



Multimodal Emotion Recognition by Fusing Video Semantic in MOOC Learning Scenarios

Yuan Zhang, Xiaomei Tao, Hanxu Ai, Tao Chen and
Yanling Gan

EasyChair preprints are intended for rapid
dissemination of research results and are
integrated with the rest of EasyChair.

April 27, 2024

Multimodal Emotion Recognition by Fusing Video Semantic in MOOC Learning Scenarios

Yuan Zhang^c, Xiaomei Tao^{a,b,c,*}, Hanxu Ai^c, Tao Chen^d and Yanling Gan^{a,b,c}

^aKey Lab of Education Blockchain and Intelligent Technology, Ministry of Education

^bGuangxi Key Lab of Multi-Source Information Mining and Security

^cSchool of Computer Science and Engineering, Guangxi Normal University

^dSchool of Computer Science, University of Birmingham

ORCID (Yuan Zhang): <https://orcid.org/0009-0004-5413-487X>, ORCID (Xiaomei Tao): <https://orcid.org/0000-0001-7850-9518>, ORCID (Hanxu Ai): <https://orcid.org/0009-0004-5413-487X>

<https://orcid.org/0000-0001-7850-9518>, ORCID (Hanxu Ai): <https://orcid.org/0009-0004-5413-487X>

Abstract.

In the Massive Open Online Courses (MOOC) learning scenario, the semantic information of instructional videos has a crucial impact on learners' emotional state. Learners mainly acquire knowledge by watching instructional videos, and the semantic information in the videos directly affects learners' emotional states. However, few studies have paid attention to the potential influence of the semantic information of instructional videos on learners' emotional states. To deeply explore the impact of video semantic information on learners' emotions, this paper innovatively proposes a multimodal emotion recognition method by fusing video semantic information and physiological signals. We generate video descriptions through a pre-trained large language model (LLM) to obtain high-level semantic information about instructional videos. Using the cross-attention mechanism for modal interaction, the semantic information is fused with the eye movement and PhotoPlethysmoGraphy (PPG) signals to obtain the features containing the critical information of the three modes. The accurate recognition of learners' emotional states is realized through the emotion classifier. The experimental results show that our method has significantly improved emotion recognition performance, providing a new perspective and efficient method for emotion recognition research in MOOC learning scenarios. The method proposed in this paper not only contributes to a deeper understanding of the impact of instructional videos on learners' emotional states but also provides a beneficial reference for future research on emotion recognition in MOOC learning scenarios.

1 Introduction

In recent years, emotion recognition in MOOC learning has received much attention. Although MOOC learning has advantages such as transcending time and space constraints and abundant learning resources, the problem of high dropout rates remains prominent [2], and emotional deficiency is one of the essential reasons [21]. Emotions play a regulatory and mediating role in cognitive processes [17]; therefore, in the fields of education and cognitive science, understanding and managing emotions is crucial for optimizing learning strategies [8]. Research has shown that the presentation of different content in videos can affect the activity of the emotion regulation areas in

the human brain, thereby influencing an individual's emotional state [27, 14]. Some studies incorporated physical level information such as brightness and saturation from videos into emotion recognition tasks [33]

Early research focused on analyzing students' engagement in learning through classroom tests and student evaluations [12, 31]. More and more studies have been conducted to identify learners' emotional states by acquiring signals such as facial expressions, eye movements, PPG, Electroencephalogram (EEG), Electrodermal Activity (EDA), etc. during the learning process [3, 18, 32, 33]. And achieve accurate identification of learners' emotional states by fusing information from multiple modal data at the data level, feature level, and decision level [3, 4, 23]. In MOOC learning scenarios, learners' emotions are closely related to learners' personal cognition and semantic information in videos, and the presentation of different contents in teaching videos will bring different feelings to learners. Therefore, we hypothesize that in MOOC learning scenarios, learners' emotions are closely related to the semantic information of instructional videos.

We propose a multimodal emotion recognition method based on the above analysis by fusing video semantic information and physiological signals. Specifically, we extract semantic information from instructional videos and fuse it with eye movement and PPG signals through the cross-attention mechanism to improve the performance of emotion recognition in MOOC learning. To the best of our knowledge, we are the first to apply the semantic information of instructional videos to emotion recognition tasks in MOOC learning, providing a new perspective for emotion recognition research in MOOC learning scenarios. The main contributions of this article can be summarized as follows:

- We hypothesize there is a close correlation between the semantic information of instructional videos and learners' emotions. To deeply explore the impact of video semantic information on learners' emotions, this study innovatively proposes a multimodal emotion recognition method that integrates video semantic information and physiological signals. The instructional video's high-level semantic information is obtained by generating video descriptions, which are fused with eye movement and PPG signals to identify the learner's emotional state. This method effectively improves the

* Corresponding Author. Email: xiaomei.tao@gxnu.edu.cn

performance of emotion recognition.

- To effectively capture the correlation and complementarity between different modal features we propose a multi-modal emotion recognition module based on cross-attention fusion. By first learning the feature representations of any two modalities separately and then further fusing the learned features to obtain the feature representations of three modalities, we achieved an effective fusion of the three modalities.
- We conducted extensive experiments and analyzed the experimental results in depth. The effectiveness and feasibility of our method have been comprehensively verified, and experimental results show that our method has achieved significant results in practice, with an accuracy improvement of over 14%.

The rest of this paper is organized as follows: Section 2 reviews previous work on emotion recognition. Section 3 provides a detailed explanation of the proposed method framework and the extraction of video semantic information. Section 4 reports and analyzes the experimental results, extensively verifying the effectiveness of the method proposed in this paper. Section 5 summarizes the experimental results and future work.

