Computer Systems
Science and Engineering

Tech Science Press

DOI: 10.32604/csse.2023.039186
Article

A Method of Multimodal Emotion Recognition in Video Learning Based on Knowledge Enhancement

Hanmin Ye[1,2], Yinghui Zhou[1] and Xiaomei Tao[3,*]

[1]School of Information Science and Engineering, Guilin University of Technology, Guilin, 541004, China
[2]Guangxi Key Laboratory of Embedded Technology and Intelligent System, Guilin University of Technology, Guilin, 541004, China
[3]Guangxi Key Lab of Multi-Source Information Mining and Security, Guangxi Normal University, Guilin, 541004, China
*Corresponding Author: Xiaomei Tao. Email: txm800505@163.com

Received: 13 January 2023; Accepted: 20 March 2023; Published: 28 July 2023

Abstract: With the popularity of online learning and due to the significant influence of emotion on the learning effect, more and more researches focus on emotion recognition in online learning. Most of the current research uses the comments of the learning platform or the learner's expression for emotion recognition. The research data on other modalities are scarce. Most of the studies also ignore the impact of instructional videos on learners and the guidance of knowledge on data. Because of the need for other modal research data, we construct a synchronous multimodal data set for analyzing learners' emotional states in online learning scenarios. The data set recorded the eye movement data and photoplethysmography (PPG) signals of 68 subjects and the instructional video they watched. For the problem of ignoring the instructional videos on learners and ignoring the knowledge, a multimodal emotion recognition method in video learning based on knowledge enhancement is proposed. This method uses the knowledge-based features extracted from instructional videos, such as brightness, hue, saturation, the videos' click-through rate, and emotion generation time, to guide the emotion recognition process of physiological signals. This method uses Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to extract deeper emotional representation and spatiotemporal information from shallow features. The model uses multi-head attention (MHA) mechanism to obtain critical information in the extracted deep features. Then, Temporal Convolutional Network (TCN) is used to learn the information in the deep features and knowledge-based features. Knowledge-based features are used to supplement and enhance the deep features of physiological signals. Finally, the fully connected layer is used for emotion recognition, and the recognition accuracy reaches 97.51%. Compared with two recent researches, the accuracy improved by 8.57% and 2.11%, respectively. On the four public data sets, our proposed method also achieves better results compared with the two recent researches. The experiment results show that the proposed multimodal emotion recognition method based on knowledge enhancement has good performance and robustness.

Keywords: Emotion recognition; video learning; physiological signal; knowledge enhancement; deep learning; CNN; LSTM; TCN

1 Introduction

In recent years, Internet technology and 5G technology have developed rapidly. With a new round of technological innovation, online learning has become a learning mode commonly used by learners of different ages. Compared with the traditional learning mode, online learning breaks the time and space constraints, provides a possibility for lifelong learning [1], and provides a way to share learning resources, so learners of different regions and ages can obtain excellent learning resources. Although online learning has many advantages that the traditional learning mode does not have, some problems must be addressed. In online learning, especially in video learning, learners cannot feel the emotional feedback and cognitive support from teachers in time, so there will be a "lack of emotion" problem. Relevant research shows that emotion plays a crucial role in decision-making, perception, and learning, and positive emotions can improve learners' understanding [2]. Suppose we can identify the learners' emotional state in real-time during the learning process, provide them with corresponding emotional feedback and cognitive support, and transform their negative emotions into positive ones. In that case, we can effectively promote the learners' learning outcomes.

Currently, most research on online learning emotion recognition uses comments of the learning platform or expressions to study learners' emotions in online learning [3–5]. However, the use of learner comments can only obtain learners' emotions after the completion of a stage of learning. It cannot identify the real-time emotion of learners in the learning process and cannot intervene according to the real-time emotion of learners in the learning process [6]. The expression can be artificially covered up or hidden, which cannot objectively reflect the learners' emotions. Physiological signals can be collected in real-time and emotion recognition can be performed in real-time. Physiological signals cannot be controlled artificially and are more objective. However, most data sets in online learning scenes are composed of expressions or learner comments, and the emotional data sets of physiological signals are less. Therefore, we conducted data collection experiments and selected eye movement and PPG physiological signals. These two physiological signals can fully reflect learners' emotional states, and eye movement data can be collected in a non-contact way. Although PPG signals are collected using ear clip sensors in this paper, some studies can use cameras to collect Remote Photoplethysmography (rPPG) signals so that PPG signals can also be collected in a non-contact way. These two kinds of non-contact collected physiological signals can minimize the impact on learners by fully reflecting learners' emotional states.

The processing of temporal data is an essential issue in Affective Computing. The text, speech, and physiological signals commonly used in Affective Computing all contain time series. Currently, most of the research uses raw data or extracted features as input when using the deep learning method to classify the emotion of temporal data. Mustaqeem et al. [7] input raw speech signals into ConvLSTM and GRU for feature extraction and emotion recognition. Du et al. [8] use EEG features extracted from different channels for emotion classification. However, using raw data as input will cause much irrelevant information to be input into the network. Using features as input will lose the temporal nature of the data itself. In addition to temporal data processing, selecting research data in video learning scenes is also very important. In the video learning scene, most studies use the information displayed by learners to identify their emotions, but learners' emotions are generated by the stimulation of instructional videos. Therefore, the video content that learners watch, the audio they hear, and the video's hue will impact learners' emotions. These factors can also be added to the emotion recognition task as knowledge. Because of the above problems, this paper proposes a construction method based on continuous multi-window multimodal temporal features and a method of multimodal emotion recognition in video learning based on knowledge enhancement. The main contributions of this paper are as follows:

- In this study, video features of instructional videos watched by learners during their learning process are used as prior knowledge to assist emotion classification tasks in emotion classification modeling. Specific knowledge includes: learning video features such as saturation, hue, brightness, the click-through rate of instructional videos, relative time, and absolute time of emotion generation.
- Aiming at the temporal characteristics of emotional data, this paper proposes a construction method based on continuous multi-window multimodal temporal features (MMTF). Based on dividing the time window of eye movement data and PPG data, eye movement features and PPG features in each time window are extracted. The feature values in consecutive multiple time windows are used as a temporal feature input that integrates the relationship between features and time. The experiment result shows that compared with the commonly used feature input or raw temporal data input, our proposed continuous multi-window multimodal temporal feature not only extracts the emotion-related representation in the raw data but also preserves the temporal relationship between the raw data and achieves better performance in the emotion recognition task.
- According to the physiological signal time series data and knowledge characteristics used in this paper, a novel method of multimodal emotion recognition in video learning based on knowledge enhancement is proposed. This method uses CNN and LSTM, which can extract spatiotemporal information and deep features, to further extract emotional information in physiological signals, uses attention mechanism to extract key information in deep representations. After the extraction of spatiotemporal information and key information, the TCN that can also extract spatiotemporal information is used to extract spatiotemporal information in knowledge-based features and integrate knowledge-based features with physiological signal features. The information in knowledge-based features is used to enhance the emotional recognition process of physiological signals. The experimental results demonstrate that our proposed method extracts spatiotemporal and key information, effectively integrates knowledge-based features and physiological signals, and the addition of knowledge information also makes the emotion recognition task get better results.

2 Related Works

2.1 Online Learning Emotion Recognition

With the popularity of online learning, more and more researchers have paid attention to the problem of "lack of emotion" in online learning, and more and more researchers have researched online learning emotion recognition. Emotion classification in online learning mainly collects unimodal or multimodal data and classifies emotion states using machine learning or deep learning methods after feature extraction.

Using the comments of the learning platform to identify learners' emotions is one of the most widely used methods in online learning emotion recognition research. Many researchers use learners' comments to identify learners' emotional states. For example, Li et al. [9] proposed a shallow BERT-CNN model consisting of a shallow pre-trained BERT, convolution layer, and self-attention pooling module to analyze the emotion of MOOC comments, with an accuracy rate of 81.3%. As the most commonly used modal in Affective Computing, the facial expression is also the modal researchers will use in online learning emotion recognition research. Bian et al. [10] established a spontaneous facial expression dataset in an online learning environment and proposed an adaptive data enhancement method based on a spatial transformer network. This method can retain discriminative regions in facial images and ignore regions unrelated to emotions. The pre-trained VGG16 was used to classify the emotion of facial expression datasets enhanced by adaptive data and random data, and the classification accuracy reached 91.6%.

