy. Table II lists the correlation coefficients between FCDE features and emotional state, as well as the average correlation coefficients between other eye movement features and emotional state. From this table, we can see that the correlation between FCDE features and emotional state is much higher than the average correlation between other eye movement features and emotional state. This indicates that FCDE features, compared with most other features, are more suitable for emotion classification in learning scenarios with video.

C. Video Feature Extraction

Instructional video is used as stimulus material to induce the learners' emotional states, so audio and image features are closely related to the emotional state of learners. We extract the features of audio and image modals from the instructional video for further emotion classification. For audio feature extraction, the mel cepstral coefficient in audio has good robustness and accuracy in emotion recognition [26], so for each frame, we extract the MFCC coefficients of audio and calculate its first-order difference, which can further reflect the dynamic characteristics of audio.

Using the method in [27], for each frame, we obtained the 13 MFCC coefficients from the audio spectrum through a set of MFCC filters using different parameters. Because in the time domain, the characteristics of the signal are difficult to be shown, the Fourier transform is used to transfer the signal to the frequency domain. The spectrum is smoothed and the influence of harmonics is eliminated through the MFCC filter. MFCC coefficients are obtained by discrete cosine transform (DCT). Then, the first-order differentials of MFCC coefficients are calculated. For a given time window, a total of 15 statistical features, including the maximum, minimum, mean, range, median, standard deviation, and variance of MFCC coefficient, the same statistical values of the first-order differentials of MFCC coefficient, and the kurtosis factor, are extracted. A total of 195 features are extracted from the 13 MFCC coefficients.

For video image feature extraction, PCRS based on the sequence of pixel change rates is proposed. Most studies use some exciting clips in the movie to induce the subjects' emotions, in which the high-speed switching of the scene and the fast movement of the target can elicit strong emotion. However, there are few high-speed switching screens and fast-moving targets in instructional videos. The feature PCRS is supposed to be related to learners' emotional arousal. PCRS is calculated in Formulas (8) and (9). Z in Formula (8) represents the level of the pixels changed in two adjacent frames of images, and C in Formula (9) indicates how fast the images are switched in a time window

$$Z = \frac{\sum_{i,j} A_{t(i,j)}}{n}. \tag{8}$$

In Formula (8), $A_t = X_t - X_{t-1}$ is the matrix of difference between two adjacent grayscale images too. Take the absolute value of the entries in the matrix. X_t represents the pixel gray value matrix in frame t. n is the total number of pixels in the gray image. $X_t(i, j)$ is the element in row i and column j of the matrix, which is a pixel gray value. Z is obtained by summing up the values of all the elements in the matrix A_t and then average them. Z represents the pixel change level between the two images. Calculate the value of Z for every two adjacent images in the window to get a sequence of pixel change level Z seq $= [Z_1, Z_2, \dots, Z_{u-1}]$, where u is the number of image frames in the window. This sequence can reflect the change of the video image. Usually, slow change in the sequence makes people feel dull or bored, and fast change sequence causes people to feel higher emotional arousal. We calculate the total number of the changed pixels for adjacent images in a sampling window, which is represented by C in the following equation:

$$C = \sum_{i=2}^{m} R^*_{A_t}. \tag{9}$$

In Formula (9), A_t is the matrix of difference between two adjacent grayscale images and m is the number of the grayscale images in a sampling window. $R^*_{A_t}$ is the number of nonzero elements in the matrix A_t. Therefore, the greater the value of C, the greater the pixel change of the image in the time window. A total of seven features of the video image include the mean, maximum, minimum, median, standard deviation, variance of the sequence Z, and the number of nonzero pixels C.

D. Feature Dimension Reduction

Principal component analysis (PCA) is used to reduce feature dimensions and eliminate redundant information. In order to reduce the interference of the test set to the training set, the data of 12 subjects who are randomly selected out from 59 subjects are used as the test set. First, the dimension reduction is carried out on the data of the remaining 47 subjects and the feature matrix is obtained. Then, the feature matrix is used to reduce the dimension of the test set. It shows the contribution rate and accumulated contribution rate of the feature principal components of eye movement signal in Fig. 4(a), principal components of MFCC coefficient feature in Fig. 4(b), and principal components of video image feature in Fig. 4(c). Since the features in the same modal are highly correlated, we reduced the dimensions of the three modals to retain the principal component whose contribution rate is greater than 1%. Finally, there are 19 principal components extracted from 52 features in eye movement signals, 19 principal components extracted from 195 features in audio signals, and three principal components extracted from seven features in video image that are retained.

V. Experiment

This section includes a single modal emotion recognition experiment and multimodal emotion recognition experiment. Shen et al. [28] found that time windows of different lengths had a great impact on experimental results, so we determined the optimal window through experiments. In [29], it was found that the best window for emotion recognition of eye movement signals is 2–3 s. Zhang et al. [30] used the audio dataset with a time window of 3 s to identify the five emotional states and the accuracy rate is 87.8%. In [31], a generation adversarial network (GAN) was used to generate 3–5-s speech samples to enhance the dataset, which improved the speech emotion recognition accuracy rate by 5%. Therefore, the dataset is divided with the time window of 2, 3, and 5 s in this study. The experiment uses the Pytorch open-source framework to train the network with Windows10 and NVIDIA GeForce RTX 2080 GPU. In parameter settings, Adam is selected as the optimizer, and the learning rate is set to 0.005, the batch size is 32, and the dropout is 0.5. In the single modal experiment, the optimal window size is selected from the three time window sizes and the features in the three modals under the optimal window are analyzed further. In the multimodal experiment, the best fusion strategy is selected by three methods: the feature-level fusion method, the decision-level fusion method, and the model-level fusion method.