2 Related Works

2.1 Emotion Recognition Using Contextual or Semantic Information

Adding context or semantic information to emotion recognition tasks has gradually become a focus of attention. After adding context information, emotion recognition systems can more accurately infer related emotions. Kosti et al. [15] believe that in addition to facial expressions and body postures, scene context also provides important information for us to perceive people's emotions. Scene context information is also a key component in understanding emotional states. Therefore, they created and released the Emotions In Context (EMOTIC) dataset and proposed a baseline CNN model for emotion recognition in scene context. Dashtipour et al. [6] proposed a context-aware multi-modal sentiment analysis framework to predict emotional states accurately.

In addition, some studies use speech transcription to text to obtain additional contextual or semantic information based on audio and visual information. Jiang et al. [13] proposed a fuzzy temporal convolutional network based on context self-attention (CSAT-FTCN) to improve the effect of emotion recognition by using speech-transcribed text as a new modality and integrates it with the original audio and visual modalities. Xia et al. [30] transcribed speech into the text as semantic information to enhance audio and visual features. Meanwhile, semantic information also serves as a new modality for decision fusion with audio and video modalities for emotion recognition. Tzirakis et al. [24] enhanced the performance and effectiveness of emotion recognition by transcribing speech into text as semantic information in speech emotion recognition tasks. The above research shows that incorporating contextual or semantic information into emotion recognition tasks has a positive effect. We believe that further incorporating the high-level semantic information of videos as a global factor into MOOC learning scenarios will also have a positive effect on learners' emotion recognition task.

2.2 Multimodal Emotion Recognition Based on Attention Mechanism

In the field of emotion recognition, multimodal fusion is a challenging task. Early research typically used traditional data level, feature level, or decision level fusion methods [16, 28]. However, with the rise of attention mechanisms, research focus has gradually shifted towards cross-modal interaction [22, 30]. For example, Wang et al. [29] utilized an attention-based fusion emotion transformer fusion (ETF) framework to integrate features from EEG and eye movement signals. Xia et al. [30] designed a semantic enhancement module based on the attention mechanism, which enhances audio and visual features through semantic information. At the same time, semantic information is also integrated with audio and video as a new modality to improve emotion recognition performance. Gong et al. [9] proposed an intra- and inter-modality attention fusion network that effectively learns the critical information between the two modalities and improves the effectiveness of emotion recognition. These studies show that using attention mechanisms can better learn the correlation and complementarity between different modalities, thus achieving more effective multimodal fusion effects.

Based on the semantic information in instructional videos, we propose a multimodal multimodal emotion recognition method by fusing video semantics and physiological signals. By generating video descriptions, we obtain the high-level semantic representation of instructional videos and use cross-attention mechanisms to fuse them with eye movement and PPG signals, effectively improving the performance of MOOC learning emotion recognition.

3 Proposed Method

3.1 Overall Framework of The Method

The multimodal emotion recognition method by fusing video semantic information and physiological signals consists of three main stages: data processing and feature extraction, cross-attention fusion, and emotion classification (as shown in Figure 1). Firstly, in the data processing and feature extraction stages, we preprocess and extract physiological signals and video semantic information respectively. The physiological signals take eye movement and PPG signals as examples, while video semantic information is derived from the extraction of video stimulus materials in video learning scenarios. Then, the extracted eye movement, PPG, and video semantic features are fed into the fusion module. The fusion module is mainly based on the cross-attention mechanism, which combines the features of three modalities in any pairwise manner and inputs them into multi-head attention to learn the corresponding feature representations. Then, these features are further fused to obtain features that contain important information for three modalities. Finally, the fused features are input into the sentiment classifier for final sentiment prediction.

3.2 Video Semantic Information Generation and Feature Extraction

In this study, a crucial task is to obtain semantic information from videos and extract features from video semantic information. To achieve this goal, we first use pre-trained LLM to generate video descriptions to obtain semantic information in instructional videos. Then, we used the pre-trained BERT model [7] to extract the features of video semantic information. The specific process is shown in Figure 2: Firstly, to ensure the stability of the subsequent running process, we preprocess the original videos and convert them

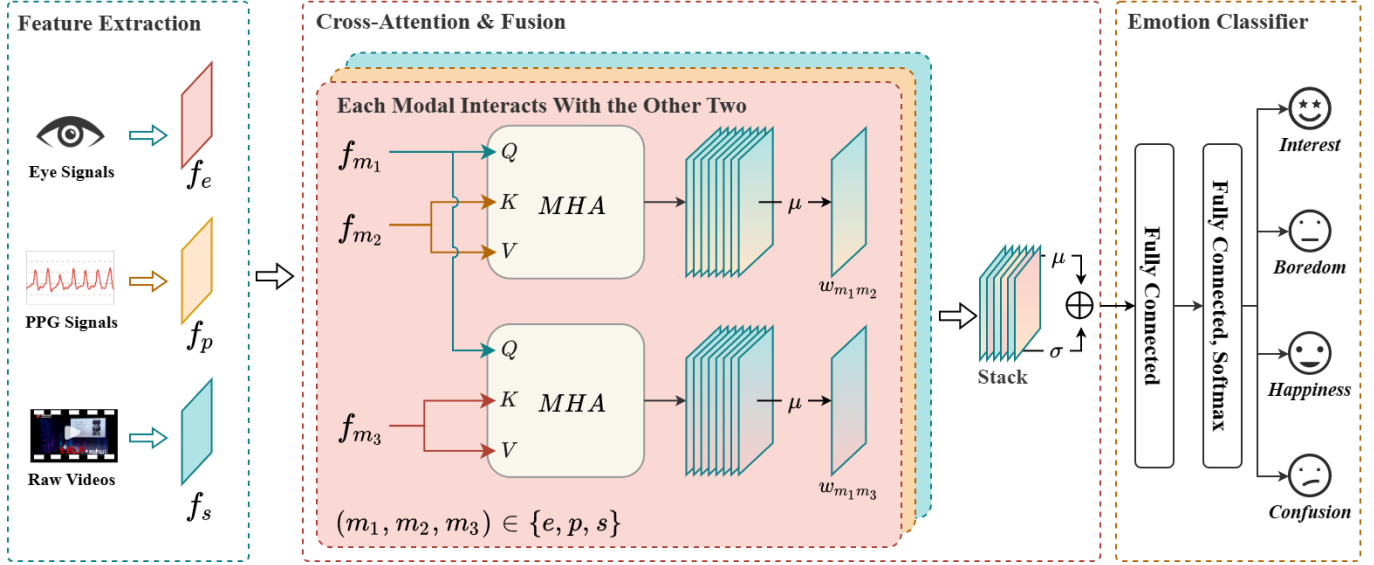


Figure 1. Overall framework of the proposed approach. It consists of three main stages: data processing and feature extraction, cross-attention fusion, and emotion classification. Among them, f_e, f_p, f_s represent eye movements, PPG and semantic features respectively, m_1, m_2, m_3 represent the three modalities, the symbol μ represents the mean, and σ represents the standard deviation