However, since the learner's comments can only be obtained after the end of a stage of learning, it is impossible to receive relevant information in real-time during the learner's learning process, and it is also impossible to conduct a real-time intervention on learners. Although facial expressions can be obtained in real-time, learners may have little expression changes during the learning process, leading to unsatisfactory recognition results. Compared with text and expression modals, physiological signals can be obtained in real-time by using sensors, and physiological signals are controlled by the autonomic nervous system and endocrine system without human control, which can more objectively reflect learners' emotions [11]. Researchers began to study emotional recognition using physiological signals. Ullah et al. [12] used EEG for emotion recognition, proposed a sparse discriminant ensemble learning algorithm for selecting the most discriminative signal subset, and used SVM to classify EEG signals. For emotional recognition research using physiological signals, sensors are first necessary to collect physiological signal data. Different types of sensors can be used to collect different physiological signal data. Liu et al. [13] use a blood oxygen sensor to collect PPG signals, a 5-lead cable sensor to collect ECG tracks, and a blood pressure cuff to measure noninvasive blood pressure. Chanel et al. [14] collected subjects' EEG, GSR, blood pressure, heart rate, respiration, and body temperature with an EEG cap, a GSR sensor, a plethysmograph, a respiration belt, and a temperature sensor. In online learning emotion recognition research, researchers also used sensors to collect physiological signals for emotion recognition research. Luo [15] used an iWatch bracelet and camera to collect heart rate and facial video data of primary and secondary school students in the process of emotion generation through emotion induction and used a parallel way of self-report and expert annotation to label data, including pleasure, focus, confusion, and boredom. After preprocessing the dataset, the pre-trained model based on ResNet18 was used for transfer learning, the frame attention network was used to learn emotion features, and the online learning emotion recognition model was trained. The recognition accuracy of the trained model for the four cognitive emotion states reached 87.804%.

In summary, online learning emotion recognition researches mostly use text and facial expression, and fewer use physiological signals. However, physiological signals are more real-time in acquisition than learners' comments on the platform. Compared with facial expressions, physiological signals are free from human control and more objective. Therefore, this paper uses an eye tracker and a wearable ear clip sensor to collect learners' eye movement data and PPG data to study learners' emotional states in video learning.

2.2 Knowledge Enhancement

At present, most emotion recognition studies only consider the relevant data of subjects when using data such as expression [16], speech [17], gesture [18], text [19], physiological signals [20,21], etc. However, subjects' emotion is stimulated by stimulus materials, which also have a more significant impact on the emotional state of subjects, so stimulus materials may also contain some information and knowledge. Chen et al. [22] analyzed the effect of different types of online instructional video lectures on continuous attention, emotion, cognitive load, and learning performance. The results show that different types of video lectures significantly impact learning effects, continuous attention, and cognitive load. Kim et al. [23] analyzed the influence of different types of multimedia content on learners' cognition. The analysis results showed that learners' cognition of multimedia content was different due to different types of multimedia content, grades, and genders. The lower grades will be more interested in multimedia content, and boys have a higher cognitive level than girls in the video, plain drill, and game multimedia content. Wang et al. [24] summarized the movie's grammar and used the audio and video elements that can stimulate the audience's emotion, such as color, lighting, rhythm, formant, and intensity, as domain knowledge to classify and regression the emotion of video content, and achieved good results. It can be seen that in the video learning scene, instructional videos greatly influence learners, so the collected relevant information of the subjects should be combined with the stimulus materials to analyze the emotional states in the study of emotion recognition in video learning. In addition to collecting physiological signals and extracting features, this paper also uses the brightness, hue, and saturation of stimulus materials, the videos' click-through rate, and the time of emotion generation as knowledge-based features to study learners' emotional states together with physiological signals. We use knowledge-based features to enhance the process of emotional recognition of physiological signals so that the information contained in knowledge-based features and the information contained in physiological signals can complement each other, thus improving the effect of emotional recognition.

3 Proposed Method

We propose a method that combines eye movements, PPG signals, and knowledge-based features extracted from instructional videos to identify learners' emotional states in the learning process. This method uses the new multi-window multimodal temporal features (MMTF) as the model's input and uses our CNN-LSTM-MHA-TCN (CLA-TCN) model to extract deep features and recognize emotions based on knowledge enhancement. We will introduce the method from two aspects: multi-window multimodal temporary feature (MMTF) construction and CLA-TCN model based on knowledge enhancement.

3.1 Multi-Window Multimodal Temporal Feature (MMTF) Construction

For temporal data, the output at the present time is related to the present and previous time input. A person's mood changes not only about current events but may also be influenced by previous events. Therefore, for the emotion recognition task, not only the input at the current moment but also the input at the previous moment should be considered. We examine the input forms of studies that use time series for emotion recognition. Du et al. [8] concatenated the features extracted from EEG signals of different channels as input and fed them into the proposed ATDD-LSTM model for emotion classification. Nie et al. [25] used neural networks to extract text, audio, and video features in the sliding window. They fed the extracted features into the multi-layer LSTM as input for emotion recognition. Xie et al. [26] used openSMILE to extract frame-level speech features in audio and input frame-level speech features into LSTM based on the attention mechanism for speech emotion recognition. Thus, the input forms commonly used by researchers for emotion recognition using time series can be roughly divided into three categories, as shown in Fig. 1: (a) The raw data is directly fed into the neural network; (b) Hand-engineered features are extracted and then fed into a neural network; (c) Researchers use a neural network to extract deep features and feed deep features into neural networks. Due to the limitations of the current input form of temporal data, this paper proposes a new form of input, as shown in Fig. 1 (d). Firstly, eye movement and PPG signals are divided into time windows, and the eye movement features and PPG features within each time window are extracted. The eye movement features and PPG features in each time window are concatenated to obtain the f-dimensional multimodal features in each time window. Suppose the feature in a time window is $W = \left[F_1, F_2, \ldots, F_f\right]$. The feature values of t consecutive time windows are combined to form a multi-window multimodal temporal feature $T = [W_1, W_2, \ldots, W_t]$. The $t \times f$ matrix is a sample, and the following sample is obtained by sliding down a time window every time, and all the obtained MMTF samples are sent to the network as input. This input form ensures the timeliness of the input data, and the temporal feature has higher accuracy and a stronger correlation to the emotion recognition task than the raw data.

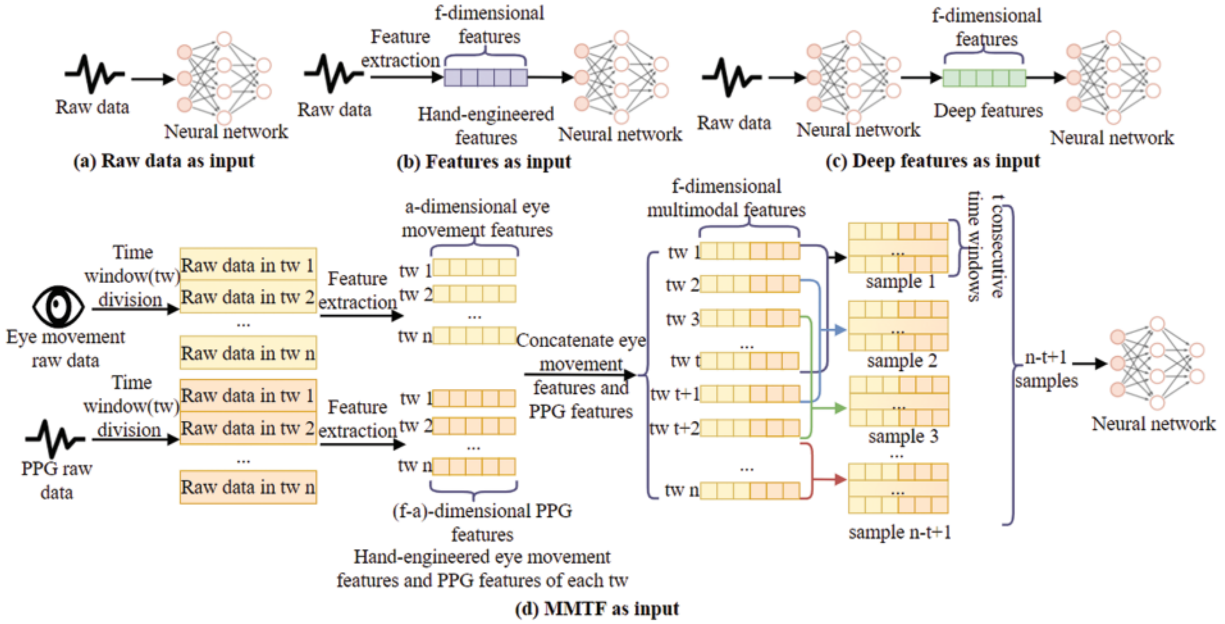


Figure 1: Three common input forms and MMTF

3.2 CLA-TCN Model Based on Knowledge Enhancement

Since the hand-engineered features only contain shallow emotional representations, which may not be sufficient to predict emotions, some researchers use DNN, CNN, and other neural networks to extract deep emotional representations to make up for the deficiency of shallow features. Ma et al. [27] used ResNet50 to extract feature representations from audio and visual data and then used the fusion network to combine the extracted features for emotion prediction. Zhu et al. [28] used LSTM to extract the deep features of three kinds of shallow feature signals: RRI, R peaks amplitude (RAMP), and respiratory signal (EDR). Muhammad et al. [29] used the attention mechanism of the deep BiLSTM network to learn the spatiotemporal features in sequential data and utilized the convolution network with residual blocks to upgrade the features to identify the human behavior in the sequence. Khan et al. [30] used 2-layer CNN to extract multi time-scale features, and then concatenated the multi time-scale features as the input of 2-layer LSTM, and the LSTM model was used to learn the dependencies in the time series. The above studies show that CNN can effectively extract the deep features and spatial information of input data, and LSTM can learn and extract the temporal relationship between input data, which is used in time series research. Therefore, CNN and LSTM are used in this paper to extract deep features, including temporal and spatial information, from the input shallow features.