A. Single Modal Emotion Recognition

1) Construction of FE-CNN and EC-CNN: Single modal emotion classification uses the combination of shallow features and deep features to classify the emotional state. The specific emotional categories include interest, happiness, boredom, and confusion. The original features after PCA dimension reduction belong to shallow features, which contain more detailed information, but the original feature has less semantic information and higher noise. The deep feature is the feature vector extracted from the original feature by the FE-CNN network. Compared with shallow features, deep features are more abstract, can represent the internal relations between different features, and contain more potential information. FE-CNN network parameters are shown in Table III. Inspired by the residual idea from the Resnet network, in the fusion process of shallow features and deep features, after each extraction of deep features, the shallow features are added to the deep features by means of addition and concatenating, so as to ensure that the original shallow features are also retained while extracting deep features. By this method, the shallow features and deep features are fully extracted and fused to provide high-quality input for the classifier. The FE-CNN network is only used for feature extraction without feature reduction and emotion classification.

Fig. 5 shows the process of fusion using original features and deep features. Taking eye movement features as an

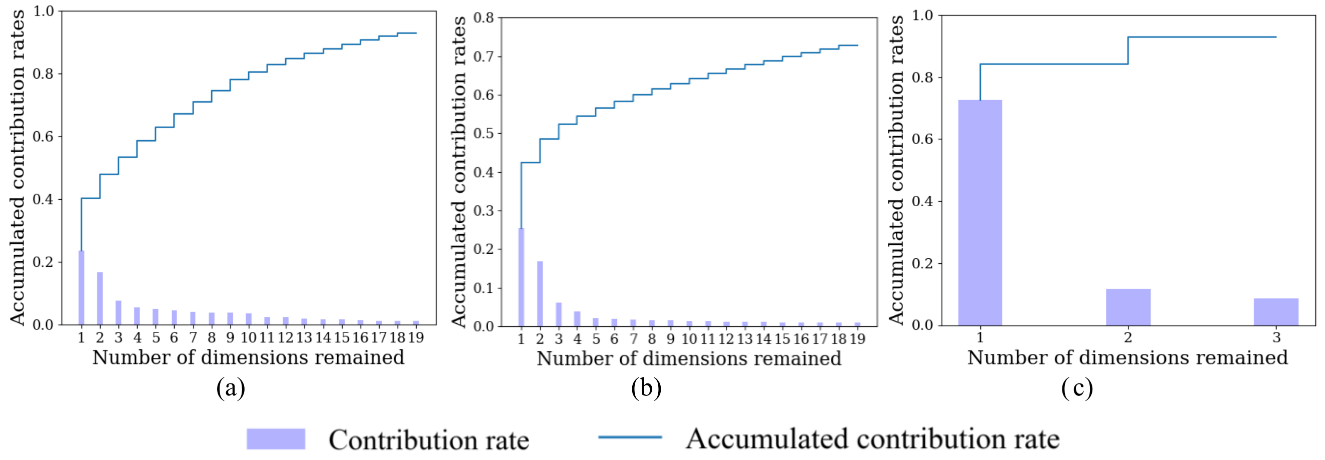


Fig. 4. Principal component contribution rates and accumulated contribution rates of PCA after dimensionality reduction. (a) Eye movement signal feature. (b) MFCC coefficient feature. (c) Video image feature.

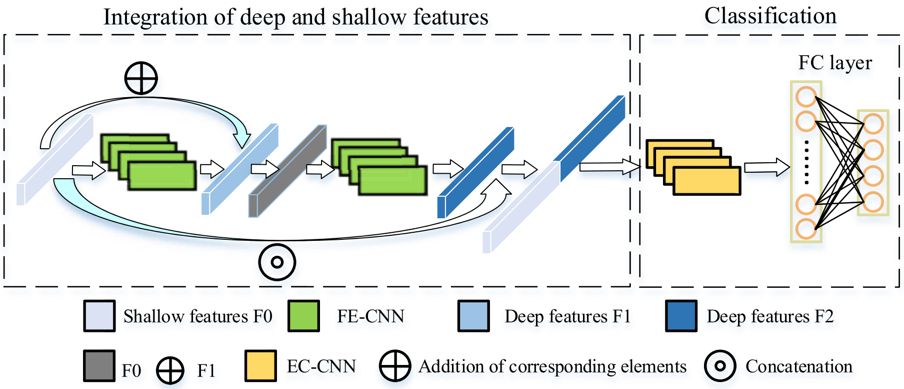


Fig. 5. Flowchart of original feature and deep feature fusion of eye movement signal, audio signal, and video image.

TABLE III
STRUCTURE AND PARAMETERS OF FE-CNN MODEL

Convolution layer	Insize	Outsize	Kersize	Stride	Padding
Layer 1	1	3	3	1	1
Layer 2	3	16	3	1	1
Layer 3	16	3	3	1	1
Layer 4	3	1	3	1	1