to a unified resolution (1280 x 720), frame rate (25fps), and target bit rate (1000k). Then, we feed the videos into the pre-trained model mPLUG-Owl¹ [34] trained on LLM for generating video descriptions to obtain semantic information. The automatic generation of semantic information in instructional videos has been achieved through the mPLUG-Owl model. As shown in the example in Figure 2, the generated semantic information includes key content such as scenes, objects, actions, and plots in the video.

For extracting video semantic features, we first perform text cleaning on the obtained semantic information and then put the preprocessed video semantic information into the Bert model for feature extraction. Semantic information is encoded in tokenized form during this process, and positional encoding is added to each token. Subsequently, after processing by multiple Transformer encoder layers, the model utilizes the self-attention mechanism and feedforward neural network to capture the semantic relationships between tokens. Next, feature representations are extracted from the last Transformer encoder layer to obtain high-quality semantic features. Due to the high dimensionality of the features extracted using the Bert model, we applied the PCA algorithm to reduce the dimensionality of semantic features. The feature dimensions were reduced to 20, 25, 50, 70, and 100, respectively, and experiments showed that the best effect was achieved at 25 dimensions. Therefore, we chose to reduce the semantic features to 25 dimensions and further use LSTM for encoding to obtain the final video semantic features for subsequent experiments.

3.3 Cross-Attention Fusion

To effectively learn important information between different modalities, we use Multi-Head Attention (MHA) to model the cross-attention fusion module and use MHA to learn information between any two modalities separately. Each MHA module requires three inputs, namely Query (Q), Key (K), and Value (V). In this article,

¹ The mPLUG-Owl model is available on the GitHub, HuggingFace, or ModelScope platforms.

when learning information from two modalities, we use one modality as the input for Q , while the other modality serves as both K and V . Each input is first projected into a different subspace using a linear layer H times, where H represents the number of heads. The projection of each subspace $h \in (0, \dots, H - 1)$ is expressed as:

$$Q_h = W_h^Q e_{m_1} \quad (1)$$

$$K_h = W_h^K e_{m_2} \quad (2)$$

$$V_h = W_h^V e_{m_2} \quad (3)$$

Where $m_1, m_2 \in \{e, p, s\}$ represents the modality used. In each subspace, scaled dot product attention operations were performed on these projections. For subspace h , the attention operation expression is as follows:

$$Att_h(Q_h, K_h, V_h) = Softmax\left(\frac{Q_h K_h}{\sqrt{d_k}}\right) V_h \quad (4)$$

Where $Att_h(\cdot)$ refers to attentional operations in subspace h , and d_k is the characteristic dimension. All H attention outputs are connected in series and passed through a linear layer to obtain the final output of the multi-head attention (MHA) module.

To achieve an effective fusion of different modalities, we input the features of any two modalities into MHA and obtained feature weights that contain common information between these two modalities. Subsequently, we perform a mean operation to obtain the final feature weights $\omega_{m_1 m_2}$ for these two modalities, as follows:

$$\omega_{m_1 m_2} = \mu(MHA(f_{m_1}, f_{m_2}, f_{m_2})) \quad (5)$$

Where $f_{m_1}, f_{m_2} \in \{f_e, f_p, f_s\}$ represents the features of any two modalities. The feature weights of any two modes are calculated to represent: $\omega_{es}, \omega_{ps}, \omega_{se}, \omega_{pe}, \omega_{sp}, \omega_{ep}$. Finally, these weights information are stacked to achieve an effective fusion of the three modalities of eye movement, PPG, and video semantic information, obtained feature weights ω_{eps} containing important information from three modalities:

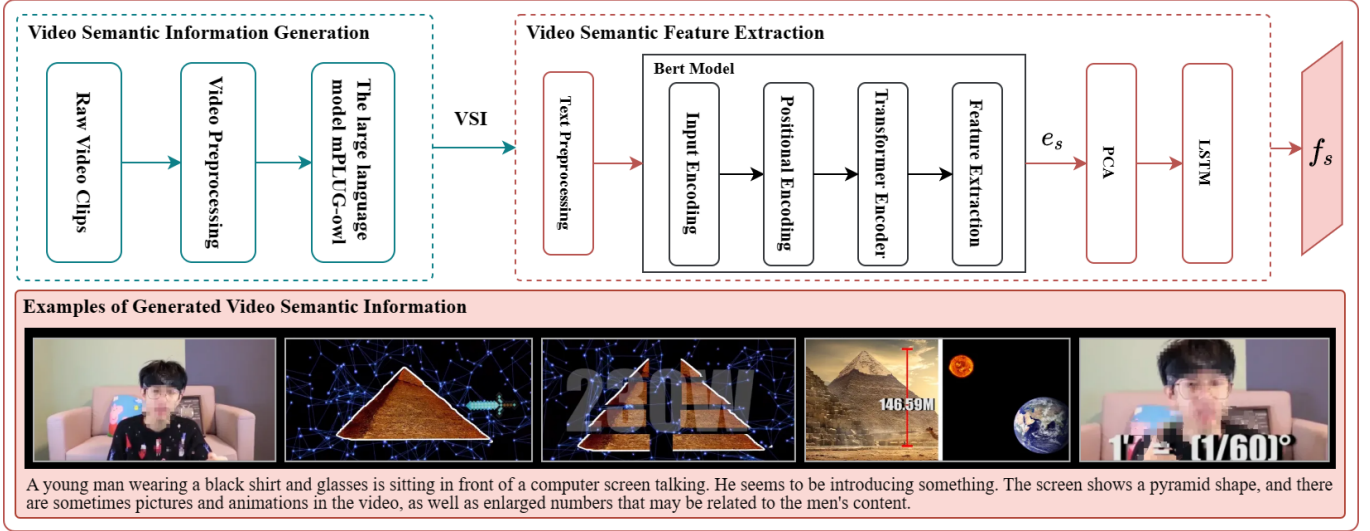


Figure 2. Schematic diagram of video semantic information generation and semantic feature extraction

$$\omega_{eps} = [\omega_{es}, \omega_{ps}, \omega_{se}, \omega_{pe}, \omega_{sp}, \omega_{ep}] \quad (6)$$

Finally, the fused multimodal features are fed into the emotion classifier for emotion prediction. The classifier is composed of two fully connected layers, and the softmax activation function is used in the second fully connected layer. The mathematical expression of emotion prediction is as follows:

$$\hat{y} = \text{Softmax}(FC_{\theta_2}(FC_{\theta_1}[\mu + \sigma])) \quad (7)$$

Where \hat{y} represents the final emotion prediction result, FC_{θ_1} and FC_{θ_2} represent fully connected layers with parameters θ_1 and θ_2 , respectively, μ and σ are the average and standard deviation calculated from the output ω_{eps} of the fusion module, and $+$ represents the concatenation operation.

4 Experiments

4.1 Datasets and Data Processing

To verify the effectiveness of our proposed method, we conducted experiments on the Video Learning Multimodal Emotion Dataset (VLMED) [3, 33], which contained the subjects' eye movement, PPG, facial expression, EDA data and the instructional videos watched by the subjects. The data was collected while the subjects watched instructional videos. This dataset simulates MOOC learning scenarios during the collection process, using 5 carefully selected instructional videos to induce different types of emotions: interest, boredom, happiness, confusion, and distraction. The experiment collected data from 68 subjects, each of whom watched 5 videos in sequence, including 4 shorter (about 2-3 minutes) and 1 longer (about 10 minutes) instructional video.

In this study, we mainly used eye movement, PPG data, and instructional videos from this dataset. Extract data with a time window of 1 second, and process and extract features from eye movement and PPG data using the same methods as in papers [3] and [33], respectively. We also extracted semantic information from instructional videos to expand the data set. The acquisition method of video semantic information and its feature extraction are introduced in Section 3.2.

During the experiment, we observed that the model performed very poorly in recognizing the emotion of Interest category, and the same problem was encountered in the work of literature 1 [3] and literature 2 [19]. We speculate that it may be caused by unbalanced samples in the data set. To solve this problem, we adopted the ADASYN sampling approach [10] to enhance the data. ADASYN is a data resampling-based method that synthesizes small sample categories in the feature space to generate high-quality new samples, thereby balancing the distribution of samples in different categories. The number of samples before and after adaptive synthesis sampling is shown in Table 1.

Table 1. Sample size before and after using ADASYN

Data	Interest	Boredom	Happiness	Confusion
Raw data	1451	2723	1761	2275
ADASYN	2848	2807	2723	2605

4.2 Experimental Setting

In this study, we used NVIDIA GeForce RTX 3070 GPU as the computing platform and constructed and trained the entire model using the TensorFlow framework. We divided the dataset into training and testing sets according to the ratio of 8:2 and trained the model using 5-fold cross-validation on the training set. At the same time, we evaluated the performance of the model using the testing set. During the experiment, we attempted different parameter combinations to determine the optimal parameter configuration as the final parameters of the model. The final network and training parameters of the model are set as follows:

Network Parameters: In the data processing and feature extraction module, we used a Conv1D and a LSTM network to encode eye movement and PPG features. Conv1D includes 16 filters of size 1 and uses ReLU as the activation function; LSTM contains 64 hidden units. Encode semantic features using a LSTM with 64 hidden units. In the cross-attention fusion module, num_heads=8, key_dim=128, and value_dim=64 in multi-head attention. The emotion classifier consists of two fully connected layers, the first consisting of 64 units,

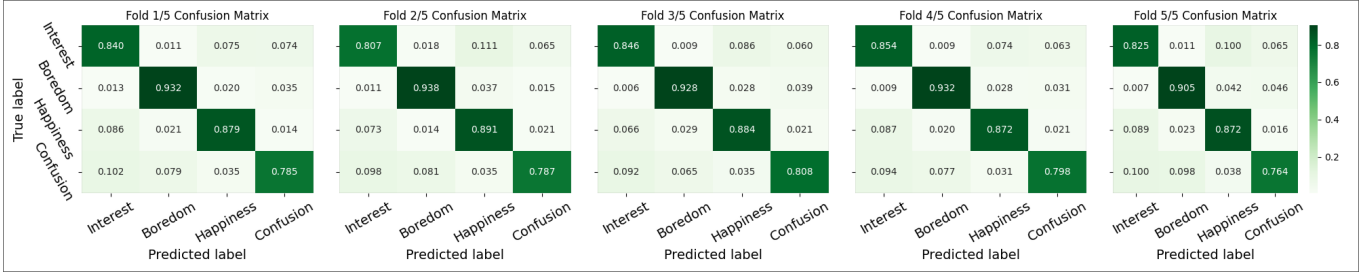


Figure 3. The confusion matrix of each fold of the model under 5-fold cross-validation

while the second consists of 4 units and uses the softmax activation function to achieve the classification of 4 emotions. In addition, we have introduced L2 regularization ($\lambda=0.001$) at various levels of the network to reduce model complexity, prevent overfitting, and enhance the model’s generalization ability.