In selecting knowledge enhancement and classification models, we chose the TCN proposed by Bai et al. [31] to fuse deep features and knowledge-based features and use the fully connected layer for emotion recognition. TCN is mainly composed of the 1D fully-convolutional network (FCN) and causal convolutions. Causal convolution is shown in Fig. 2a. In causal convolution, future data reading can be abandoned in training. The value at time t of the ith layer only depends on the influence of time t of layer $i - 1$ and the values before it, and the following results can be obtained only with the previous causes. However, simple causal convolution still has the problem of traditional convolution neural networks, that is, the modeling length of time series is limited by the size of the convolution kernel, and it is difficult to obtain long-term dependencies. Therefore, TCN uses dilated convolutions instead of simple causal convolutions. Dilated convolution allows for interval sampling of the input during convolution. The receptive field can cover more values in the input sequence through interval sampling. Then we can obtain long-term dependency. To ensure that the TCN can remain stable when the number of layers becomes deeper, Bai et al. replaced the convolution layer with a generic residual block. As shown in Fig. 2b, the residual block contains two layers of dilated causal convolution and non-linearity. WeightNorm and Dropout are added to each layer to regularize the network. The 1×1 convolution layer ensures that the outputs of the lower layer and the upper layer have the same shape when merging. The dilated causal convolution and residual block enable TCN to extract dependencies and causal relationships from long sequences, avoid the problem of exploding/vanishing gradients, and save training time and required memory.

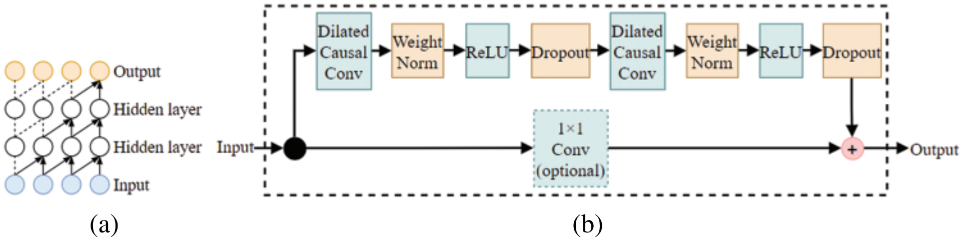


Figure 2: Architectural elements in a TCN, (a) is the visualization of causal convolutional; (b) is the residual block structure

The specific structure of our proposed model is shown in Fig. 3. We input the constructed multi-window multimodal physiological signal temporal features into two-layer CNN and two-layer LSTM to extract the spatiotemporal information and deeper emotional representation of the physiological signal temporal features. We use the multi-head attention mechanism to make the model pay more attention to critical information. The extracted deep temporal features and knowledge-based temporal features are concatenated and sent to TCN, where the knowledge-based temporal features and the deep features of physiological signals are fused with TCN. The information contained in knowledge-based temporal features is used to supplement and enhance the deep features of physiological signals. Finally, the fully connected layer is used to classify the four kinds of emotions.

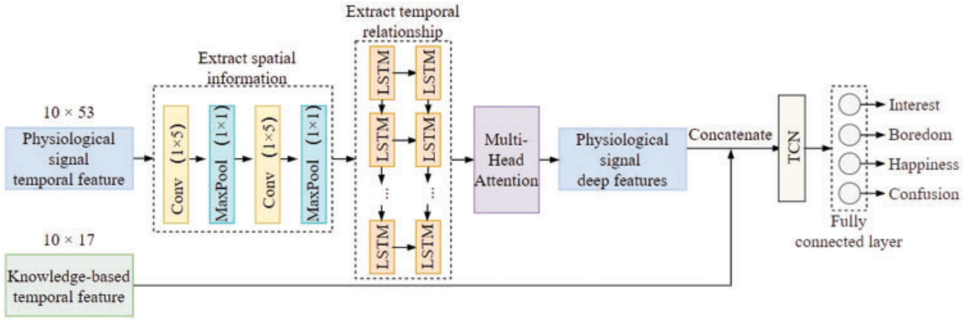

10 × 53
Physiological signal temporal feature
Extract spatial information
Conv (1×5)
MaxPool (1×1)
Conv (1×5)
MaxPool (1×1)
Extract temporal relationship
LSTM
Multi-Head Attention
Physiological signal deep features
Concatenate
TCN
Interest
Boredom
Happiness
Confusion
Fully connected layer
10 × 17
Knowledge-based temporal feature

Figure 3: CLA-TCN model based on knowledge enhancement

4 Experiment

4.1 Data Collection

We have designed an experimental scheme based on video learning to let subjects watch the instructional videos and stimulate subjects to generate emotions. We use the Tobii TX300 table-mounted eye tracker and wearable ear clip sensor to collect the real-time eye movement and PPG data generated by subjects during the learning process. The eye tracker will generate the subjects' fixation, saccade, pupil size, eye movement trajectory, and other data during the experiment, and the ear clip sensor will generate the subjects' PPG signals during the experiment.

Because the experiment is to study the learners' emotions in the video learning scene, before the experiment, the experimenters screened a large number of stimulating materials and selected four instructional videos to stimulate the learners' four academic emotions: interest, boredom, happiness, and confusion.

We recruited 68 participants for the testing procedure. They were all college students with normal vision or correction. Tests were performed with 34 males and 34 females. The experiment was conducted in a quiet room without noise and harassment. The experiment used a desktop computer, a table-mounted eye tracker, and a wearable ear clip sensor. The participants were briefed about the experiment verbally before testing.

Before the experiment, eye calibration was performed after wearing the sensor device. Then the subjects watched the crosshairs on the screen for 30 s to obtain the baseline values of eye movement data and PPG data in the neutral state. During the experiment, four 2 min video clips were played in random order. The subjects watch the video clip on the computer screen. After the video is played, the subjects need to label the emotion generated when watching the video by pressing the key and then watch the next video clip and label the emotion by pressing the key.

After the experiment, participants were introduced to the meaning of arousal and four emotional words: interest, boredom, happiness, and confusion. Then, the subjects needed to watch the instructional videos again and recall the emotional state generated at that time. According to the emotional words in the emotion classification model, they selected their emotional state to label the four videos and scored A1 (weak) to A5 (strong) according to the emotional intensity at that time.

We retained the data of subjects with no missing data and eye movement calibration accuracy and PPG calibration accuracy higher than 70%. Finally, 45 subjects with emotion intensity from A3 to A5 were selected as the data set. The sample number of each emotion in the data set is shown in Table 1.

The data set can be obtained from https://github.com/zhou9794/video-learning-multimodal-emotion-dataset.

Table 1: The sample number of four emotions

Emotion	Sample number
Interest	1451
Boredom	2723
Happiness	1761
Confusion	2275

4.2 Data Preprocessing

We preprocess the multimodal data collected by the sensor. For eye movement data, we removed the lost eye movement data due to blinking, eye closure, lower head, and other reasons during the experiment. We subtracted the average of the corresponding baseline values of the subject's eye movement data from the remaining data to exclude the differences between the subjects. For missing values, use the linear interpolation method to supplement, and the linear interpolation formula is shown in Eq. (1).

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

where (x_0, x_1) is the time point of adjacent frames before and after the missing value, (y_0, y_1) is the corresponding eye movement data value of (x_0, x_1), x is the corresponding time point of the missing eye movement data, and y is the missing eye movement data.