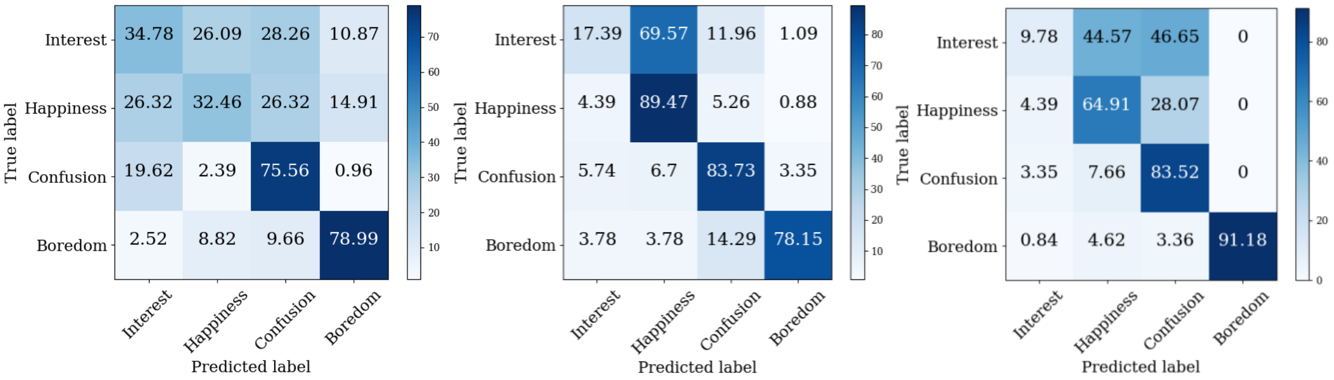
TABLE IV
STRUCTURE AND PARAMETERS OF EC-CNN MODEL

Convolution layer	Insize	Outsize	Kersize	Stride	Padding
Layer 1	1	16	3	1	1
Layer 2	16	32	3	1	1
Layer 3	32	64	3	1	1
Layer 4	64	128	3	1	1

example, as shown in Fig. 5, the original features $F0 = [\text{PCA}_1, \text{PCA}_2, \ldots, \text{PCA}_{19}]$ are sent into FE-CNN and the deep features $F1 = [F1_1, F1_2, F1_3, \ldots, F1_{19}]$ are obtained after the first feature extraction. The deep feature $F1$ is added to the shallow features to supplement the abstract information of the shallow features, $F0 + F1 = [\text{PCA}_1 + F1_1, \text{PCA}_2 + F1_2, \ldots, \text{PCA}_{19} + F1_{19}]$. The feature vector $F0 + F1$ is sent into FE-CNN again for the second extraction to obtain deep feature $F2 = [F2_1, F2_2, F2_3, \ldots, F2_{19}]$. Feature vector $F3 = [\text{PCA1}, \text{PCA2}, \ldots, \text{PCA19}, F1, F2, \ldots, F19]$ is obtained by concatenating the deep features $F2$ and the shallow features $F0$. The modals of audio and video image are processed in the same way.

In this experiment, a convolution neural network called emotion classification CNN (EC-CNN) is designed for classification. Because the shallow features in the feature sequence are obtained by PCA dimension reduction, the redundant information is removed, and deep features have been extracted twice by FE-CNN, so the classifier does not need too many network layers. The EC-CNN model has four-layer convolution, and the BatchNorm1d layer and RELU activation function are added between convolution layers. The parameters of EC-CNN are shown in Table IV.

2) Experimental Results and Evaluation: As mentioned before, 2, 3, and 5 s are used as time window sizes for the data segmentation of the three modals. Cross validation is suitable for small sample datasets, so we use the evaluation method



(a) (b) (c)

Fig. 6. Confusion matrix of classification results by using different modal features. (a) Confusion matrix of eye movement modal. (b) Confusion matrix of audio modal. (c) Confusion matrix of video image modal.

TABLE V

Emotional States Recognition Accuracy With Different Single Modals/Different Time Windows/Methods

Modal / Model		Eye movement signal(%)	Audio signal(%)	Video images(%)
2s	SVM	41	72	34
	RF	38	62	22
	1D-ResNet18	50.74	73.8	44.34
	FE-CNN+EC-CNN	53.68	**74.94**	71.51
3s	SVM	59	63	69
	RF	50	60	58
	1D-ResNet18	58.4	65.3	70.8
	FE-CNN+EC-CNN	**64.32**	**74.67**	**71.88**
5s	SVM	33	56	40
	RF	34	57	33
	1D-ResNet18	36.7	68	46.8
	FE-CNN+EC-CNN	43.24	70.28	70.16

of tenfold cross validation [32]. The best time window size is determined based on the average recognition accuracy on the test set

Table V shows the recognition accuracy of the three modals using four different machine learning models by three different time window sizes. The results show that the accuracy of emotion recognition using only audio or video images is higher than that using only eye movement signals, so it is confirmed that the audio and video image features of MOOC videos are closely related to the emotion generated by learners. Combining individual eye movement features with environmental features can further improve the accuracy of emotion recognition. This will be discussed in Section V-B.

From the aspect of modal, the audio signal has the highest recognition accuracy. From the aspect of time window size, the eye movement signal and the video image achieve the highest accuracy in the 3-s window. The recognition effect of the eye movement signal is particularly prominent in the 3-s window. The audio signal has the best result in the 2-s window, but the recognition accuracy of 74.94% is very close to the accuracy value of 74.67% in the 3-s window, so we chose 3 s as the time window size to build the dataset of all modals.

For each modal, ten models were obtained from the tenfold cross validation and sorted according to the recognition accuracy on the test set, and the fifth model in the middle was selected for analysis. The reason why we choose the model with the accuracy ranking in the middle is to make the output result more consistent with the general actual situation [33].

Fig. 6 shows the confusion matrix of classification results by eye movement signal, audio signal, and video image with the model. Table VI shows the precision (P), recall (R), specific (S), and $F1$ scores of each modal. As can be seen from Fig. 6 and Table VI, among the three modals, the classification accuracy and other evaluation results of boredom are very high, which is the easiest to identify compared to the other emotions of interest, happiness, and confusion. In audio modal and video image modal, confusion and happiness also have a high recognition accuracy. However, the accuracy of interest is low in the three modals. In the eye movement modal, interest is easily misidentified as happiness and confusion, and in the audio modal and video image modal, interest is mostly misidentified as happiness.

