




Enhancing Online Teaching Effectiveness Through Computer Vision Analysis of Teacher Expressions and Gestures in Educational Videos

Ziqiao Wang^{1,2}, Baoqian Yang¹, Shihan Wang¹, Jinfeng Sun³, Zhefeng Yin^{4*} 

¹ Faculty of Education, Northeast Normal University, Changchun 130024, China

² Academic Affairs Office of Yanbian University, Yanji 133002, China

³ Key Laboratory of Natural Medicines of the Changbai Mountain, Ministry of Education, Yanbian University, Yanji 133002, China

⁴ College of Engineering, Yanbian University, Yanji 133002, China

Corresponding Author Email: zfyin@ybu.edu.cn

Copyright: ©2024 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ts.410309>

ABSTRACT

Received: 30 November 2023

Revised: 15 April 2024

Accepted: 26 May 2024

Available online: 26 June 2024

Keywords:

online education, computer vision, teacher expression recognition, facial action units, temporal attention, spatiotemporal feature disentanglement

With the proliferation of online education, improving the interactivity and effectiveness of online teaching has become a pressing issue. Computer vision technology, with its powerful capabilities in video and image analysis, can be used to deeply analyze teachers' facial expressions and body movements in educational videos, thereby assessing their impact on teaching effectiveness. Although some studies have attempted to apply these techniques, most methods overlook the temporal and spatial features of facial expressions and movements, leading to insufficient recognition accuracy. This paper proposes two innovative methods: a facial expression recognition method for teachers based on facial action units and temporal attention, and a gesture recognition method based on spatiotemporal feature disentanglement. These methods can more accurately capture and analyze the dynamic expressions and movements of teachers, providing new technical support for online education, with the expectation of significantly improving online teaching effectiveness.

1. INTRODUCTION

With the rapid development of online education, more and more students are acquiring knowledge through internet platforms. This mode of learning not only breaks the limitations of time and space but also greatly enriches educational resources [1-4]. However, in an online learning environment, teachers cannot directly observe students' reactions, and students find it difficult to intuitively feel the teacher's emotions and teaching enthusiasm, which to some extent affects the teaching effectiveness [5-9]. Therefore, how to use advanced technical means to enhance the interactivity and effectiveness of online teaching has become an important research topic in the field of education.

Computer vision technology has significant advantages in analyzing video images, capable of accurately capturing and analyzing facial expressions and body movements in videos [10, 11]. By applying these technologies, we can gain a deeper understanding of the way teachers express themselves during the online teaching process and its impact on teaching effectiveness. This has important practical significance for improving the quality of online education and enhancing students' learning experiences [12, 13].

Although many studies have attempted to use computer

vision technology to analyze teachers' facial expressions and movements, these methods often rely too much on static features and fail to fully consider the temporal changes and spatial distribution of expressions [14-17]. Additionally, some methods have low accuracy in recognizing complex expressions and movements, making it difficult to comprehensively evaluate teachers' teaching behaviors [18-23]. These limitations indicate the urgent need for more flexible and accurate analysis methods to improve the depth and breadth of research.

This paper proposes two innovative methods to address the shortcomings of existing research. Firstly, the teacher facial expression recognition method based on facial action units and temporal attention can more accurately capture and analyze the dynamic changes in teachers' facial expressions. Secondly, the teacher gesture recognition method based on spatiotemporal feature disentanglement provides a more detailed analysis of teachers' gestures by separating temporal and spatial features. These methods not only help improve the accuracy of teacher expression recognition but also provide new technical means to enhance the effectiveness of online teaching. Through these studies, we hope to provide strong technical support for online education and promote the comprehensive improvement of educational quality.

2. TEACHER FACIAL EXPRESSION RECOGNITION BASED ON FACIAL ACTION UNITS AND TEMPORAL ATTENTION

This chapter proposes a teacher facial expression recognition network model based on facial action units and temporal attention, aiming to accurately capture and analyze teachers' facial expressions in online teaching videos to improve the ability to assess their teaching effectiveness. The proposed model consists of three parts: a global feature extraction network, a Region of Interest (RoI) feature extraction network, and a label correction and classification module. The global feature extraction network is used to extract the global expression features of each frame of the teacher's facial image. After the global feature extraction network, the RoI feature extraction network is set up to make full use of the details of some key areas of the face, thereby improving the recognition accuracy. The model also sets up a label correction and classification prediction module to correct the erroneous labels in the training samples of teacher facial images and complete the classification of teacher facial expression categories. Figure 1 shows the architecture of the teacher facial expression recognition network.

2.1 Action units grouping rules and RoI area division

Considering the problem of inaccurate division of important areas of teacher facial expressions using facial feature points,

this chapter proposes a method to assist in the division of important facial expression areas using facial action units. First, the Dlib tool is used to detect the feature points of the teacher's facial image. Further, according to the action units division rules, the teacher's facial image is divided into 8 groups, which are 8 areas composed of multiple basic RoI.

2.2 RoI feature extraction network

In the teacher facial expression recognition network based on facial action units and temporal attention, for the input video sequence, the global features of the teacher's face are first extracted through the global feature extraction network. Next, these global features are input into the specially designed RoI feature extraction network, extracting the feature maps of 9 specific facial action units from the global features.