Training Parameters: When training, we use sparse categorical cross-entropy as the loss function, Adam as the optimizer, and set the random seed to 7 to ensure the repeatability of the experimental results. Set batch size = 32, epoch = 500, and learning rate = $1e-3$. To avoid overfitting, we set the learning rate decay and early stop criteria for model training. If the model does not show improvement for 5 consecutive epochs, the learning rate is attenuated to the original 0.1. When the model does not show better performance for 10 consecutive epochs, we determine that the model is overfitted and terminate the training.

To evaluate the effectiveness of the model, we tested its performance extensively using 5-fold cross-validation. We calculated the average accuracy (Avg_{acc}), average recall (Avg_{recall}), and average F1 score (Avg_{f1}) as evaluation metrics.

4.3 Experimental Results and Analysis

Figure 3 shows the confusion matrix for each fold of our model under 5-fold cross-validation. It can be observed that the difference between the results of each fold is not large, which indicates that our proposed model shows effective and stable performance in MOOC learning scenarios. However, we found that the model had relatively low accuracy in identifying the two categories of Interest and Confusion. Specifically, the Interest category is easily misclassified as either Happiness or Confusion, which may be due to both Interest and Happiness representing positive emotions, so they are easy to confuse when classifying emotions. Similarly, Confusion and Boredom are both negative emotions, resulting in some Confusion samples being incorrectly classified as Boredom. In addition, compared with Happiness and Boredom, Interest and Confusion are neutral emotions with low emotional intensity, and their emotion scores are similar. Therefore, some samples of Interest and Confusion have similar feature distributions on the two physiological signals of eye movement and PPG, which is difficult to distinguish effectively. The above analysis is confirmed in the visualization results of feature distribution in Figure 5.

4.3.1 Effectiveness of adaptive synthetic sampling

To overcome the potential impact of imbalanced data distribution, we adopted the Adaptive Synthesis (ADASYN) sampling method for

data augmentation, and its effectiveness was verified through experiments. The experimental results are shown in Figure 4, where (a) and (b) are the ROC curves before and after using ADASYN, respectively. It can be found that without ADASYN processing, the model performs poorly in recognizing the emotion of Interest category. After ADASYN processing, the model has significantly improved its recognition of various emotions. This method effectively alleviates the problem caused by the unbalanced data distribution and improves the model’s overall performance. It should be emphasized that the enhanced data were used in other experiments in this paper.

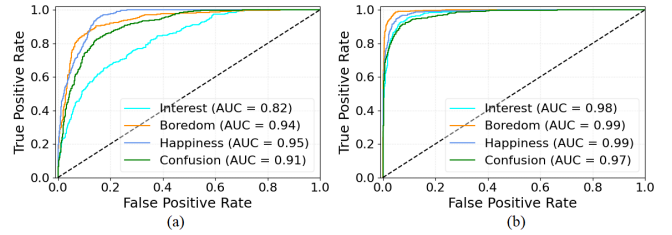


Figure 4. The ROC curves with (a) and without (b) ADASYN

4.3.2 Compare with other models

To better demonstrate the effectiveness of our model, using the data collected in this paper, we reproduce six baseline classifiers and compare them with the methods proposed in this paper, including the traditional machine learning method K nearest neighbor (KNN) [5], deep learning methods of LSTM [1] and CNN-LSTM [11], as well as Transformer [26], CNN-LSTM-MHA-TCN (CLA-TCN) [33] and Cascade Multi-Head Attention (CMHA) [35] using attention mechanisms. The results are shown in Table 2. The effect obtained by using the deep learning method is significantly better than that obtained by the machine learning method, indicating that the deep learning method can extract deeper features. When the attention mechanism is used, the effect is further improved, and our method achieves the best performance, indicating that our model can effectively learn important information between different modalities and more efficiently fuse information from different modalities.

4.3.3 Comparison with different semantic information

Unlike most studies that transcribe audio into text as semantic information, we use generated video descriptions as semantic information.

Table 2. Results compared with other models

Model	Acc \pm std(%)	Recall(%)	F1-score
KNN [5]	59.56 \pm 2.7	59.67	58.98
LSTM [1]	67.59 \pm 2.0	67.89	67.38
CNN-LSTM [11]	73.23 \pm 1.7	73.14	73.06
Transformer [26]	78.72 \pm 1.8	78.96	78.47
CLA-TCN [33]	82.52 \pm 2.1	81.03	81.15
CMHA [35]	83.97 \pm 1.2	84.01	83.88
Ours	86.69 \pm 0.7	86.62	86.59

To demonstrate the effectiveness of the proposed method, we conducted experiments using caption semantic² and semantic information generated by BiliGPT³. Table 3 shows the experimental results. We can see that the emotion recognition effect is improved after using subtitle semantics, but it is not as good as the summary subtitle with reduced redundant information after the summary. When description semantics is used, better results are obtained, because the learners’ emotional production in the learning process is affected by the visual content stimulation, and the video description contains this information. Therefore, using video description as semantic information can get better results, which also reflects the innovation of the method in this paper.

Table 3. The results of using different semantic information

Semantic Type	Acc \pm std(%)	Recall(%)	F1-score
Without Semantic	63.57 \pm 1.9	63.58	0.63
Caption Semantic	76.12 \pm 1.6	76.17	0.76
Caption Summary Semantic	78.44 \pm 1.5	78.54	0.78
Description Semantic	86.69 \pm 0.7	86.62	0.86

4.4 Ablation Studies

4.4.1 Effectiveness of multimodal fusion

We conducted experiments using unimodal, bimodal, and trimodal, and the results are shown in Table 4. We observed that emotion recognition improved significantly when multimodal data was used. This result shows that integrating multi-modal data helps to capture learners’ emotional states more comprehensively, thus achieving higher performance affective perception. In addition, comparing experiments I and II in Table 4, we found that the use of eye movement could obtain better results than PPG signals, indicating that there is a strong correlation between learners’ emotions and eye movement signals, possibly because in MOOC learning scenarios, learners mainly watch instructional videos through vision. This result also suggests that learners’ emotions can be affected by visual stimuli.