As can be seen from Table VI, the $F1$ score of happiness, confusion, and boredom in audio and video image modals is all high, indicating that audio and video image modals are effective in the recognition of happiness, confusion, and boredom. In addition, the precision and recall of these three emotions in the two modals of audio and video images are all greater than 0.9, and in the eye movement modal, the precision and recall of confusion and boredom are also high. This indicates that the model has a strong ability to classify confusion and boredom. Compared with the other three emotions, the precision and recall of interest in the three modals are relatively low, indicating that interest can easily be identified as other emotions by this model. Fig. 7 shows the receiver operating characteristic (ROC) curves of the three modals. As can be seen from Fig. 7, the average ROC curve area under curve (AUC) of the three modals is all greater than 0.8, which proves that the three modals are relatively stable and have strong generalization ability. Also, the three modals all have 1–3 AUC values greater than 0.9, which means that the three modals have a better recognition effect on one or several emotions. The average AUC of the audio modal is the largest, while that of the eye movement modal is the smallest. AUC values of boredom in the three modals are all greater than 0.9. The AUC values of confusion and happiness states

TABLE VI

EVALUATION OF SINGLE MODAL EMOTION RECOGNITION MODEL IN DIFFERENT AFFECTIVE STATE

	Eye movement modal				Audio modal				Video image modal			
	P	R	S	F1	P	R	S	F1	P	R	S	F1
Interest	0.71	0.63	0.83	0.67	0.79	0.58	0.95	0.67	0.78	0.45	0.97	0.57
Happiness	0.76	0.6	0.88	0.67	0.82	0.96	0.81	0.88	0.8	0.88	0.86	0.84
Confusion	0.79	0.92	0.76	0.85	0.89	0.94	0.86	0.91	0.81	0.96	0.79	0.88
Boredom	0.9	0.92	0.89	0.96	0.98	0.92	0.97	0.95	1	0.97	1	0.98

in the audio modal are both greater than 0.9, and their AUC values in the video image modal are also higher than 0.8. Similar to the confusion matrix, the recognition of boredom is the most stable. Confusion and happiness also showed good performance in audio signals and video images.

Pupil diameter is a very important index in eye movement signal, so we further analyze the original data of pupil diameter. Fig. 8 shows the comparison of pupil diameter in the state pairs of confusion–interest, happiness–interest, and confusion–boredom. The pupil diameter data of each frame in the emotional state pair are put into a coordinate axis with the pupil diameter of the left eye as the vertical axis and the pupil diameter of the right eye as the horizontal axis. The overlap of the two colors in the figure is the overlap of the pupil diameters of the two emotions. As can be seen from Fig. 8(a) and (b), the pupil diameter distribution of the interest state overlaps considerably with the confusion and happiness states. This is the same as the performance of the confusion matrix, where interest is easily misidentified as happiness and confusion.

According to Fig. 8(a)–(c), the pupil diameter distribution under boredom has less overlap with the pupil diameter distribution of the other three emotions. This is the same as the previous analysis results, and boredom has a good performance in each evaluation index.

Both audio signal features and video image features belong to video content. In these two modals, a large part of "interest" is misidentified as "happiness." In terms of stimulus material, a video that makes people feel happy will usually also be interesting, but a video that makes people feel interested will not necessarily make people feel happy and may be confusing. The performance in the confusion matrix also confirms this point. The emotion of interest is most easily misidentified as happiness, followed by confusion, while only a small part of happiness is misidentified as interest.

3) Computational Complexity Analysis: The recognition accuracy of the deep learning model is significantly better than that of the machine learning model. Therefore, we analyze the computational complexity of the deep learning model. We compared floating-point operations (FLOPs), memory access cost (MAC), and the number of parameters (NP) between the FE-CNN+EC-CNN model and the ResNet18 model. FLOPs are the number of FLOPs in one training turn, which is used to measure the time complexity of the model. One MFLOP equals one million FLOPs. MAC represents the memory usage, which is used to evaluate the memory usage of the model at runtime. NP represents the total NP inside the model, which is used to measure the size of the model. The three metrics are calculated using the open-source library ptflops.

TABLE VII

COMPUTATIONAL COMPLEXITY ANALYSIS OF THE MODELS

Convolution layer	MFLOPs	MAC(M)	NP(M)
ResNet18	8.36	15.05	5.02
FE-CNN+EC-CNN	3.87	6.25	3.85

As shown in Table VII, our model is better than ResNet18 in time complexity, memory usage, and NP.

B. Multimodal Fusion Emotion Recognition

The multimodal fusion method is the core of multimodal emotion recognition. The common fusion methods include feature-level fusion, decision-level fusion, and model-level fusion [34]. We studied three fusion methods to determine the best fusion model. Fig. 9 shows three fusion methods, and IDSF represents the process of integration of deep and shallow features in Fig. 5.

1) Feature-Level Fusion: The feature-level fusion is shown in Fig. 9(a). After feature extraction, the principal components of the three modes are obtained. By concatenating the principal components of the three modals, $0 = [\text{PCA1}_e, \ldots, \text{PCA19}_e, \text{PCA1}_a, \ldots, \text{PCA19}_a, \text{PCA1}_v, \text{PCA2}_v, \text{PCA3}_v]$. $F0$ is sent into the IDSF module for the fusion of shallow features and deep features, and then, the fused feature vector is sent into the EC-CNN module for classification. Finally, the recognition accuracy of 76.02% is obtained.

2) Decision-Level Fusion: The decision-level fusion is shown in Fig. 9(b). The classification vectors (shape: [1, 4]) obtained from the three modals are weighted and fused according to a certain proportion to obtain the fused classification vectors.