The PPG signal is similar to most physiological signals and belongs to low-frequency signals. The signal frequency band range is 0–20 Hz, and most energy is concentrated within 10 Hz [32]. Due to the low-frequency and weak characteristics of the PPG signal, it is easily affected by power frequency interference, baseline wander, and other noises in the data collection process, so it is necessary to filter and denoise the PPG signal in the preprocessing process. Power frequency interference refers to the noise generated by the distributed capacitance of the human body and the influence of electric and magnetic fields, with a frequency of about 50 Hz. Baseline wander is a low-frequency noise caused by respiratory artifact and motion artifact, with a frequency of about 1 Hz. For this reason, a high-pass filter with a threshold of 1 Hz is set to filter out baseline wander, and a low-pass filter with a threshold of 10 Hz is set to filter out power frequency interference and retain the primary information of PPG. For the filtered PPG signal, the average value of the baseline PPG signal of the corresponding subjects was subtracted to exclude the differences among subjects.

4.3 Feature Extraction

We extracted standard features of eye movement and PPG. For eye movement data, we extracted 29 time-domain features of pupil diameter, fixation, and saccade, and for PPG data, we extracted 42 time-domain features, frequency-domain features, and nonlinear features. Then we calculated the Pearson correlation coefficient between the above features and emotional states, and the calculation formula is shown in Eq. (2).

$$r_{ij} = \frac{\sum_1^n \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_1^n \left(X_i - \overline{X}\right)^2 \sum_1^n \left(Y_i - \overline{Y}\right)^2}} \tag{2}$$

where X and Y are samples for which correlation coefficients need to be calculated, $\overline{X}$ and $\overline{Y}$ are the mean values of samples X and Y, and n is the number of samples. However, due to random sampling, sample imbalance, and other problems, the correlation coefficient has a certain degree of contingency, so it is necessary to carry out a significance test to find the significance coefficient p. It is generally believed that there is no significant correlation when $p > 0.05$, the two groups of samples are significantly correlated when $p < 0.05$, and the two groups of samples are very significantly correlated when $p < 0.01$.

According to the value of the significant coefficient, we selected 24 eye movement features and 29 PPG features with a significant coefficient less than 0.05, and the final selected features are shown in Table 2.

Table 2: Eye movement features and PPG features

Modality	Feature
Eye movement	Fixation times, saccade times, minimum and mean of fixation velocity and saccade velocity, maximum of fixation velocity, maximum and mean of fixation time, maximum, minimum, mean, standard deviation and variance of left pupil diameter, right pupil diameter and mean pupil diameter
PPG	Maximum, minimum, mean, and median of HR, IBI, and R peaks, and standard deviation of HR, SDSD, NNI_20, PNNI_20, RMSSD, range_NNI, CVSD, CVNNI, lf, hf, vlf, lf_hf_ratio, lfnu, hfnu, total_power, triangular_index, sd1

After extracting eye movement features and PPG features, we analyzed the videos watched by learners, divided the instructional videos into 20 frames per second, calculated the hue, saturation, and brightness of each frame image, calculated the mean, maximum, minimum, variance and standard deviation of hue, saturation, brightness of multiple images in each time window. We also extracted the video's click-through rate on the video website, the time of the learner's emotional generation in the whole experimental process (absolute time), and the time of the learner's emotional generation compared with the current video (relative time). A total of 18 features are used as knowledge-based features. The Pearson correlation coefficient and significance coefficient between 18 knowledge-based features and emotional states are calculated. The hue, saturation, and brightness feature values are calculated as shown in Algorithm 1, and the 17 knowledge-based features are retained according to the significance coefficient, as shown in Table 3.

Algorithm 1: Calculation of hue, saturation, and brightness feature values

Input: Instructional videos, start time of the time window W_b, end time of the time window W_e

Output: Feature values of hue, saturation, and brightness: H, S, Br

```
The instructional video is divided into multiple images at 20 frames per second
for $W_b$ to $W_e$ do
  if $W_b <$ time of the current frame $< W_e$ then
    for i = 0 to the height of the frame do
      for j = 0 to the length of the frame do
        Brp ← $\max(R, G, B)$
        Sp ← $\begin{cases} \frac{Br - \min(R, G, B)}{Br} & \textit{if } Br \neq 0 \\ 0 & \textit{otherwise} \end{cases}$
        Hp ← $\begin{cases} \frac{60(G-B)}{Br - \min(R,G,B)} & \textit{if } Br = R \\ 120 + \frac{60(B-R)}{Br - \min(R,G,B)} & \textit{if } Br = G \\ 240 + \frac{60(R-G)}{Br - \min(R,G,B)} & \textit{if } Br = B \end{cases}$
      end for
    end for
    H ← $mean\left(\sum H_p\right)$
    S ← $mean\left(\sum S_p\right)$
    Br ← $mean\left(\sum Br_p\right)$
  end if
  Calculate the mean, maximum, minimum, variance, and standard deviation of H, S, and Br of multi-frame images in the time window
end for
```

Table 3: Knowledge-based features

Kind	Feature
Hue	Mean, maximum, minimum, variance, standard deviation
Saturation	Mean, maximum, variance, standard deviation
Brightness	Mean, maximum, minimum, variance, standard deviation
Click-through rate	The click-through rate of instructional video
Time	Relative time, absolute time

4.4 Experiment Setup

Our implementation is based on the TensorFlow deep learning framework. We divided the training set, validation set, and test set according to the ratio of 7:1:2 and carried out 5-fold cross-validation. 500 epochs are set to ensure adequate training. The optimization algorithm selects the Adam algorithm. The cross-entropy is selected as the loss function. The initial learning rate is set to 0.001. To prevent the model from overfitting, we use the learning rate decay and early stopping criteria to complete the learning in the network training process: whenever a model performs better than the previous best model in training, the model will be saved. If the model does not get better results in the subsequent five epochs, the learning rate will decay to the original 0.2. If the model does not get better results in the subsequent ten epochs, it is considered to be the model is overfitting, and the training process will be stopped.

4.5 Comparison Experiments and Results

This section analyzes the recognition results of different input forms, the recognition results of multimodal physiological signals fused with knowledge-based features, the recognition results of using the shallow feature, the deep feature, and knowledge enhancement, the recognition results of using MHA, the comparison between the proposed model and other models, and the comparison on public data sets.

4.5.1 Analysis of Recognition Results in Different Input Sequence Lengths

We studied and compared the amount of time information contained in different input sequence lengths and the impact of different input sequence lengths on recognition accuracy. We respectively tried to send 0.5, 1, and 2 s time windows with different numbers of continuous time windows as input and sent them to the two-layer LSTM network. The results are shown in Fig. 4, where *n* is the continuous number of time windows.

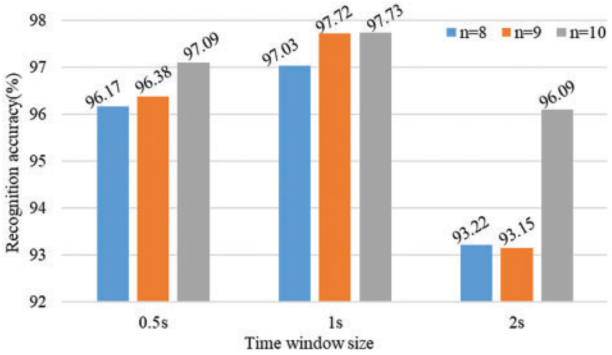
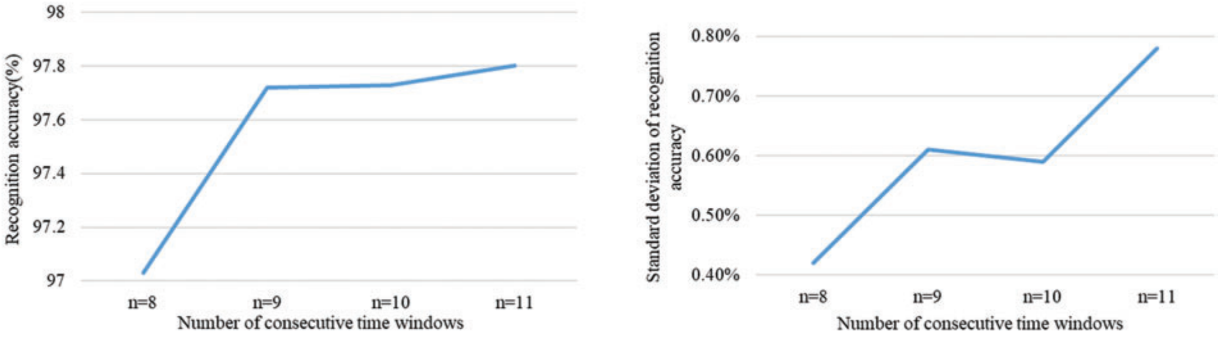


Figure 4: Recognition accuracy of different input sequence lengths

It can be seen from the Fig. 4 that, compared with the 2 s window, the relatively short time windows of 0.5 and 1 s have higher emotion recognition accuracy, and the recognition accuracy of the 1 s time window as input is also higher than that of 0.5 s time window as input. The result may be because we have extracted the features in the time window and taken the features in the continuous time window as the temporal data. The window of 2 s is long, and some temporal relationships of data may be lost in the feature extraction process, while the window of 0.5 s is too small to include changes in IBI and R peak in some cases. The recognition accuracy of the 0.5 s window may be lower than that of the 1 s time window due to the lack of part of PPG signal information. Compared with the 2 and 0.5 s windows, the 1 s time window is a short time to obtain the original time relationship and the PPG information. Therefore, we finally choose 1 s as the time window value for feature extraction.