4.4.2 Effectiveness of Video Semantics

As shown in Table 4, the experimental effect has been significantly improved after incorporating video semantic information, whether it is bimodal or trimodal. Further comparing experiments IV and V in Table 4, we found that using eye movement signals and video semantic information for emotion recognition is more effective than using PPG signals and video semantic information. This indicates a stronger correlation between eye movement signals and semantic

² Audio transcription into text subtitles as semantic information.

³ <https://bibigpt.co>, First transcribe the audio into text and then further summarize it as semantic information.

information. In MOOC learning scenarios, learners acquire knowledge by watching instructional videos, so eye movement signals are naturally directly affected by the instructional videos. This also confirms the importance of integrating video semantic information into MOOC learning emotion recognition tasks.

4.4.3 Effectiveness of cross-attention

To verify the effectiveness of our proposed cross-attention fusion method, we also conducted experiments by directly concatenating features without using cross-attention fusion. The experimental results show that when cross-attention is used, the accuracy of emotion recognition is significantly improved (See experiments VI and VII in Table 4). This indicates that our model can effectively capture the correlation and complementarity between different modalities, thereby improving the performance of emotion recognition. In our method, the data from three modalities is combined pairwise and fed into MHA, so that each modality can learn information related to the other two modalities. Then, the learned features are further fused to obtain features containing important information about the three modalities. Finally, the fused features are used for emotion recognition. By this method, we achieved the effective fusion of multimodal data, which can make full use of the effective information of each modality and improve the accuracy of emotion recognition.

Table 4. The experimental results of using eye movement (EM), PPG, and video semantic information (VSI) data, as well as whether the cross-attention mechanism (CA) is used

Experiment	Modality	CA	Acc(%)	Recall(%)	F1
I	EM	-	60.15	60.00	0.59
II	PPG	-	44.62	40.79	0.39
III	EM & PPG	✓	63.57	63.54	0.61
IV	PPG & VSI	✓	69.88	69.87	0.69
V	EM & VSI	✓	72.77	72.79	0.73
VI	EM & PPG & VSI	-	72.10	72.23	0.72
VII	EM & PPG & VSI	✓	86.69	86.62	0.86

4.5 Effects on Public Dataset

To demonstrate the generalization ability of our method, we conducted experiments on the publicly available dataset MAHNOB-HCI [20]. This dataset is a multimodal database that synchronously records the data of 27 subjects’ EEG, eye movements, facial video, audio signals, and peripheral physiological signals while they watched 20 emotional videos. We used the EEG and eye movement signals, along with extracted semantic information from the videos in this dataset, for experimentation. We compared the results in terms of arousal (including Calm, Medium arousal, Excited/Activated) and valence (including Unpleasant, Neutral valence, and Pleasant) dimensions with the baseline. The experimental results are shown in Table 5. We can see that compared to using single-modal EEG and eye-tracking data, the performance significantly improves when adding video semantic information. The best results are achieved when using all three modalities simultaneously. This experiment further demonstrates the positive impact of incorporating video semantic information on emotion recognition. It also validates the strong generalization capability of our method, making it suitable for emotion recognition tasks induced by stimulus materials.

Table 5. Experimental results using EEG, Eye Movement (EM), and Video Semantic Information (VSI) in the MAHNOB-HCI dataset

Method	Modality	Acc (%)		F1-score	
		arousal	valence	arousal	valence
Baseline	EEG	52.4	57.0	0.42	0.56
	EM	63.5	68.8	0.60	0.68
	EEG & EM	67.7	76.1	0.62	0.74
Ours	EEG	53.2	55.9	0.49	0.53
	EM	62.9	68.4	0.57	0.63
	EEG & VSI	67.1	72.4	0.61	0.64
	EM & VSI	68.8	73.9	0.67	0.71
	EEG&EM&VSI	82.3	82.8	0.80	0.79

4.6 Visualization

To demonstrate the effectiveness of our method more clearly, we visualized the feature distributions learned by the classifier in the second-to-last layer of our model using t-SNE [25] in three different settings. As shown in Figure 5 (a), it is difficult for the model to effectively distinguish different categories of emotions using only eye movement and PPG data. As shown in Figure 5 (b), with the addition of video semantic information, it can be observed that the feature distribution distinguishes different emotion categories becomes more clear, which further verifies that fusing video semantic information has a positive effect on improving emotion recognition performance. When the cross-attention mechanism is further applied, the feature distribution becomes more obvious (as shown in Figure 5 (c)), indicating that the cross-attention mechanism can effectively learn information between different modalities and improve emotion recognition performance. In addition, in Figure 5, we can also find that it is difficult to distinguish the emotional categories of Interest and Confusion, which explains why the recognition accuracy of Interest and Confusion is low (as shown in Figure 3 and Figure 4 (a)).

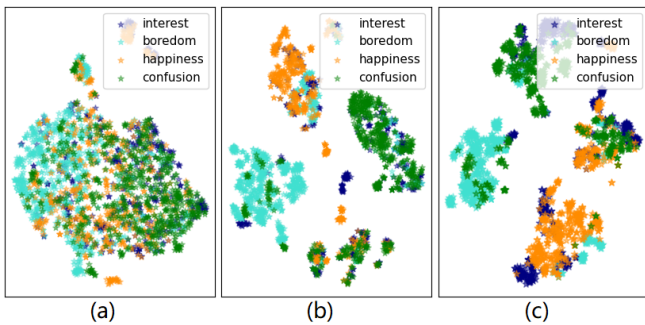


Figure 5. Visualization results of feature distribution in three settings. In Figure (a), only eye movement and PPG data are used. In Figure (b), video semantic information is further integrated on the basis of (a), but no cross-attention mechanism is adopted; In Figure (c), on the basis of (b), the cross-attention mechanism is further introduced