The vectors fc1,fc2, andfc3 output by the full connection (FC) layer of EC-CNN network in the three models are assigned with the different weight value $w1$,$w2$, and $w3$ and fused to get the final classification vector output $= \text{fc1} * w1 + \text{fc2} * w2 + \text{fc3} * w3, (w1 + w2 + w3 = 1)$. After normalization by the softmax function, the final classification results are obtained. Also, the recognition result after fusion is obtained.

$w1$,$w2$, and $w3$ are set empirically and adjusted by experimental results. First, the three modals are fused in pairs, such as in the following equation:

$$\begin{cases} w2 + w3 = 1, \quad w1 = 0, w2, w3 \neq 0 \\ w1 + w3 = 1, \quad w2 = 0, w1, w3 \neq 0 \\ w1 + w2 = 1, \quad w3 = 0, w1, w2 \neq 0. \end{cases} \tag{10}$$

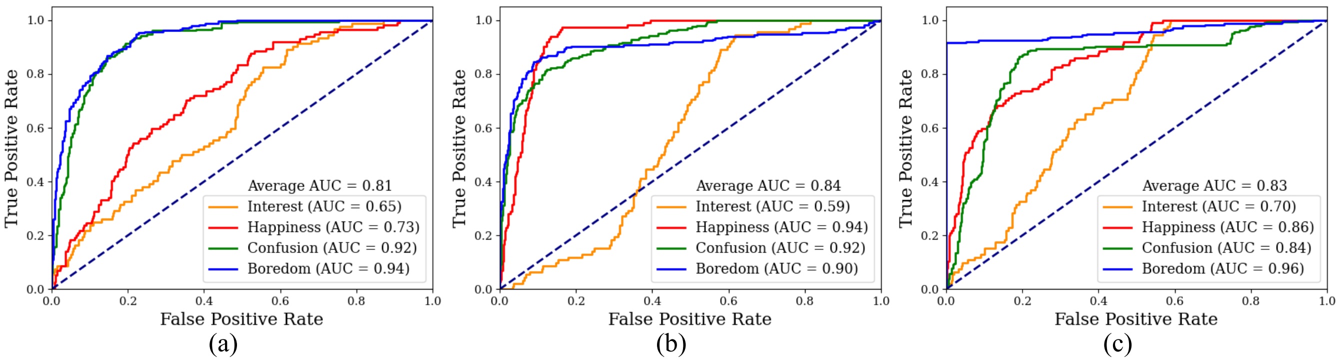


Fig. 7. ROC curve of classification results by using different modal features. (a) ROC curve of eye movement modal. (b) ROC curve of audio signal modal. (c) ROC curve of video image modal.

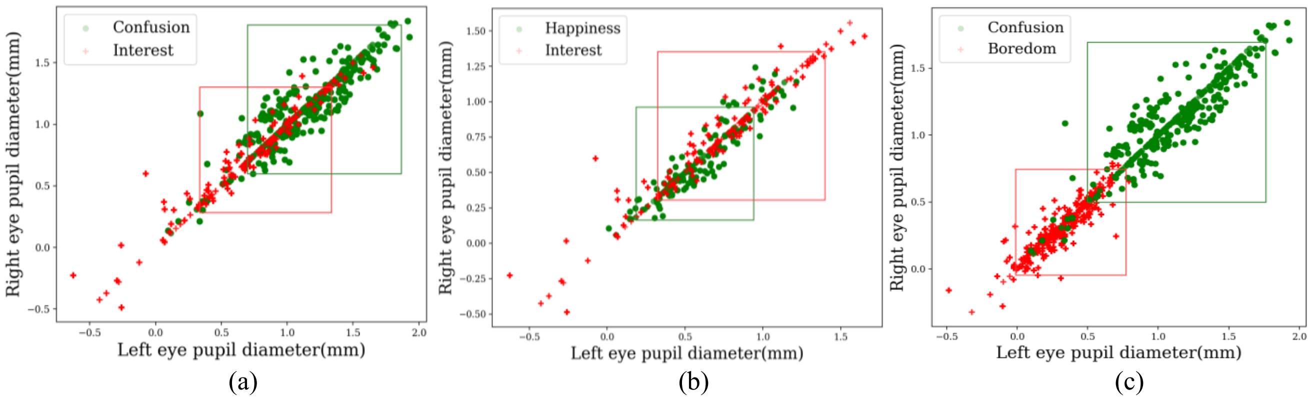


Fig. 8. Comparison of the pupil diameter distribution in different emotion pairs: (a) interest and confusion, (b) interest and happiness, and (c) confusion and boredom.

TABLE VIII

EMOTIONAL STATES RECOGNITION ACCURACY OF SINGLE MODAL/BIMODAL

	Eye movement signal parameters(w1)	Audio signal parameters(w2)	Video image parameters(w3)	Recognition （%）
Model 1	1	0	0	64.32
Model 2	0	1	0	74.67
Model 3	0	0	1	71.88
Model 4	0.25	0	0.75	75
Model 5	0.25	0.75	0	80.68
Model 6	0	0.7	0.3	80.22

The recognition accuracy of the fusion model in Table VIII shows that the recognition effect of the fusion by two modals is higher than that of the single modal, which proves that the three modals all have certain complementarity. The decision-level fusion between eye movement and audio achieved the highest recognition accuracy of 80.68%.

According to Models 4 and 5 in Table VIII, when the eye movement features are combined with the audio and video images of the learning scenario features, the recognition accuracy of the learned emotion is increased by 10.68% and 16.36%, respectively. This indicates that the combination of scenario features can effectively improve the accuracy of emotion recognition in MOOC scenarios.