The recognition accuracy and the standard deviation of accuracy were compared when 1 s windows of different continuous lengths were used as input and fed into LSTM. The results are shown in Fig. 5. The recognition accuracy of 9 ∗ 1, 10 ∗ 1, and 11 ∗ 1 s as input is similar, and the standard deviation of the recognition rate of 10 ∗ 1 s as input is lower. It shows that the model trained with 10 ∗ 1 s as input has better recognition performance and more stability, so 10 ∗ 1 s is selected as the final input form.



(a) (b)

Figure 5: Accuracy and standard deviation of 1 s windows with different continuous lengths as input, (a) is recognition accuracy; (b) is the standard deviation of recognition accuracy of ten-fold cross-validation

4.5.2 Analysis of Recognition Results in Different Input Forms

To compare the effectiveness and superiority of the multi-window multimodal temporary feature input form (MMTF) we proposed, we used our data set and PPG data in DEAP data set and compared the raw data and features of 1 and 10 s as input. We input the data of three input forms into the LSTM network. Because LSTM has advantages in processing temporal data, we use LSTM to compare the emotional recognition results of the three input forms. The result can reflect the amount of temporal information in different input forms. The recognition results are shown in Table 4.

Table 4: Recognition accuracy of MMTF, raw data, and features as input

Data set	Input form	Accuracy
Ours	Feature (1 s)	77.48%
	Raw data (1 s)	72.43%
	Feature (10 s)	65.91%
	Raw data (10 s)	74.65%
	MMTF (10 ∗ 1 s)	**97.38%**
DEAP	Feature (1 s)	34.79%
	Raw data (1 s)	29.14%
	Feature (10 s)	27.60%
	Raw data (10 s)	30.21%
	MMTF (10 ∗ 1 s)	**43.66%**

We compare the features and raw data in 1 and 10 s as input with the 10 ∗ 1 s MMTF as input. We respectively sent different forms of input to LSTM for classification. In our data set, the MMTF input form has a recognition accuracy of 97.38%, which is 19.9% and 31.47% higher than 1 and 10 s feature forms and 24.95% and 22.78% higher than 1 and 10 s raw data input forms. When the input time is 1 s, the recognition accuracy of the feature form input is 5.05% higher than that of the raw data sequence input. When the input time is 10 s, the recognition accuracy of the feature form input is 8.74% lower than that of the raw data sequence input. We obtain similar results in the DEAP data set, where MMTF outperforms 1 and 10 s raw data and feature input, 1 s feature input outperforms 1 s raw data input, and 10 s raw data input outperforms 10 s feature input. It can be seen that when the length of time of the input data is short, although there is a specific temporal relationship between the raw data, the time relationship in a short time does not contain much information related to emotion. After extracting features from the raw data, the information related to emotion is retained, and the information unrelated to emotion is discarded, which improves recognition accuracy. When the length of time of the input data is long, the raw data contains more temporal relationships, which can play a particular auxiliary role in emotion recognition. After extracting features from raw data, due to the long-time window, most of the information will be lost after extracting features, thus leading to the decline of the recognition accuracy of feature input. However, our proposed input form of MMTF is composed of features of multiple continuous windows, whether input short-time data or long-time data. MMTF not only guarantees the time relationship between data to a certain extent but also extracts emotion-related features, effectively integrates the advantages of two common input forms of feature and raw data sequence, and achieves higher recognition accuracy.

4.5.3 Analysis of Multimodal Physiological Signal Recognition Results Based on Knowledge Enhancement

To compare the influence of unimodal and multimodal on recognition accuracy, we input eye movement data, PPG data, and eye movement and PPG data (multimodal) into CLA-TCN, respectively. The recognition accuracy of unimodal and multimodal are shown in Fig. 6.

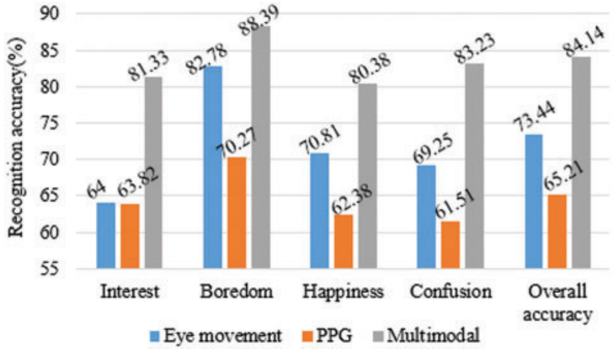


Figure 6: Recognition accuracy of unimodal and multimodal

As can be seen from Fig. 6, multimodal fusion can improve recognition accuracy. Compared with only using PPG data, the recognition accuracy of multimodal fusion is significantly enhanced. The recognition accuracy of the four emotions and the overall recognition accuracy have increased by more than 17%, indicating that the eye movement data contains much information that is not in the PPG data. Eye movement data plays a complementary role in PPG data. Compared with only using eye movement data, the recognition accuracy of multimodal fusion is also improved. The recognition accuracy of four emotions and the overall recognition accuracy increased by 17.33%, 5.61%, 9.57%, 13.98%, and 10.7%, respectively, which indicates that PPG data also has a specific complementary effect on eye movement data. Especially in recognizing interest, happiness, and confusion, the recognition accuracy of the three emotions after integrating eye movement and PPG data is greatly improved. The result may be because eye movement and PPG contain certain information that can distinguish these three emotions, and the information contained in multimodal can complement each other. The multimodal fusion method also effectively integrates the information contained in multimodal, thus improving emotion recognition accuracy. In addition, it can be seen from Fig. 6 that eye movement has a higher recognition accuracy for boredom in recognizing four emotions. The result may be because learners fix their eyes at a point or look outside the screen when bored, which will be reflected in eye movement data. Therefore, compared with the other three emotions, eye movement data may be easier to recognize boredom.

To compare the effects of the shallow feature, the deep feature, and knowledge enhancement on the prediction results, we respectively compared four cases: (1) The hand-engineered shallow features are input into TCN and the fully connected layer for emotion recognition; (2) Two-layer CNN, two-layer LSTM, and multi-head attention mechanisms are used to extract deep features, and the deep features are input into TCN and the fully connected layer for emotion recognition; (3) The fused deep features and knowledge-based features are input into TCN and the fully connected layer for emotion recognition (deep + knowledge); (4) The fused shallow features, deep features, and knowledge-based features are input into TCN and the fully connected layer for emotion recognition (shallow + deep + knowledge). The recognition accuracy of the four cases is shown in Fig. 7.