To further refine these RoI features, the fourth SE-Res module in the SE-ResNet-50 network is used to construct the RoI fine-grained feature extraction network to process the 9 RoI feature maps and obtain high-level semantic features. Given that the expression of teachers' facial expressions is formed by the movement of specific RoIs, the model should focus on the ROI features related to the teacher's facial expression for each expression. For this purpose, a spatial attention mechanism is adopted, which compresses and integrates the temporal and spatial features of RoIs through temporal feature compression and spatial feature compression mechanisms.

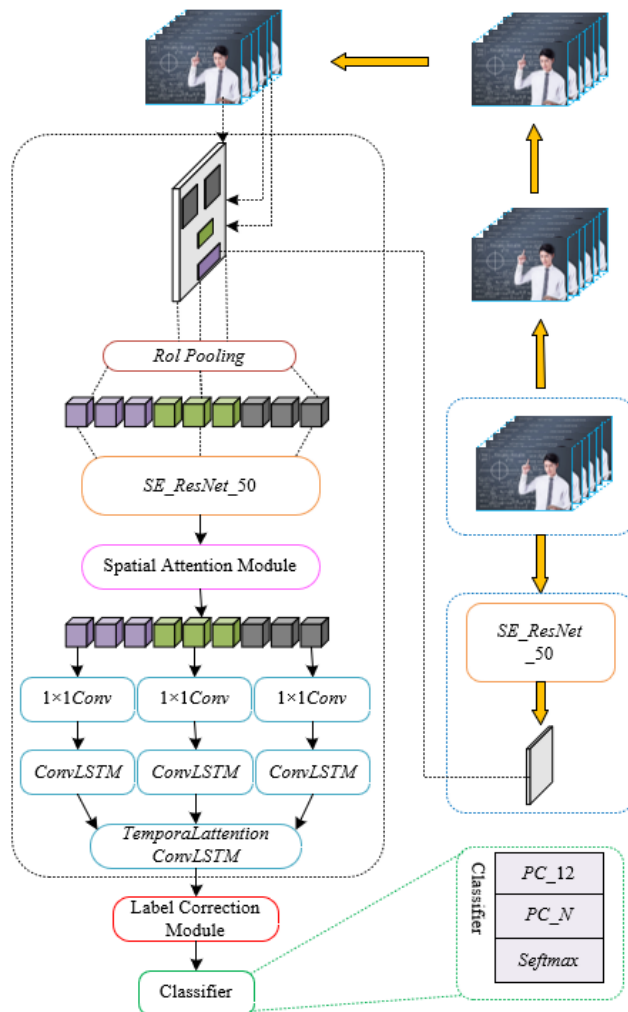


Figure 1. Teacher facial expression recognition network

Additionally, to reduce feature dimensions while retaining important feature information, a channel feature compression mechanism is proposed. This uses a two-layer fully connected network to compress and integrate the channel features of the RoIs and perform concatenation operations, which are then input into subsequent network layers. Based on the positional relationships of RoIs in the facial image, the obtained RoI feature sequences are divided into three groups representing the facial expression features of the upper, middle, and lower parts of the face, respectively. After dimensionality reduction using a 1×1 convolutional layer, these features are input into ConvLSTM. Finally, the output of ConvLSTM is concatenated to obtain the overall facial features.

2.3 Temporal attention ConvLSTM network

In the analysis of teacher expressions and teaching effectiveness in online learning videos, teachers' facial expressions are an important form of non-verbal communication, significantly affecting students' learning experiences and outcomes. To effectively recognize and analyze the emotional features of teachers' facial expressions during teaching, traditional temporal models such as LSTM, while capable of handling time-series data, have limitations in capturing spatial information in video frames. To address this issue, this paper proposes a teacher facial expression recognition network based on facial action units and temporal attention, introducing the temporal attention ConvLSTM network. ConvLSTM combines the advantages of convolutional neural networks and long short-term memory networks, possessing good temporal feature extraction capabilities and effectively modeling spatial information in video frames. Specifically, the output of ConvLSTM at each time step is a three-dimensional vector, meaning it can operate on feature maps in the temporal dimension to generate a four-dimensional tensor. This structure allows ConvLSTM to

capture the complex spatiotemporal relationships in video sequences. However, directly using ConvLSTM for expression classification can lead to a large number of model parameters, increasing computational costs and the risk of overfitting. Therefore, this paper designs a temporal attention module to compress the video feature sequence, reducing feature dimensions and network parameters while retaining the rich feature information in the video sequence. Figure 2 shows the architecture of the temporal attention ConvLSTM network.