Then, the weights of the three modals in weighted decision fusion are considered. The weights of the modals of eye movement, audio fusion, and video image are determined by

$$\begin{cases} w1 = 0.25 * w4 \\ w2 = 0.75 * w4 \\ w3 + w4 = 1. \end{cases} \tag{11}$$

In Formula (11), $w1$ is the weight of the eye movement model, $w2$ is the weight of the audio model, $w3$ is the weight of the video model, and $w4$ is the new weight of the eye movement and audio fusion model. The experimental results show that when $w4 = 0.9$, $w3 = 0.1$, combined in the optimal ration in Table VIII, namely, the decision-level fusion with (w1 = 0.225, w2 = 0.675, and w3 = 0.1) achieves the optimal recognition accuracy of 81.90%.

3) Model-Level Fusion: The model-level fusion is shown in Fig. 9(c). The features of the three modals are processed by the corresponding IDSF module to get the feature vectors. Then, the feature vectors are sent to the corresponding EC-CNN module, which is without an FC layer, and the vectors $D1$, $D2$, and D3 containing classification information are obtained. The complete classification information vector D is obtained by concatenating $D1$, $D2$, and $D3$. Then, the local information in vector D is captured through a layer of convolution. The input dimension and output dimension of the convolution are both 1 and kernal_size is 3. The output vector from CNN is sent to an FC layer for final emotion recognition, and the recognition accuracy is 80.16%.

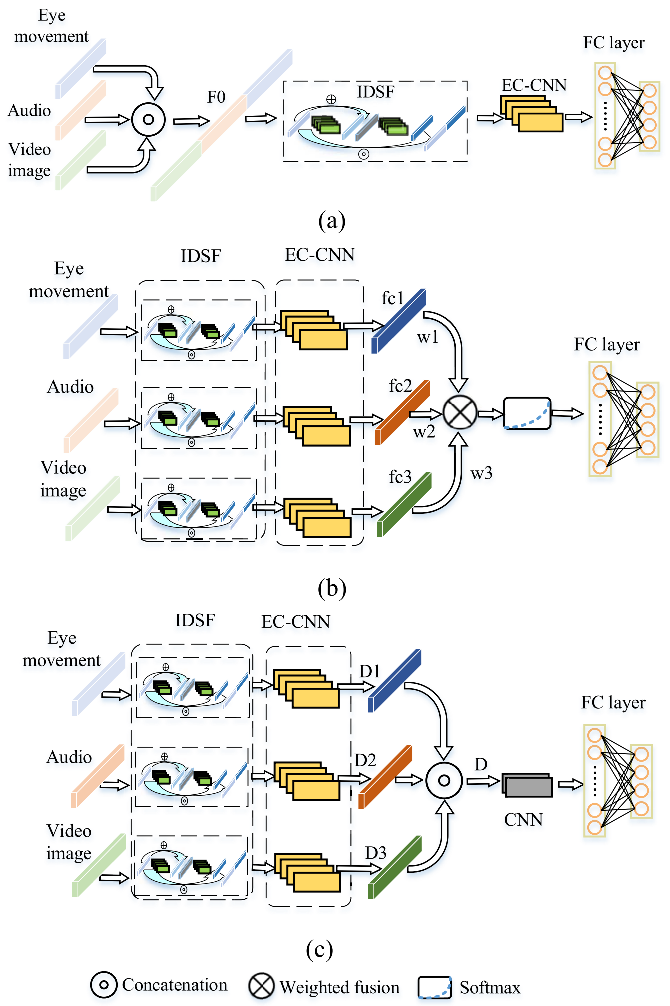


Fig. 9. Flowchart of different multimodal fusion methods: (a) feature-level fusion, (b) decision-level fusion, and (c) model-level fusion.

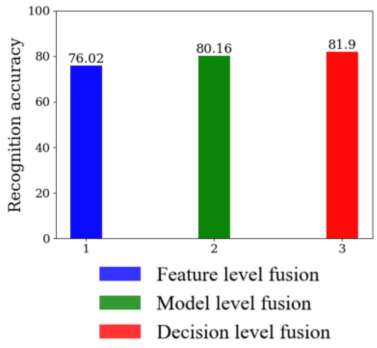


Fig. 10. Recognition accuracy of feature-level fusion, model-level fusion, and decision-level fusion.

Fig. 10 shows the recognition accuracy of the three fusion methods. It can be seen from the figure that the recognition accuracy of feature-level fusion is the lowest among the three methods, indicating that the feature-level fusion method cannot make full use of the complementary information in the three modals. It is more suitable to use the decision-level fusion method.

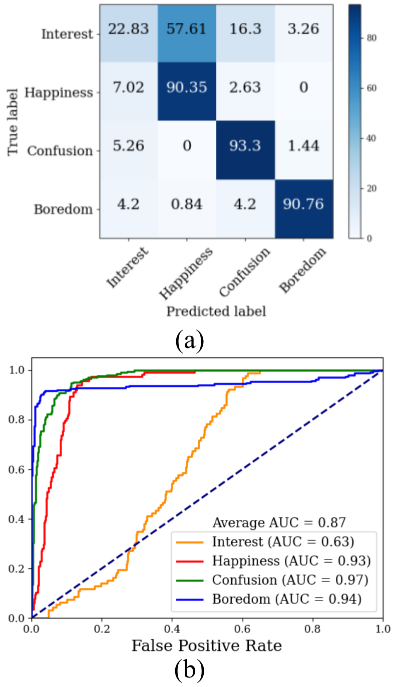


Fig. 11. Confusion matrix and ROC curve of decision-level fusion model: (a) confusion matrix and (b) ROC curve of decision-level fusion.

C. Discussion

In this section, we discuss the impact of different modals and components on recognition results.