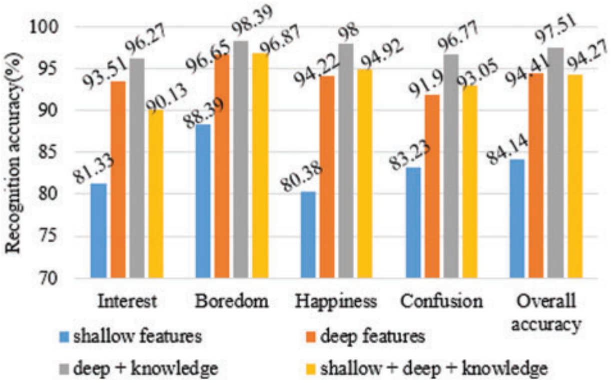
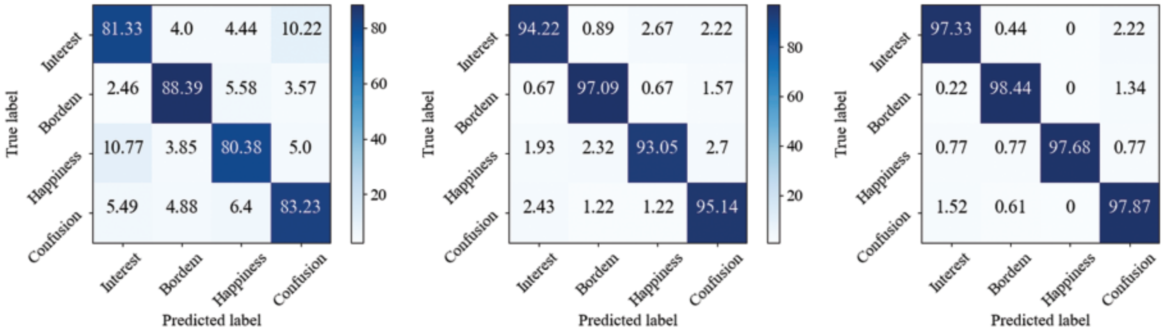


Figure 7: Recognition accuracy of shallow feature, deep feature, and knowledge enhancement

As can be seen from Fig. 7, compared with only using shallow features to recognize emotions, the recognition accuracy of the four emotions and the overall recognition accuracy after extracting deep features has been significantly improved. The recognition accuracy of the four emotions has increased by 12.18%, 8.26%, 13.84%, and 8.67%, respectively. The overall recognition accuracy has increased by 10.27%. The result shows that CNN, LSTM, and multi-head attention mechanisms extract the spatiotemporal information in shallow features. The extracted deep features contain more effective and deeper emotional representations not found in shallow features, which effectively makes up for the deficiency of shallow features. After knowledge enhancement, the recognition accuracy of the four emotions and the overall recognition accuracy have been further improved. The recognition accuracy of the four emotions has increased by 2.76%, 1.74%, 3.78%, and 4.87%, respectively. The overall recognition accuracy has increased by 3.1%. The result indicates that knowledge-based features contain information not included in physiological signals. In addition, the method of knowledge enhancement can effectively integrate physiological signals and knowledge-based features so that the two kinds of features complement each other and then get a better recognition effect.

Moreover, Fig. 7 also shows that the recognition accuracy of the fused shallow features, deep features, and knowledge-based features as input is lower than using the fused deep features and knowledge-based features. The result may be because the shallow features have gone through the neural network to extract deeper features containing more information and less noise, improving the recognition accuracy. The shallow features may include more noise, and the information in the shallow features may overlap with the deep features. Therefore, the shallow features may not bring more practical information to the deep features, but bring noise, which reduces the recognition accuracy.

Fig. 8 shows the confusion matrix of using the shallow feature classification, using the deep feature classification, and using the deep feature and knowledge-based feature classification in an experiment randomly selected in the five rounds of experiments of 5-fold cross-validation. As can be seen from Figs. 8a and 8b, the use of deep features improves the recognition accuracy of the four emotions and the overall recognition accuracy, and the use of deep features reduces the misclassification probability among the four emotions. The result indicates that the deep features extracted by CNN-LSTM-MHA contain more spatiotemporal information and emotional representation. The deep features also enhance the model's ability to distinguish the four kinds of emotions. It can be seen from Figs. 8b and 8c that after using knowledge to enhance the emotion recognition process of physiological signals, the emotion of interest and happiness got higher recognition accuracy improvement. The result may be because the knowledge-based features consist of the hue, brightness, and saturation of the instructional video, the video's click-through rate, and the time of emotion generation. The video that stimulates the learners' happiness and interest may have colorful and changeable pictures. The video that makes the learners bored may have monotonous and dull images. A video with a high click-through rate could also stimulate the learners' interest. Moreover, learners' emotions may also be related to the teaching stage. The introduction stage may be more able to stimulate learners' interest. The longer the learning time is, the more likely learners are to have negative emotions. The knowledge-based features we extracted contain this information, so the recognition accuracy of the two emotions of interest and happiness has been improved significantly. It can be seen from Figs. 8a–8c that boredom achieves the highest recognition accuracy no matter using shallow features, deep features, or deep features and knowledge-based features. The result may be because boredom has the largest number of samples among the four emotions, so the model can learn more information from more samples and better distinguish boredom from other emotions. In addition, it can be seen that when only physiological signal features are used, the recognition accuracy of interest and happiness differs significantly from that of the other two emotions. After using knowledge-based features, the difference between the recognition accuracy of interest and happiness and that of the other two emotions becomes smaller. The result indicates that knowledge-based features contain information not in physiological signals and can complement physiological signals to improve the recognition effect.



(a) (b) (c)

Figure 8: The confusion matrix of shallow feature, deep feature, deep + knowledge-based feature. (a) is the confusion matrix of shallow feature classification, (b) is the confusion matrix of deep feature classification, and (c) is the confusion matrix of deep + knowledge-based feature classification

4.5.4 Analysis of Recognition Results by Adding Multi-Head Attention Mechanism

We compared the effect of using and not using MHA on the recognition results. The results are shown in Table 5. CLA-TCN uses MHA to obtain critical information after CNN-LSTM, and CL-TCN does not use MHA after CNN-LSTM.

Table 5: Comparison of with or without the use of multi-head attention mechanism

Model	Accuracy
CLA-TCN	**97.51%**
CL-TCN	95.40%

It can be seen from Table 5 that after MHA is added to the model, the recognition accuracy of the model is improved by 2.11%, which indicates that MHA takes into account the correlation between deep features and learns the dependency between deep features. MHA enables the network to focus on multiple regions at the same time, expands the network's attention range, and enables the network to focus on both local information and global information.

4.5.5 Compared with Other Model Recognition Results

To prove the superiority of the proposed method, we compare the proposed method with three machine learning methods, three classical deep learning methods, and two recent approaches. The comparison results are shown in Table 6. In the following comparison methods, the Decision Tree [33], SVM [34], and Random Forest [35] all use grid search to find the optimal parameters. The hyper-parameters of ResNet18 [36], VGG16 [37], and Xception [38] are consistent with the hyperparameters of our proposed model. CNN [39] and CNN-LSTM [40], respectively, use the hyperparameters in the corresponding references. All deep learning models use learning rate decay and early stopping criteria to prevent overfitting.

Table 6: Comparison with other models

Model	Accuracy	F1-score	AUC
Decision tree [33]	52.02%	0.6718	0.8780
SVM [34]	58.34%	0.399	0.7573
Random Forest [35]	62.42%	0.5437	0.8092
ResNet18 [36]	85.32%	0.8466	0.9735
VGG16 [37]	94.29%	0.9387	0.9786
Xception [38]	83.20%	0.8100	0.9540
CNN [39]	88.94%	0.8803	0.9744
CNN-LSTM [40]	95.40%	0.9595	0.9805
Ours	**97.51%**	**0.9734**	**0.9971**

As seen from Table 6, our proposed method's accuracy, F1-score, and AUC are better than other methods. Our proposed method achieved the highest recognition accuracy of 97.51%, which is improved by more than 28% compared with the three machine learning methods of Decision Tree, SVM, and Random Forest. Compared with the classical neural networks ResNet, VGG16, and Xception, the recognition accuracy of the proposed method is improved by 12.19%, 3.22%, and 14.78%, respectively. Compared with CNN [39], the recognition accuracy of the model proposed in this paper is improved by 8.57%. Compared with CNN-LSTM [40] model, the recognition accuracy of the model proposed in this paper is improved by 2.11%. The result indicates that our proposed method extracts emotion representations containing more spatiotemporal information from shallow features and pays attention to both local and global information.

In addition, we also compare the parameters and computational complexity of the model. We calculate the total parameters and Floating-point Operations (FLOPs) of the five neural networks compared in Table 6 and the proposed network. FLOPs refer to the number of floating-point operations in one training turn, which is used to measure the complexity of the model. The total parameters and FLOPs index of the five networks is shown in Table 7, where m after the number represents one million FLOPs and k represents one thousand FLOPs.

Table 7: Computational complexity and parameter quantity comparison of the models

Model	FLOPs	Total parameters
ResNet18 [36]	11.18 m	11.18 m
VGG16 [37]	1135.44 m	1,135.42 m
Xception [38]	20.86 m	20.87 m
CNN [39]	8.16 m	8.16 m
CNN-LSTM [40]	474.78 k	464.82 k
Ours	312.98 k	249.12 k

As seen from Table 7, ResNet18, VGG16, and Xception are all neural networks with millions of computations and parameters. CNN [39] has 8.16 m parameters and requires 8.16 m operations in one training turn. CNN-LSTM [40] has 464.82 k parameters and requires 474.78 k operations in one training turn. Our proposed model has 249.12 k parameters and only requires 312.98 k operations in one training turn. Our model has minor total parameters, the smallest amount of operations, the lowest complexity, and the best performance compared with the previous network. This also shows that the model's performance does not necessarily increase with the increase of network layers. Appropriately increasing the number of network layers can improve the network's performance, but a too-deep network may lead to overfitting or a local minimum.