Specifically, the feature maps output by ConvLSTM at each time step are concatenated in the channel dimension, resulting in a sequence of features that integrate features from different time steps. This operation ensures that spatial information is fully considered and integrated before temporal modeling, enabling subsequent temporal processing to more effectively capture the dynamic changes of features. Suppose the output features of ConvLSTM at each time step are represented by $G=[G^{(1)}, G^{(2)}, \dots, G^{(s)}]$, and the feature sequence extracted by ConvLSTM is represented by C , then:

$$G = D_{T-CL}(C) \quad (1)$$

For the feature map output at each time step, global average pooling is used to compress it in the channel dimension. Specifically, this operation compresses the feature map of each channel into a single value, representing the overall feature of that channel at the current time step. This step aims to reduce data dimensionality, thereby decreasing computational complexity while retaining important feature information. Suppose the feature of the z -th channel of the feature map output by ConvLSTM at time step s is represented by $G_z^{(s)}$, and the compressed value is represented by $o_z^{(s)}$, then:

$$o_z^{(s)} = D_{CO}(G_z^{(s)}) = \frac{1}{A \times B} \sum_{u=1}^A \sum_{k=1}^B G_z^{(s)}(u, k) \quad (2)$$

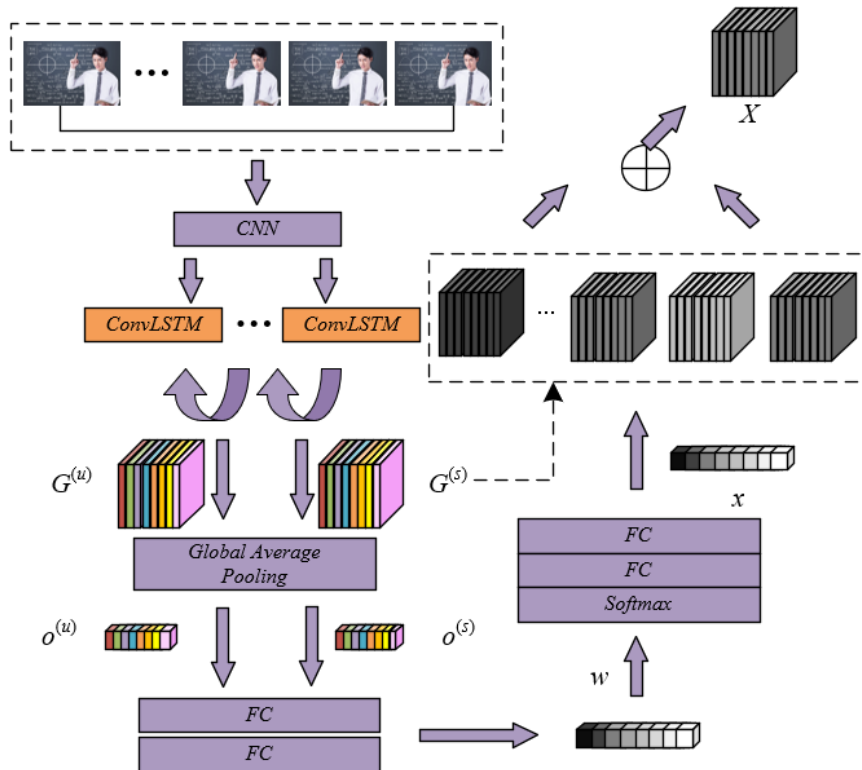


Figure 2. Temporal attention ConvLSTM network architecture

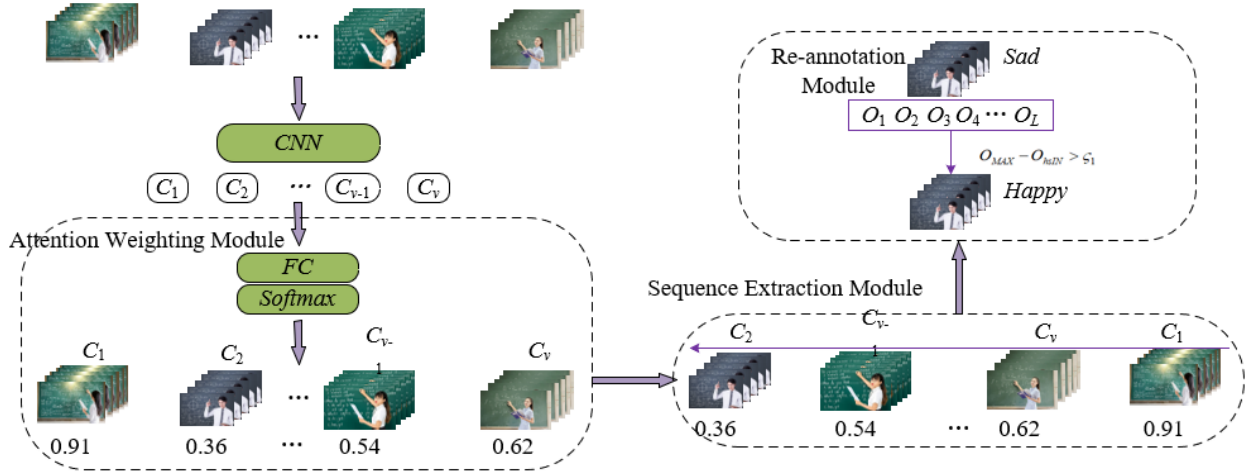


Figure 3. Label correction network workflow

At each time step, all channel-compressed values are concatenated to form a vector $o^{(s)}$. To further integrate the feature information of all channels, this vector $w^{(s)}$ is compressed again to obtain a scalar $w^{(s)}$. This scalar $w^{(s)}$ represents the integrated feature of the feature map output by ConvLSTM at time step s . Suppose the ReLU function is represented by σ , and the weights of the fully connected layers are represented by Q_1 and Q_2 , then:

$$w^{(s)} = D_{RE} \left(o^{(s)}, Q_{RE} \right) = \sigma Q_2 \sigma \left(Q_1 o^{(s)} \right) \quad (3)$$