The experimental results show that, compared with the single modal, the effect of the fusion of any two modals is better, so the fusion method is effective. By the single modal emotion recognition with each of the three modals, the recognition accuracy of interest is the lowest, and the recognition accuracy of happiness is average.

After the decision-level fusion, the recognition accuracy of the model on happiness has been significantly improved. The confusion matrix and ROC curve are shown in Fig. 11. In addition, it can be seen from Fig. 11(b) that the AUC is 0.87 in the fused model. Compared with the single modal (AUC is 0.81 in the eye movement modal, AUC is 0.83 in the video image modal, and AUC is 0.84 in the audio signal modal in Fig. 7), the stability and generalization ability of the fused model is better. However, the interest recognition effect of the fused model is still poor, which proves that there are limitations to recognize the emotion of interest by the three modals.

In order to evaluate the effectiveness and necessity of FCDE features, FE-CNN model, and EC-CNN model, a group of ablation experiments are designed.

The experimental results are shown in Table IX. The configuration for each experiment is described as follows.

1) *ALL:* Complete configuration of the optimal model, including eye movement features (including FCDE), audio and video features, IDSF module (deep feature and shallow feature fusion twice), and EC-CNN classification module.

TABLE IX
RESULTS OF ABLATION EXPERIMENT OF THE MODEL

Experiment number	Model	Recognition accuracy (%)
1	All/FCDE	81.2
2	ALL + FE-CNN	79.15
3	All/IDSF	75.26
4	All/FE-CNN	78.47
5	EC-CNN/SVM	73
6	EC-CNN/ResNet18	74.58
7	All	**81.9**

2) *All/FCDE:* Remove FCDE features from the eye movement feature.
3) *ALL + FE-CNN:* Based on the IDSF module, add an FE-CNN module to extract the deep features again.
4) *All/IDSF:* Remove the IDSF module.
5) *All/FE-CNN:* Remove the second FE-CNN module.
6) *EC-CNN/SVM:* EC-CNN is replaced by SVM.
7) *EC-CNN/ResNet18:* EC-CNN is replaced by ResNet18.

Table IX lists the ablation experimental results for each experiment. It can be seen from experiments 1 and 7, the recognition accuracy of input data containing FCDE features is better than that of input data without FCDE features. After adding FCDE features, the accuracy of model is improved by 0.7%. From experiments 2, 3, 4, and 7, it can be seen that the model with two FE-CNN modules can fully extract the deep features and fully integrate the deep features with the shallow features, and it achieves higher accuracy than the models with zero, one, or three FE-CNN modules. Experiments 5–7 show that the EC-CNN model is more effective in emotion classification than some baseline machine learning methods.

The experimental results in Table VIII show that the emotion recognition effect of modal fusion is better than single modal. The ablation experiment in Table IX shows that all the modules, including the FCDE features, FE-CNN, and EC-CNN modules.

VI. CONCLUSION

In this article, through the single modal experiment, it is found that eye movement signal, audio signal, and video image are all suitable for identifying learning emotion in MOOC learning scenarios, and the best performing modal is the audio signal. Among the four learning emotions, interest is the most difficult one to identify. In the multimodal fusion experiment, the fusion effect of feature level is not ideal, while the fusion of decision level can make better use of the complementary information of the three modals, achieving 81.90% recognition accuracy. The fusion model has a strong ability to distinguish between confusion, boredom, and happiness.

Future work can focus on the following aspects. First, in the data collection stage, more stimulus materials, which can induce the state of interest, should be added to improve the sample quality of category interest. Second, more modals, which can be obtained by noninterventional means, such as micro-expression and photo plethysmo graph (PPG) signals, can be adopted for emotion recognition. Third, the video semantic information and learners' cognitive state can also be combined to further analyze emotion recognition, especially in learning scenarios.

VII. ACKNOWLEDGMENT

The authors would like to thank Beijing Jinfa Company for providing the data acquisition platform and all the subjects who participated in the data collection experiment.

REFERENCES

[1] J. Riel and K. Lawless, "Developments in MOOC technologies and participation since 2012: Changes since the year of the MOOC," in *Encyclopedia of Information Science and Technology*, M. Khosrow-Pour, Ed., 4th ed. Hershey, PA, USA: IGI Global, 2017, pp. 1–10.

[2] X. Yang, T. Zhang, and C. Xu, "Text2Video: An end-to-end learning framework for expressing text with videos," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2360–2370, Sep. 2018.

[3] S. Zhao, X. Huang, and X. Lu, "A study on the prediction of emotional index to the achievement of MOOC students," *China Univ. Teaching*, vol. 5, no. 5, pp. 66–71, 2019.

[4] C. H. Ik, "The relationship between emotion and academic achievement: The mediation effect of emotion regulation and learning strategy," *Korean J. Child Educ.*, vol. 22, no. 1, pp. 313–324, 2013.

[5] M. Li et al., "Method of depression classification based on behavioral and physiological signals of eye movement," *Complexity*, vol. 2020, pp. 1–9, Jan. 2020.

[6] P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak, "Eye-tracking analysis for emotion recognition," *Comput. Intell. Neurosci.*, vol. 2020, pp. 1–13, Sep. 2020.

[7] T. Li, J. Yang, D. Bai, and Y. Wang, "A new directional intention identification approach for intelligent wheelchair based on fusion of EOG signal and eye movement signal," in *Proc. IEEE Int. Conf. Intell. Saf. Robot. (ISR)*, Aug. 2018, pp. 470–474.

[8] X. Zhang et al., "Fusing of electroencephalogram and eye movement with group sparse canonical correlation analysis for anxiety detection," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 958–971, Apr./Jun. 2022.