4.5.6 Compared with Other Publicly Data Sets

To compare the generalization ability of our model, we contrast CNN [39], CNN-LSTM [40], and our method on four public datasets: WESAD [41], MAHNOB-HCI [42], DEAP [43], and SEED [44].

The WESAD dataset collects physiological signals such as PPG, ECG, and EDA of 15 subjects through wrist and chest-worn sensors and labels arousal, valence, and three emotions: neutral, stress, and entertainment. The MAHNOB-HCI dataset recorded EEG, eye movement, face and body video, and peripheral physiological signals generated by 27 subjects while viewing images and videos. The data are annotated with emotion keywords, arousal, valence, dominance, and predictability. The DEAP dataset recorded EEG and peripheral physiological signals including PPG from 32 participants while they watched forty 1-min music video clips. The data are labeled with arousal, valence, like/dislike, dominance, and familiarity. The SEED dataset used 15 movie clips to stimulate the subjects' three emotions, positive, neutral, and negative. It contained the EEG and eye movement signals generated by the 15 subjects while watching the movie clips. However, our dataset focuses on video learning scenarios. We select instructional videos as stimulus materials, select eye movement and PPG signals that can be collected in a non-contact method, and collect the physiological signals of 68 subjects. The data are labeled with arousal and four emotions: interest, boredom, happiness, and confusion.

We use two recent approaches in Table 6 to compare with our model. In WESAD, MAHNOB-HCI, and DEAP datasets, we performed four classifications of high/low valence and high/low arousal. In the SEED dataset, we performed three classifications: positive, neutral, and negative. The comparison results are shown in Table 8.

Table 8: Comparative experiments on WESAD, MAHNOB-HCI, DEAP, and SEED

Data sets	Modalities	Accuracy		
		CNN [39]	CNN-LSTM [40]	Ours
WESAD	PPG	74.98%	74.83%	**76.41%**
MAHNOB-HCI	Eye movement, video	47.13%	50.41%	**52.50%**
DEAP	PPG, video	35.27%	41.83%	**46.08%**
SEED	Eye movement, video	70.21%	76.63%	**83.76%**

In the WESAD data set, there is little difference in the accuracy of the three methods, indicating that the three methods have extracted deeper features in PPG. On MAHNOB-HCI, DEAP, and SEED data sets, the accuracy of CNN-LSTM is higher than that of CNN, and the accuracy of our method is higher than that of CNN-LSTM, which indicates that LSTM extracts the temporal information that

CNN does not extract. The TCN we use better integrates the spatiotemporal information and deep representation extracted by CNN and LSTM and obtains a better recognition effect.

The accuracy gap on different data sets may be because different data sets process data differently, and different processing methods also lead to different data set quality. For example, our data set filters the data to remove the data with low calibration accuracy so that the data in our data set has higher quality. The annotation granularity of emotion also affects accuracy. WESAD, MAHNOB-HCI, and DEAP are all classified using the level of valence and arousal, and coarse-grained annotation may lead to many samples being misclassified. In addition, the other three datasets, except SEED, are four-class classifications, and the SEED dataset is three-class classifications, which may be the reason why the SEED dataset has the highest accuracy.

5 Discussion

This paper proposes a construction method based on continuous MMTF and a multimodal video learning emotion recognition CLA-TCN model based on knowledge enhancement.

For the MMTF input form we proposed, we compared it with the commonly used raw data input form and feature input form. The MMTF input form we proposed as the input obtained significantly better results than the other two input forms. This is because the MMTF input form we proposed is to extract the features of physiological signals in a short time window and connects the feature of multiple continuous time windows as input. This form of input not only extracts the emotion-related features, discards the emotion-independent information in raw data, but also preserves the temporal nature of raw data, which makes MMTF more suitable for temporal data. However, the selection of the number of continuous time windows in MMTF may be different due to different data quality, so MMTF can be combined with optimization algorithms to continue to improve the generalization ability of MMTF in the following work.

For our proposed knowledge enhancement method, considering that the stimulation of instructional video generates the emotion of learners in the video learning scene, most of the current research only focuses on the use of learners' external performance for emotion recognition while ignoring the impact of instructional video on learners' emotion. We extract a series of knowledge such as hue, brightness, saturation, the click-through rate of videos, and emotion generation time as knowledge-based features to enhance the process of emotion recognition from physiological signals. Because the stimulation of instructional video generates learners' emotions, the content and color of instructional video will have an intuitive impact on learners' emotions, so after using the method of knowledge enhancement, the recognition accuracy of our model has been significantly improved, which indicates that there is complementary information between knowledge-based features and physiological signal features, and the addition of knowledge-based features can enhance the ability of the model to distinguish between the four emotions. In addition to significantly improving the accuracy of emotion recognition, the method of knowledge enhancement can also ensure that the model can still have a specific recognition ability even when the physiological signal is lost, or the noise is too large. However, this paper only considers the image features in the instructional video, and the audio and semantic information are not considered, which may achieve better results if such knowledge is added.

For the CLA-TCN model we proposed, we use CNN and LSTM to extract deep features and spatiotemporal information. MHA is used to make the model focuses more on crucial information. TCN is used to fuse physiological signals with information in knowledge-based features, and the fully connected layer is used to classify emotions. According to the comparative experiment, CNN and LSTM extracted the deeper emotional representation and spatiotemporal information from the temporal features of physiological signals. The emotional recognition results have significantly improved compared to only shallow features. The addition of MHA also makes the model pay more attention to the global and local critical information, effectively improving the recognition accuracy. The addition of knowledge-based features also significantly enhances the recognition results. In addition, we have carried out comparative experiments on DEAP, MAHNOB-HCI, SEED, and WESAD data sets, and our methods have obtained good results on four public data sets.

6 Conclusions and Future Work

This paper proposes a multimodal emotion recognition method in video learning based on knowledge enhancement, which uses knowledge-based features to enhance the emotion recognition process of physiological signals. To be specific, we propose a construction method based on continuous MMTF. We extract the combination of hue, brightness, and saturation in instructional videos, the click-through rate of videos, and the time of emotion generation as knowledge-based features and use the CLA-TCN model for knowledge-enhanced emotion recognition. The experimental results show that the proposed multimodal temporal feature construction method extracts the emotion-related representation from the raw data and effectively preserves the time relationship between the data. Compared with the raw data as input and the features as input, the recognition accuracy is increased by more than 19%. The extracted knowledge-based features contain a wealth of emotional information. Compared with only using physiological signals for emotion recognition, the recognition accuracy of the four kinds of emotions and the overall emotion recognition accuracy is significantly improved after knowledge enhancement. The CLA-TCN model we proposed entirely extracts the deeper emotional representation and the spatiotemporal information contained in the shallow features and pays attention to the local and global information simultaneously. When using the deep features and knowledge-based features for emotional recognition, TCN also fully uses the information in the deep features and knowledge-based features and complements them. The accuracy has reached 97.51%.

This experiment uses knowledge to guide the emotion recognition process of physiological signals. Therefore, a series of knowledge-based features are extracted from the instructional video to enhance the emotion recognition results of physiological signals. However, in this experiment, we only used the video features of the stimulus material and did not use the audio of the stimulus material. In future work, we can consider using audio and more knowledge forms to enhance the emotion recognition process. Furthermore, this paper adopts a direct concatenate method for fuse eye movement features, PPG features, and knowledge-based features. In the subsequent work, we can try more fusion methods, such as feature fusion according to the amount of information contained by different modalities, the amount of noise contained by different modalities, or the degree of complementarity between modalities. In addition, although non-contact methods can collect eye movement data and PPG signals, the PPG signals used in this paper are obtained by contact sensors. Therefore, if we want to extend the proposed method to practical applications, we need to replace PPG with rPPG. However, rPPG still needs to improve with low recognition accuracy. Improving the recognition accuracy of rPPG is also a problem we need to consider in the next step.

Acknowledgement: Thanks for the data acquisition platform provided by Beijing Jinfa Company. Thanks to all the subjects who participated in the data collection experiment.

Funding Statement: This work has been supported by the National Science Foundation of China (Grant Nos. 62267001, 61906051).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] R. Wei, Y. Ding, L. Zhang and Z. Wu, "Design and implementation of emotion recognition module of online learning system," *Modern Educational Technology*, vol. 24, no. 3, pp. 115–122, 2014.