To obtain the intrinsic association between the feature maps output by ConvLSTM at each time step, the scalar $w^{(s)}$ compressed at each time step is concatenated to form a vector w . Then, a two-layer fully connected neural network and Softmax activation function are used to calculate the attention weight x for each time step. This process aims to assign appropriate weights to the feature maps at each time step to highlight the feature information at key moments. Suppose the weight parameters of the fully connected layers are represented by Q_3 and Q_4 , then:

$$x = \text{Soft max} \left(Q_4 \sigma \left(Q_3 w \right) \right) \quad (4)$$

Finally, the feature maps of all time steps are weighted and summed with their corresponding attention weights to obtain the integrated feature representation. This step combines important temporal feature information, dynamically adjusting the importance of each time step through the attention mechanism, thereby generating a more representative integrated feature for subsequent expression recognition tasks.

$$\hat{G} = \sum_{s=1}^S G^{(s)} \cdot x^{(s)} \quad (5)$$

2.4 Label correction network

In natural environments, datasets often have some inaccuracies in facial expression labeling. These uncertainties can lead to mislabeling in the dataset, causing the model to learn incorrect expression features, severely affecting the model's generalization performance and recognition accuracy. In the scenario of teacher facial expression recognition, these issues are particularly prominent because teachers' expressions

are rich and subtle, easily causing misjudgments by annotators. For example, a teacher's smile, frown, or other subtle expression changes during teaching may be incorrectly labeled as different emotional states, thereby interfering with the model's learning and recognition ability. To avoid the interference of these uncertain samples with the model, this chapter designs a label correction module and embeds it into the network model. This module improves label accuracy through automated and semi-automated methods for initial annotation correction. Figure 3 shows the workflow of the label correction network.

The structure of the label correction module consists of three sub-modules: the self-attention weight module, the weight sorting module, and the label re-annotation module. The specific process is as follows: first, input the teacher facial expression video sequences that may contain labeling errors into the convolutional neural network for feature extraction. These features include facial action units, reflecting specific expression changes in the teacher's face. Next, the self-attention weight module is used to perform deterministic evaluation on each sample. This is done by calculating the importance weight of each sample in the time series to determine which samples might be mislabeled. This module utilizes a temporal attention mechanism to effectively capture key features at different moments in the video sequence, thereby identifying samples whose labels may be biased. Subsequently, the weight sorting module sorts the attention weights of the samples, prioritizing those samples considered to be mislabeled. In this way, the system can concentrate resources and computing power to specifically correct possible labeling errors, thereby improving the overall accuracy of the labels. Finally, the label re-annotation module re-labels the samples suspected of being mislabeled based on the outputs of the previous two modules. This module may combine contextual information of the teacher's expression recognition and features from adjacent moments for more accurate label correction.

Specifically, the self-attention weight module consists of two fully connected layers and a Sigmoid activation function. Its main function is to evaluate each sample and generate weight scores. In the task of teacher facial expression recognition, this module evaluates the matching degree between each sample's label and its actual expression by capturing facial action units and temporal features. Suppose the feature of the u -th sample is represented by C_u , the weight

parameters of the fully connected layers in the self-attention weight module are represented by Q_5 , and the weight score of the u -th sample labeled by the self-attention weight module is represented by x_u , then:

$$x_u = \text{Sigmoid}(Q_5(C_u)) \quad (6)$$

The role of the weight sorting module is to sort the samples based on the weight scores generated by the self-attention weight module. After sorting, the samples are divided into high-weight and low-weight groups. The high-weight group contains samples with higher weight scores, and their labels are considered more accurate; the low-weight group contains samples with lower weight scores, and their labels may contain errors. The proportion of the high-weight group is set to η , and experiments have shown that setting η to 0.7 yields the best results. This means that 70% of the samples are considered to have reliable labels, while the remaining 30% of the samples need further label inspection and correction.

The label re-annotation module is specifically used to correct the labels of samples in the low-weight group. The specific operation is that for each sample in the low-weight group, if the difference between the maximum predicted probability of its label and the predicted probability of its true label exceeds a preset threshold, the module will re-label the sample. For example, if the current label prediction probability of a sample is significantly lower than the probability it should have for its actual label, the sample's label may be incorrect, and the label re-annotation module will re-evaluate and correct the label. This process utilizes deep analysis of facial action units and temporal features to ensure the accuracy of label correction. Suppose the corrected sample label is represented by h' , the threshold by ζ_1 , the maximum value of the model's prediction probability by O_{MAX} , the model's prediction probability on the labeled label by O_{hsIN} , and the original label and the label corresponding to the maximum prediction probability of the sample by h_{ORG} and h_{MAX} , then:

$$h' = \begin{cases} h_{MAX} & \text{If } O_{MAX} - O_{hsIN} > \zeta_1 \\ h_{ORG} & \text{Otherwise} \end{cases} \quad (7)$$