[9] L. Gan et al., "A cross-culture study on multimodal emotion recognition using deep learning," in *Proc. Int. Conf. Neural Inf. Process.*, 2019, pp. 670–680.

[10] L.-M. Zhao, R. Li, W.-L. Zheng, and B.-L. Lu, "Classification of five emotions from EEG and eye movement signals: Complementary representation properties," in *Proc. 9th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, Mar. 2019, pp. 611–614.

[11] T.-H. Li, W. Liu, W.-L. Zheng, and B.-L. Lu, "Classification of five emotions from EEG and eye movement signals: Discrimination ability and stability over time," in *Proc. 9th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, Mar. 2019, pp. 607–610.

[12] J.-J. Guo, R. Zhou, L.-M. Zhao, and B.-L. Lu, "Multimodal emotion recognition from eye image, eye movement and EEG using deep neural networks," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 3071–3074.

[13] M. S. Likitha, S. R. R. Gupta, K. Hasitha, and A. U. Raju, "Speech based human emotion recognition using MFCC," in *Proc. Int. Conf. Wireless Commun., Signal Process. Netw. (WiSPNET)*, Mar. 2017, pp. 2257–2260.

[14] H. Zheng and Y. Yang, "An improved speech emotion recognition algorithm based on deep belief network," in *Proc. IEEE Int. Conf. Power, Intell. Comput. Syst. (ICPICS)*, Jul. 2019, pp. 493–497.

[15] J. Wang and Z. Han, "Research on speech emotion recognition technology based on deep and shallow neural network," in *Proc. Chin. Control Conf. (CCC)*, Jul. 2019, pp. 3555–3558.

[16] Q. Mao, Q. Zhu, Q. Rao, H. Jia, and S. Luo, "Learning hierarchical emotion context for continuous dimensional emotion recognition from video sequences," *IEEE Access*, vol. 7, pp. 62894–62903, 2019.

[17] B. Xing et al., "Exploiting EEG signals and audiovisual feature fusion for video emotion recognition," *IEEE Access*, vol. 7, pp. 59844–59861, 2019.

[18] S. Zhang et al., "Learning affective features with a hybrid deep model for audio–visual emotion recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 3030–3043, Oct. 2018.

[19] C. Zhang, J.-N. Chi, Z.-H. Zhang, and Z.-L. Wang, "The research on eye tracking for gaze tracking system," *Acta Autom. Sinica*, vol. 36, no. 8, pp. 1051–1061, Sep. 2010.

[20] X. Zhang et al., "Fatigue detection with covariance manifolds of electroencephalography in transportation industry," *IEEE Trans. Ind. Informat.*, vol. 17, no. 5, pp. 3497–3507, May 2021.

[21] X. Zhang et al., "Individual similarity guided transfer modeling for EEG-based emotion recognition," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2019, pp. 1156–1161.

[22] B. Hu, J. Shen, L. Zhu, Q. Dong, H. Cai, and K. Qian, "Fundamentals of computational psychophysiology: Theory and methodology," *IEEE Trans. Computat. Social Syst.*, vol. 9, no. 2, pp. 349–355, Apr. 2022.

[23] H. Chen et al., "Personal-Zscore: Eliminating individual difference for eeg-based cross-subject emotion recognition," *IEEE Trans. Affect. Comput.*, early access, Dec. 23, 2021, doi: 10.1109/TAFFC.2021.3137857.

[24] W. Wang et al., "Human–computer interaction: Intention recognition based on EEG and eye tracking," *Acta Aeronaut. Astronaut. Sinica*, vol. 42, no. 2, pp. 324290–324301, 2021.

[25] S. Baker and I. Matthews, "Lucas–Kanade 20 years on: A unifying framework," *Int. J. Comput. Vis.*, vol. 56, no. 3, pp. 221–255, 2004.

[26] Y. Zhu et al., "Research on feature selection method in speech emotion recognition," *J. Appl. Acoust.*, vol. 39, no. 2, pp. 216–222, 2020.

[27] S. R. Bandela and T. K. Kumar, "Stressed speech emotion recognition using feature fusion of Teager energy operator and MFCC," in *Proc. 8th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2017, pp. 1–5.

[28] J. Shen, S. Zhao, Y. Yao, Y. Wang, and L. Feng, "A novel depression detection method based on pervasive EEG and EEG splitting criterion," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2017, pp. 1879–1886.

[29] Y. Wang, "Research and application on emotion recognition algorithm based on multi-modal eye movement information," M.S. thesis, School Comput. Sci. Technol., Anhui Univ., Hefei, China, 2019. [Online]. Available: https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201902&filename=1019130693.nh

[30] J. Zhang, X. Wang, and W. Jing, "Speech emotion recognition from spectrograms with deep convolutional neural network," *J. Changchun Univ. Sci. Technol.*, vol. 2020, no. 1, pp. 76–81, 2020.

[31] Y. Gao, Y. Cui, and C. Sun, "Speech emotion recognition method of small sample based on generative adversarial networks," *Comput. Eng. Des.*, vol. 2020, no. 12, pp. 3550–3556, 2020.

[32] J. Shen, X. Zhang, B. Hu, G. Wang, and Z. Ding, "An improved empirical mode decomposition of electroencephalogram signals for depression detection," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 262–271, Jan./Mar. 2022.

[33] J. Shen et al., "An optimal channel selection for EEG-based depression detection via kernel-target alignment," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 7, pp. 2545–2556, Jul. 2021.

[34] H. Cai, Z. Qu, Z. Li, Y. Zhang, X. Hu, and B. Hu, "Feature-level fusion approaches based on multimodal EEG data for depression recognition," *Inf. Fusion*, vol. 59, pp. 127–138, Jul. 2020.