5 Conclusions and Future Works

In this work, we propose a multimodal emotion recognition method that integrates video semantic information and physiological signals, aiming at the particularity of the MOOC learning scenario. This is the first attempt to apply semantic information from instructional videos to emotion recognition tasks in MOOC learning. We use a method of

generating video descriptions to extract high-level semantic information from educational videos, thereby expanding the dataset. Experimental results indicate that incorporating video semantic information has a significantly positive impact on emotion recognition. We use cross-attention to capture semantic correlations between different sequences and have designed a multimodal fusion method based on cross-attention. This method successfully fuses video semantic information with physiological signals, achieving an accuracy improvement of over 14%. Additionally, we adopted adaptive synthetic sampling for data augmentation, effectively eliminating the impact of data distribution imbalance. To validate the generalization ability of our approach, we further conducted experiments on the publicly available HCI dataset. The results indicate that our method can significantly improve the performance of emotion recognition. Overall, through extensive experimentation, we have demonstrated the effectiveness and feasibility of the proposed method, providing new perspectives and effective approaches for emotion recognition studies induced by stimuli materials.

In the current study, we used the global semantic information of the instructional video for analysis. In future work, we will try to extract more fine-grained video semantic information to conduct experiments. In addition, we will also explore more effective multi-modal fusion strategies to fully utilize information from different modalities to achieve higher emotion recognition performance.

6 Data Availability Statement

Due to privacy, copyright, commercial confidentiality, or legal restrictions, the dataset used in this article is not publicly available. If you have specific research or analysis needs for this data set, you can contact the maintainer or relevant person in charge of the data set through the following link: <https://github.com/zhou9794/video-learning-multimodal-emotion-dataset>. Typically, you may need to sign a confidentiality agreement or a data use agreement to gain access to the dataset.

Acknowledgements

We sincerely appreciate all the editors and reviewers' insightful comments and constructive suggestions. This research work was supported by the National Natural Science Foundation of China under Grant No.62267001 and No.62307009.

References

- [1] S. Alhagry, A. A. Fahmy, and R. A. El-Khoribi. Emotion recognition based on eeg using lstm recurrent neural network. *International Journal of Advanced Computer Science and Applications*, 8(10), 2017. doi: 10.14569/IJACSA.2017.081046. URL <http://dx.doi.org/10.14569/IJACSA.2017.081046>.
- [2] K. A. Azhar, N. Iqbal, Z. Shah, and H. Ahmed. Understanding high dropout rates in moocs a qualitative case study from pakistan. *Innovations in Education and Teaching International*, 2023. URL <https://api.semanticscholar.org/CorpusID:258085455>.
- [3] J. Bao, X. Tao, and Y. Zhou. An emotion recognition method based on eye movement and audiovisual features in mooc learning environment. *IEEE Transactions on Computational Social Systems*, 11(1):171–183, 2024. doi: 10.1109/TCSS.2022.3221128.
- [4] 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. *Information Fusion*, 59:127–138, 2020. ISSN 1566-2535. doi: <https://doi.org/10.1016/j.inffus.2020.01.008>. URL <https://www.sciencedirect.com/science/article/pii/S1566253519302143>.
- [5] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967. doi: 10.1109/TIT.1967.1053964.