2.5 Loss function

To enhance the robustness and accuracy of the model, we need to specially design the loss function to distinguish between high-weight samples and low-weight samples. During training, to avoid overfitting the network to the low-weight group samples, this paper uses the sample weight scores generated by the label correction module as sample weights, assigning different weight values to each sample. This means that high-weight samples will contribute more to the loss function, while the impact of low-weight samples will be reduced. Suppose the logic-weighted cross-entropy loss function is represented by M_{QZR} . When the u -th sample is input into the network, the k -th value of the final fully connected network's logic output is represented by $m_{u,k}$, the total number of expression categories is represented by L , and the number of samples trained in the same batch is represented by V . The expression is:

$$M_{QZR} = -\frac{1}{V} S \sum_{s=1}^T \log \frac{e^{x_u m_{u,k}}}{\sum_{k=1}^L e^{x_u m_{u,k}}} \quad (8)$$

To ensure that the average weight scores of high-weight samples and low-weight samples differ by more than a preset threshold, this paper introduces a weight regularization term. Suppose the average weight scores of the high-weight group and low-weight group samples are represented by β_G and β_M , respectively. The weight regularization term can be expressed as a max function, where the threshold minus the difference between the two means. Considering both the basic loss and the weight regularization term, the final loss function can be written as the basic loss plus the weighted sum of the regularization term, where the weighting parameter is a hyperparameter used to control the influence of the weight regularization term on the total loss. Suppose the threshold is represented by ζ_2 , then:

$$M_{EE} = \text{MAX} \{0, \zeta_2 - (\beta_G - \beta_M)\} \quad (9)$$

The total loss function of the network is balanced based on γ for the two loss functions. Thus, the overall loss function of the network is:

$$M = \gamma M_{EE} + (1 - \gamma) M_{QZR} \quad (10)$$

3. TEACHER ACTION EXPRESSION RECOGNITION BASED ON SPATIOTEMPORAL FEATURE DISENTANGLEMENT

This paper proposes a teacher action expression recognition model based on spatiotemporal feature disentanglement. To fully utilize the action semantic information contained in teaching video sequences, a self-attention mechanism is introduced. This mechanism can effectively capture key actions and detailed features in video sequences, enhancing the model's ability to understand teachers' actions. Considering the specificity of spatiotemporal features in teaching videos, we designed a spatiotemporal feature disentanglement network to ensure the independence of features in the temporal and spatial dimensions, thereby improving the model's ability to analyze complex action sequences. Specifically, the spatiotemporal feature disentanglement network separates temporal and spatial information, allowing the model to handle and understand these two types of features' unique attributes separately. Temporal features reflect the continuity and trend of teachers' actions, while spatial features capture the details and positional information of specific actions. Through this disentanglement process, the model can more accurately recognize teachers' action expressions, enabling more detailed analysis of actions during the teaching process. Additionally, to enhance the model's representation ability for teachers' actions, we incorporated a regional layer network. The regional layer network focuses on obtaining detailed feature information of teachers' actions, especially when teachers perform specific teaching actions, finely capturing subtle changes and local features of these actions.

3.1 Global feature extraction regional layer network for teacher actions

In teacher action expression recognition, the similarity in the manifestation of different actions is a challenge, as these actions often have highly similar feature information in a local area of the teacher's body. To address this issue, a network

structure is needed that can simultaneously focus on both local and global features of the teacher's body, enabling the model to comprehensively understand and perceive teacher actions. In the teacher action expression recognition model based on spatiotemporal feature disentanglement, the global feature extraction network includes a regional layer network specifically designed to capture action information in important areas of the teacher's body. This network first uses the first three modules of the SE-ResNet-50 network to extract global features of the teacher's body, which can capture the overall form and dynamic changes of teacher actions. However, relying solely on global features may not be sufficient to distinguish similar actions, as certain local action features exhibit high similarity across different teacher action expressions. Therefore, the regional layer network further refines the feature extraction process. It not only focuses on the local action features of the teacher's body but also extends to larger critical areas of the teacher's body to capture the action details of these areas. In this way, the model can organically combine local and global features.

In teacher action expression recognition, standard convolutional layers typically use shared convolutional kernels or filters to process the entire image. However, due to the higher structural characteristics of the teacher's body, this approach is insufficient to effectively capture the local and subtle appearance changes in important regions of the teacher's body actions. Moreover, teachers' action expressions often contain rich local information, such as gestures, facial expressions, and body postures, which hold different meanings and importance in various teaching contexts. Hence, relying solely on globally shared convolutional kernels cannot adequately capture these subtle yet significant changes. To address this shortcoming and fully utilize the correlations between important regions of teachers' actions, we introduce a regional layer network in the global feature extraction network for teacher actions, specifically designed to capture detailed features in key regions of the teacher's body actions. This network can perform more precise feature extraction for key areas such as the head, hands, and upper body of the teacher. This not only enhances the model's ability to perceive action details in these key regions but also better captures the dynamic correlations between these regions.

The regional layer network structure mainly includes three steps: patch cropping, local convolution, and identity addition, to achieve the extraction and importance weighting of teacher action features in the input spatiotemporal feature sequence. First, the input spatiotemporal feature sequence is divided into multiple patches, each representing a small spatiotemporal region to capture the feature changes in local areas of the teacher's body. Next, the local convolution layer performs feature learning on each patch, redistributing weights based on each patch's contribution to the action, highlighting important features. Finally, through identity addition, the patches processed by local convolution are recombined, matched back to their original positions, generating an action information weight feature map. This feature map emphasizes regions with rich action feature information on the teacher's body, suppressing regions with less information, achieving effective extraction and expression of teacher action features.