[2] C. Wu, Y. Huang and J. Hwang, "Review of affective computing in education/learning: Trends and challenges," *British Journal of Educational Technology*, vol. 47, no. 6, pp. 1304–1323, 2016.

[3] X. Feng, Y. Wei, X. Pan, L. Qiu and Y. Ma, "Academic emotion classification and recognition method for large-scale online learning environment—Based on A-CNN and LSTM-ATT deep learning pipeline method," *International Journal of Environmental Research and Public Health*, vol. 17, no. 6, pp. 1–16, 2020.

[4] J. Ye, Z. Liao, J. Song, W. Tang, P. Ge *et al.,* "Research on learner emotion recognition method in online learning community," *Journal of Chinese Computer Systems*, vol. 42, no. 5, pp. 912–918, 2021.

[5] X. Pan, B. Hu, Z. Zhou and X. Feng, "Are students happier the more they learn?–Research on the influence of course progress on academic emotion in online learning," *Interactive Learning Environments*, vol. 2022, pp. 1–21, 2022.

[6] J. Zhou, "Research on online learner's learning effect analysis based on micro-expression emotion recognition," M.S. Dissertation, Hubei University, Wuhan, 2020.

[7] Mustaqeem and S. Kwon, "CLSTM: Deep feature-based speech emotion recognition using the hierarchical ConvLSTM network," *Mathematics*, vol. 8, no. 12, pp. 2133, 2020.

[8] X. Du, C. Ma, G. Zhang, J. Li, Y. Lai *et al.,* "An efficient LSTM network for emotion recognition from multichannel EEG signals," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1528–1540, 2022.

[9] X. Li, Y. Ouyang, W. Rong, H. Zhang and X. Zhang, "A shallow BERT-CNN model for sentiment analysis on moocs comments," in *Proc. IEEE Int. Conf. on Engineering*, Yogyakarta, Indonesia, pp. 1–6, 2019.

[10] C. Bian, Y. Zhang, F. Yang, W. Bi and W. Lu, "Spontaneous facial expression database for academic emotion inference in online learning," *IET Computer Vision*, vol. 13, no. 3, pp. 329–337, 2019.

[11] Y. Tian, X. Zhou, M. Zhou and D. Chen, "A review of the methods of learning affective analysis," *Chinese Journal of ICT in Education*, vol. 2021, no. 22, pp. 1–6, 2021.

[12] H. Ullah, M. Uzair, A. Mahmood, M. Ullah, S. T. D. Khan *et al.,* "Internal emotion classification using EEG signal with sparse discriminative ensemble," *IEEE Access*, vol. 7, pp. 40144–40153, 2019.

[13] Q. Liu, L. Ma, S. Z. Fan, M. F. Abbod, C. W. Lu *et al.,* "Design and evaluation of a real time physiological signals acquisition system implemented in multi-operating rooms for anesthesia," *Journal of Medical Systems*, vol. 42, no. 8, pp. 1–19, 2018.

[14] G. Chanel, J. Kronegg, D. Grandjean and T. Pun, "Emotion assessment: Arousal evaluation using EEG's and peripheral physiological signals," in *Proc. Int. Workshop on Multimedia Content Representation, Classification and Security*, Berlin, Heidelberg, pp. 530–537, 2006.

[15] J. Luo, "A research K-12 students' e-learning emotion recognition based on video," M.S. Dissertation, Central China Normal University, Wuhan, 2021.

[16] D. Y. Liliana, T. Basaruddin, M. R. Widyanto and I. I. D. Oriza, "Fuzzy emotion: A natural approach to automatic facial expression recognition from psychological perspective using fuzzy system," *Cognitive Processing*, vol. 20, no. 4, pp. 391–403, 2019.

[17] Mustaqeem and S. Kwon, "Att-Net: Enhanced emotion recognition system using lightweight self-attention module," *Applied Soft Computing*, vol. 102, pp. 107101, 2021.

[18] S. T. Ly, G. Lee, S. Kim and H. Yang, "Gesture-based emotion recognition by 3D-CNN and LSTM with keyframes selection," *International Journal of Contents*, vol. 15, no. 4, pp. 59–64, 2019.

[19] L. Cai, Y. Hu, J. Dong and S. Zhou, "Audio-textual emotion recognition based on improved neural networks," *Mathematical Problems in Engineering*, vol. 2019, pp. 1–9, 2019.

[20] H. Zhang, “Expression-EEG based collaborative multimodal emotion recognition using deep autoencoder,” *IEEE Access*, vol. 8, pp. 164130–164143, 2020.

[21] T. Chen, H. Yin, X. Yuan, Y. Gu, F. Ren *et al.,* “Emotion recognition based on fusion of long short-term memory networks and SVMs,” *Digital Signal Processing*, vol. 117, pp. 1–10, 2021.

[22] C. Chen and C. Wu, “Effects of different video lecture types on sustained attention, emotion, cognitive load, and learning performance,” *Computers & Education*, vol. 80, pp. 108–121, 2015.

[23] S. Kim, J. Cheon, S. Han and H. Kim, “Examining differences of users’ perceptions of multimedia content types in a national online learning system,” *Asia-Pacific Education Researcher*, vol. 20, no. 3, pp. 621–628, 2011.

[24] S. Wang, L. Hao and Q. Ji, “Knowledge-augmented multimodal deep regression Bayesian networks for emotion video tagging,” *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 1084–1097, 2019.

[25] W. Nie, Y. Yan, D. Song and K. Wang, “Multi-modal feature fusion based on multi-layers LSTM for video emotion recognition,” *Multimedia Tools and Applications*, vol. 80, pp. 16205–16214, 2021.

[26] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou *et al.,* “Speech emotion classification using attention-based LSTM,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1675–1685, 2019.

[27] F. Ma, W. Zhang, Y. Li, S. Huang and L. Zhang, “Learning better representations for audio-visual emotion recognition with common information,” *Applied Sciences*, vol. 10, no. 20, pp. 1–23, 2020.

[28] Z. Zhu and J. Li, “Multi-feature information fusion LSTM-RNN detection for OSA,” *Journal of Computer Research and Development*, vol. 57, no. 12, pp. 2547–2555, 2020.

[29] K. Muhammad, Mustaqeem, A. Ullah, A. S. Imran, M. Sajjad *et al.,* “Human action recognition using attention based LSTM network with dilated CNN features,” *Future Generation Computer Systems*, vol. 125, pp. 820–830, 2021.

[30] S. D. Khan, L. Alarabi and S. Basalamah, “Toward smart lockdown: A novel approach for COVID-19 hotspots prediction using a deep hybrid neural network,” *Computers*, vol. 9, no. 4, pp. 99, 2020.

[31] S. Bai, J. Z. Kolter and V. Koltun, “An empirical evaluation of generic convolutional and recurrent networks for sequence modeling,” arXiv preprint arXiv:1803.01271, pp. 1–14, 2018.

[32] Z. Liu, “Noninvasive continuous blood pressure measurement based on ECG and pulse wave signals,” M.S. Dissertation, Chongqing University, Chongqing, 2017.

[33] J. Quinlan, “Induction of decision trees,” *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[34] W. S. Noble, “What is a support vector machine?,” *Nature Biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[35] L. Bbeiman, “Bagging predictors,” *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[36] K. He, X. Zhang, S. Ren and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 770–778, 2016.

[37] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. Int. Conf. on Learning Representations*, San Diego, CA, USA, pp. 1–14, 2015.

[38] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 1251–1258, 2017.

[39] T. Kang, “Emotion recognition using short-term multi-physiological signals,” *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 16, no. 3, pp. 1076–1094, 2022.

[40] G. Lu, W. Cong, J. Wei and J. Yan, “EEG-based emotion recognition using CNN and LSTM,” *Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition)*, vol. 41, no. 1, pp. 58–64, 2021.

[41] P. Schmidt, A. Reiss, R. Dürichen and K. V. Laerhoven, “Introducing wesad, a multimodal dataset for wearable stress and affect detection,” in *Proc. of the 20th ACM Int. Conf. on Multimodal Interaction*, Boulder, CO, USA, pp. 400–408, 2018.

[42] M. Soleymani, J. Lichtenauer, T. Pun and M. Pantic, “A multimodal database for affect recognition and implicit tagging,” *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2011.

[43] S. Koelstra, C. Muhl, M. Soleymani, J. Lee, A. Yazdani *et al.,* “DEAP: A database for emotion analysis using physiological signals,” *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2011.

[44] W. Zheng and B. Lu, “Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks,” *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.