- [6] K. Dashtipour, M. Gogate, E. Cambria, and A. Hussain. A novel context-aware multimodal framework for persian sentiment analysis. *Neurocomputing*, 457:377–388, 2021. ISSN 0925-2312. doi: <https://doi.org/10.1016/j.neucom.2021.02.020>. URL <https://www.sciencedirect.com/science/article/pii/S0925231221002666>.
- [7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://aclanthology.org/N19-1423>.
- [8] A. R. Faria, A. Almeida, C. Martins, and R. Gonçalves. Emotion effects on online learning. In P. Novais, D. Camacho, C. Analide, A. El Falah Seghrouchni, and C. Badica, editors, *Intelligent Distributed Computing IX*, pages 375–385, Cham, 2016. Springer International Publishing.
- [9] L. Gong, W. Chen, M. Li, and T. Zhang. Emotion recognition from multiple physiological signals using intra- and inter-modality attention fusion network. *Digit. Signal Process.*, 144(C), feb 2024. ISSN 1051-2004. doi: 10.1016/j.dsp.2023.104278. URL <https://doi.org/10.1016/j.dsp.2023.104278>.
- [10] H. He, Y. Bai, E. A. Garcia, and S. Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328, 2008. doi: 10.1109/IJCNN.2008.4633969.
- [11] R. K. Jeevan, V. M. R. S.P., P. Shiva Kumar, and M. Srivikas. Eeg-based emotion recognition using lstm-rnn machine learning algorithm. In *2019 1st International Conference on Innovations in Information and Communication Technology (ICICT)*, pages 1–4, 2019. doi: 10.1109/ICICT1.2019.8741506.
- [12] L. X. Jensen, M. Bearman, and D. Boud. Understanding feedback in online learning a critical review and metaphor analysis. *Computers Education*, 173:104271, 2021. ISSN 0360-1315. doi: <https://doi.org/10.1016/j.compedu.2021.104271>. URL <https://www.sciencedirect.com/science/article/pii/S0360131521001482>.
- [13] D. Jiang, H. Liu, R. Wei, and G. Tu. CSAT-FTCN: A Fuzzy-Oriented Model with Contextual Self-attention Network for Multimodal Emotion Recognition. *Cognitive Computation*, 15(3):1082–1091, May 2023. ISSN 1866-9964. doi: 10.1007/s12559-023-10119-6. URL <https://doi.org/10.1007/s12559-023-10119-6>.
- [14] M. J. Kim, R. A. Loucks, A. L. Palmer, A. C. Brown, K. M. Solomon, A. N. Marchante, and P. J. Whalen. The structural and functional connectivity of the amygdala: From normal emotion to pathological anxiety. *Behavioural Brain Research*, 223(2):403–410, 2011. ISSN 0166-4328. doi: <https://doi.org/10.1016/j.bbr.2011.04.025>. URL <https://www.sciencedirect.com/science/article/pii/S0166432811003354>.
- [15] R. Kostj, J. M. Alvarez, A. Recasens, and A. Lapedriza. Context based emotion recognition using emotic dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11):2755–2766, 2020. doi: 10.1109/TPAMI.2019.2916866.
- [16] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, and A. Košir. Audio-visual emotion fusion (avef): A deep efficient weighted approach. *Inf. Fusion*, 46(C):184192, mar 2019. ISSN 1566-2535. doi: 10.1016/j.inffus.2018.06.003. URL <https://doi.org/10.1016/j.inffus.2018.06.003>.
- [17] C. Meei Tyng, H. U. Amin, M. N. MOHAMAD SAAD, and A. Malik. The influences of emotion on learning and memory. *Frontiers in Psychology*, 8:1454, 08 2017. doi: 10.3389/fpsyg.2017.01454.
- [18] J. Shen, X. Zhang, G. Wang, Z. Ding, and B. Hu. An improved empirical mode decomposition of electroencephalogram signals for depression detection. *IEEE Transactions on Affective Computing*, 13(1):262–271, 2022. doi: 10.1109/TAFFC.2019.2934412.
- [19] X. Shen, J. Bao, X. Tao, and Z. Li. Research on emotion recognition method based on adaptive window and fine-grained features in mooc learning. *Sensors*, 22(19), 2022. ISSN 1424-8220. doi: 10.3390/s22197321. URL <https://www.mdpi.com/1424-8220/22/19/7321>.
- [20] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2012. doi: 10.1109/TAFFC.2011.25.
- [21] H.-L. Sun, T. Sun, F.-Y. Sha, X.-Y. Gu, X.-R. Hou, F.-Y. Zhu, and P.-T. Fang. The influence of teacherstudent interaction on the effects of online learning: Based on a serial mediating model. *Frontiers in Psychology*, 13, 2022. ISSN 1664-1078. doi: 10.3389/fpsyg.2022.779217. URL <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.779217>.
- [22] M. Sun, W. Cui, Y. Zhang, S. Yu, X. Liao, B. Hu, and Y. Li. Attention-rectified and texture-enhanced cross-attention transformer feature fusion network for facial expression recognition. *IEEE Transactions on Industrial Informatics*, 19(12):11823–11832, 2023. doi: 10.1109/TII.2023.3253188.
- [23] X. Tao and Y. Zhang. A multimodal intelligent emotion perception framework by data-driven and knowledge-guided. In *2022 2nd International Conference on Electronic Information Engineering and Computer Technology (EIECT)*, pages 70–73, 2022. doi: 10.1109/EIECT58010.2022.00019.
- [24] P. Tzirakis, A. Nguyen, S. Zafeiriou, and B. W. Schuller. Speech emotion recognition using semantic information. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6279–6283, 2021. doi: 10.1109/ICASSP39728.2021.9414866.
- [25] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL <http://jmlr.org/papers/v9/vandermaaten08a.html>.
- [26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 60006010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.
- [27] P. Vuilleumier. How brains beware: neural mechanisms of emotional attention. *Trends in Cognitive Sciences*, 9(12):585–594, 2005. ISSN 1364-6613. doi: <https://doi.org/10.1016/j.tics.2005.10.011>. URL <https://www.sciencedirect.com/science/article/pii/S1364661305003025>.
- [28] Y. Wang and X. Guan. Multimodal feature fusion and emotion recognition based on variational autoencoder. In *2023 IEEE 5th International Conference on Civil Aviation Safety and Information Technology (ICCASIT)*, pages 819–823, 2023. doi: 10.1109/ICCASIT58768.2023.10351725.
- [29] Y. Wang, W.-B. Jiang, R. Li, and B.-L. Lu. Emotion transformer fusion: Complementary representation properties of eeg and eye movements on recognizing anger and surprise. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1575–1578, 2021. doi: 10.1109/BIBM52615.2021.9669556.
- [30] X. Xia, Y. Zhao, and D. Jiang. Multimodal interaction enhanced representation learning for video emotion recognition. *Frontiers in Neuroscience*, 16, 2022. ISSN 1662-453X. doi: 10.3389/fnins.2022.1086380. URL <https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2022.1086380>.
- [31] Z. Z. Xianglin Pan, Bihao Hu 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*, 31(10):6869–6889, 2023. doi: 10.1080/10494820.2022.2052110. URL <https://doi.org/10.1080/10494820.2022.2052110>.
- [32] M. Yang, Y. Wu, Y. Tao, X. Hu, and B. Hu. Trial selection tensor canonical correlation analysis (tscca) for depression recognition with facial expression and pupil diameter. *IEEE Journal of Biomedical and Health Informatics*, PP:1–12, 10 2023. doi: 10.1109/JBHI.2023.3322271.
- [33] H. Ye, Y. Zhou, and X. Tao. A Method of Multimodal Emotion Recognition in Video Learning Based on Knowledge Enhancement. *Computer Systems Science and Engineering*, 47(2):1709–1732, 2023. ISSN 0267-6192. doi: 10.32604/csse.2023.039186. URL <https://www.techscience.com/csse/v47n2/53638>. Publisher: Tech Science Press.
- [34] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi, C. Li, Y. Xu, H. Chen, J. Tian, Q. Qi, J. Zhang, and F. Huang. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality. 2023. doi: 10.48550/ARXIV.2304.14178. Publisher: [object Object] Version Number: 1.
- [35] J. Zheng, S. Zhang, Z. Wang, X. Wang, and Z. Zeng. Multi-channel weight-sharing autoencoder based on cascade multi-head attention for multimodal emotion recognition. *IEEE Transactions on Multimedia*, 25: 2213–2225, 2023. doi: 10.1109/TMM.2022.3144885.