3.2 Self-attention and spatiotemporal feature disentanglement network

In the teacher action expression recognition model for

online learning videos, teacher action expression needs to focus on both temporal and spatial feature information, which have significant differences in their characteristics. Previous methods often treat temporal and spatial features equally, which is evidently unreasonable. Moreover, this approach has relatively high computational complexity, which is not conducive to efficient processing in practical applications. To address this issue, this paper proposes a disentanglement operation for temporal and spatial dimensions in the study of the teacher action expression recognition model based on spatiotemporal feature disentanglement. The main goal of this method is to specialize the processing of temporal and spatial features separately while preserving the differences between the two dimensions, thereby improving the model's recognition performance. Through this disentanglement operation, the model can more precisely capture the dynamic changes of teacher actions in the temporal dimension as well as detailed position and posture information in the spatial dimension. Specifically, temporal feature processing can better capture the continuity and rhythm of actions, while spatial feature processing can more clearly identify the specific details and structure of actions. Suppose the attention map of the s -th frame is represented by X^s , and $A_s \in E^{V \times V}$. Two embedding functions are represented by δ and ψ , and the transposed matrix is represented by $'$. The specific attention map obtained is given by the following formula:

$$X^s = \text{Softmax} \left(\delta(A_s) \psi(A_s)' \right) \quad (11)$$

In traditional video action recognition models, attention is usually calculated only for spatial features within each video frame, neglecting the interconnections between different video frames, resulting in insufficient modeling capabilities. Furthermore, the computational complexity of this method is $P(SI^2Z)$, which is inefficient for processing long video sequences. To address this issue, suppose there are n important regions of teacher body actions, their attention maps are represented by $X_v \in R^{V \times Z}$, and the model input is represented by $A_v \in R^{S \times Z}$. The calculation of temporal attention is similar to spatial attention but is not limited to the intra-frame relationships; it also includes cross-frame relationships. The overall computational complexity of this method is $P(SI^2Z + VS^2Z)$, capturing detailed changes in teacher actions in both temporal and spatial dimensions by summing the computational complexities of temporal and spatial attention calculations.

$$X^s = \text{Softmax} \left(\sum_s^s \sum_\pi^s \delta(A_s) \psi(A_\pi)' \right) \quad (12)$$

Although this method can more comprehensively capture detailed changes in teacher actions in both temporal and spatial dimensions, its computational complexity is $P(S^2I^2Z + I^2S^2Z)$, which is relatively large and not suitable for practical applications. Therefore, this paper proposes a compromise solution, calculating attention maps only for the important regions of teacher body actions in the current video frame, performing attention map calculations frame by frame, and finally averaging the attention maps of all video frames. This approach effectively captures important action features in each video frame while reducing the model overfitting phenomenon, thereby enhancing the model's generalization ability. The

computational complexity of this scheme is $P(SV^2Z+VS^2Z)$, significantly reducing the demand for computational resources.

In the model, we combine the self-attention mechanism and spatiotemporal feature disentanglement network to improve the analysis ability of teacher action expressions and teaching effectiveness in teaching videos. Specifically, the model captures subtle action expressions by focusing on the spatial and temporal attention of important regions of the teacher's body. This method differs from traditional video action recognition, which usually focuses only on the actions of the entire video frame, neglecting the importance of details and specific regions. Based on the above analysis, the model calculates attention maps for important regions of teacher body actions in the video frame by frame and uses the Tanh function to calculate these attention maps. Subsequently, the calculated attention maps are multiplied by the original input features to obtain the final output features. This method not only accurately captures important action features in each video frame but also reduces the model overfitting phenomenon through frame-by-frame averaging, thereby enhancing the model's generalization ability. Figure 4 shows the principle diagram of the spatiotemporal feature disentanglement module.

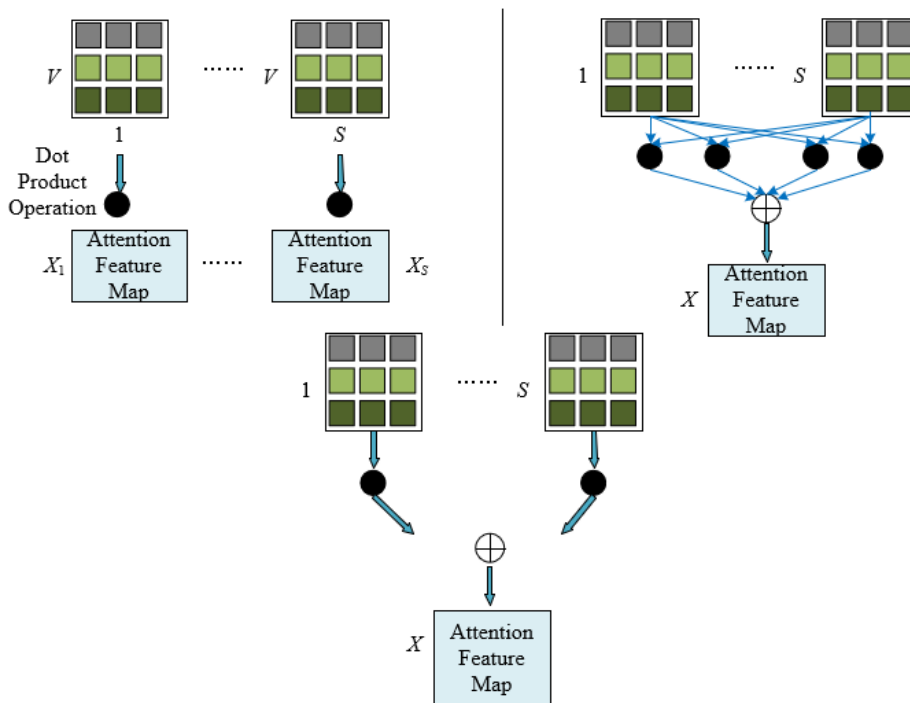


Figure 4. Principle diagram of the spatiotemporal feature disentanglement module

4. EXPERIMENTAL RESULTS AND ANALYSIS

According to the data in Table 1, the method based on facial action unit division consistently outperforms the method based on important facial region division across different types of teaching videos. In classroom lecture videos, the accuracy of the facial action unit division method is 88.23%, compared to 83.26% for the important facial region division method. In Q&A interaction videos, the former's accuracy is 90.23%, significantly higher than the latter's 88.97%. Even in the relatively challenging post-class summary videos, the facial action unit division method still slightly outperforms the important facial region division method with an accuracy of

Specifically, the network input of the spatiotemporal feature disentanglement module is a video feature sequence containing V regions of interest, S frames, and Z channels. In each layer, the input features are first viewed as a $V \times SC$ matrix, i.e., V elements with SC channels, and input into the spatial attention module to obtain the spatial correspondence between each RoI . After processing by the spatial attention module, the output matrix is rearranged into S elements, each containing V Z -channels, which is equivalent to an $S \times VZ$ matrix. Next, these rearranged features are input into the temporal attention module to extract the temporal features between each video frame. In this way, we can capture the changes and continuity of teacher actions at different time points, thereby better understanding the key actions and expressions during the teaching process. The entire spatiotemporal feature disentanglement module consists of L stacked layers to gradually update and refine the spatiotemporal features. After these features are processed through multiple layers, they finally pass through a global average pooling layer to reduce feature dimensions, then through a fully connected layer for further processing, and finally into the classification layer to obtain the final prediction results of teacher action expressions.

52.46% versus 50.21%. These data indicate that the facial action unit division method achieves higher recognition accuracy in various video scenarios. The experimental results show that the teacher facial expression recognition method based on facial action units and temporal attention significantly improves the accuracy of facial expression recognition. This method can more accurately capture the dynamic changes of facial expressions, especially in teaching videos that contain complex expressions and diverse actions. Compared to the method based solely on important facial region division, the facial action unit division method can finely capture the subtle movements and changes of facial muscles, providing more precise facial expression analysis.

Table 1. The impact of RoI division rules on teacher facial expression recognition accuracy (%)

Dataset	Division Based on Facial Action Units	Division Based on Important Facial Regions
Classroom Lecture Video	88.23	83.26
Q&A Interaction Video	90.23	88.97
Post-Class Summary Video	52.46	50.21

Table 2. The impact of temporal attention module on teacher facial expression recognition accuracy (%)

Dataset	With Temporal Attention Module	Without Temporal Attention Module
Classroom Lecture Video	87.23	84.26
Q&A Interaction Video	90.54	88.78
Post-Class Summary Video	52.12	50.43

Table 3. The impact of label correction module on teacher facial expression recognition accuracy (%)

Dataset	With Label Correction Module	Without Label Correction Module
Classroom Lecture Video	87.12	84.41
Q&A Interaction Video	90.33	88.78
Post-Class Summary Video	52.43	48.26

Table 4. The impact of regional layer network on teacher action expression recognition accuracy (%)

Method	Classroom Lecture Video	Q&A Interaction Video	Post-Class Summary Video
Without Regional Layer Network	85.32	88.95	51.23
With Regional Layer Network	86.34	89.32	51.64

Table 2 shows the impact of the temporal attention module on teacher facial expression recognition accuracy. The data indicates a significant improvement in recognition accuracy with the addition of the temporal attention module. In classroom lecture videos, the accuracy with the temporal attention module is 87.23%, compared to 84.26% without it, an increase of about 3 percentage points. In Q&A interaction videos, the accuracy with the temporal attention module is 90.54%, higher than 88.78% without it. Even in post-class summary videos, the temporal attention module improves accuracy from 50.43% to 52.12%. These data indicate that the temporal attention module effectively enhances teacher facial expression recognition accuracy across different types of teaching videos. The analysis of experimental results shows that the temporal attention module plays a significant enhancing role in teacher facial expression recognition. By capturing the temporal dynamic changes of facial expressions, the temporal attention module can more finely analyze the subtle changes and continuity characteristics of teacher facial expressions. Compared to models without the temporal attention module, the addition of this module enables better understanding and differentiation of expression changes between consecutive frames, thereby improving recognition accuracy.

Table 3 shows the impact of the label correction module on teacher facial expression recognition accuracy. The data indicates a significant improvement in recognition accuracy with the addition of the label correction module. In classroom lecture videos, the accuracy with the label correction module is 87.12%, compared to 84.41% without it, an increase of about 2.71 percentage points. In Q&A interaction videos, the accuracy with the label correction module is 90.33%, higher than 88.78% without it. In post-class summary videos, the label correction module improves accuracy from 48.26% to 52.43%, a notable increase. These data indicate that the label correction module significantly enhances teacher facial expression recognition accuracy across different types of teaching videos. The analysis of experimental results shows

that the label correction module plays a crucial enhancing role in teacher facial expression recognition. By dynamically adjusting and optimizing labels, the label correction module can more accurately reflect the actual changes in teacher facial expressions, thereby reducing errors and noise interference and improving recognition accuracy. Combining the data from Table 3, it is evident that the label correction module effectively improves recognition performance in various teaching scenarios, particularly in the complex and variable post-class summary videos.

Table 4 shows the impact of the regional layer network on teacher action expression recognition accuracy. The data indicates that recognition accuracy improves with the addition of the regional layer network. In classroom lecture videos, the accuracy with the regional layer network is 86.34%, compared to 85.32% without it, an increase of 1.02 percentage points. In Q&A interaction videos, the accuracy with the regional layer network is 89.32%, higher than 88.95% without it. Even in post-class summary videos, the regional layer network improves accuracy from 51.23% to 51.64%. Although the improvement margin is relatively small, these data indicate that the regional layer network helps improve teacher action expression recognition accuracy across different types of teaching videos. The analysis of experimental results shows that the regional layer network plays a positive role in teacher action expression recognition. By separating and analyzing the temporal and spatial features in videos, the regional layer network can more finely capture changes in teacher actions, thereby improving recognition accuracy. Combining the data from Table 4, it can be seen that the addition of the regional layer network, although resulting in small improvements, consistently enhances performance across all types of teaching videos. This indicates that the teacher action expression recognition method based on spatiotemporal feature disentanglement improves the model's ability to capture action details by more effectively separating and processing the temporal and spatial dimensions of actions.

Table 5 shows the impact of the spatiotemporal feature

disentanglement module on teacher action expression recognition accuracy. The data indicates that recognition accuracy improves with the addition of the spatiotemporal feature disentanglement module. In classroom lecture videos, the accuracy with the spatiotemporal feature disentanglement module is 86.34%, compared to 85.32% without it, an increase of 1.02 percentage points. In Q&A interaction videos, the accuracy with the spatiotemporal feature disentanglement module is 89.32%, higher than 88.95% without it. In post-class summary videos, the spatiotemporal feature disentanglement module improves accuracy from 51.23% to 51.64%. Although the improvement margin is relatively small, these data indicate that the spatiotemporal feature disentanglement module helps improve teacher action expression recognition accuracy across different types of teaching videos. The analysis of experimental results shows that the spatiotemporal feature disentanglement module plays a positive role in teacher action expression recognition. By separately processing the temporal and spatial features in videos, the spatiotemporal feature disentanglement module can more accurately and finely capture changes in teacher actions, thereby improving recognition accuracy. Combining the data from Table 5, it can be seen that the addition of the spatiotemporal feature disentanglement module, although resulting in small improvements, consistently enhances performance across all types of teaching videos. This indicates that the teacher action expression recognition method based on spatiotemporal feature disentanglement improves the model's ability to capture action details by more effectively separating and processing the temporal and spatial dimensions of actions.

Table 6 shows the analysis of covariance results for the

effect of different facial and action expressions of teachers on student engagement. In terms of classroom performance, facial expressions have a significant impact on student engagement, with an F-value of 205.69, p-value less than 0.001, and an effect size (η^2) of 0.68, indicating a very strong influence of facial expressions. However, the impact of body expressions is not significant, with an F-value of 0.62, p-value of 0.423, and a very small effect size (η^2 of 0.01). The interaction effect between facial expressions and body expressions is also not significant, with an F-value of 0.01 and a p-value of 0.985. In terms of classroom evaluation feedback, none of the variables, including facial expressions, body expressions, and their interaction effect, have a significant impact on student engagement, with all p-values greater than 0.05. These results indicate that teachers' facial expressions have a significant impact on student engagement in classroom performance, while body expressions and their interaction have a minimal effect. The analysis of experimental results shows that facial expressions play a significant role in influencing student engagement in the classroom, whereas the impact of body expressions is relatively weak. The teacher facial expression recognition method based on facial action units and temporal attention, by more accurately capturing and analyzing the dynamic changes of teachers' facial expressions, can effectively improve the accuracy of facial expression recognition, thereby better understanding and analyzing its impact on student engagement. This finding validates the effectiveness of the proposed method in this paper, highlighting the importance of accurate facial expression recognition in the context of significant impact on student engagement.

Table 5. The impact of spatiotemporal feature disentanglement module on teacher action expression recognition accuracy (%)

Method	Classroom Lecture Video	Q&A Interaction Video	Post-Class Summary Video
Without Spatiotemporal Feature Disentanglement Module	85.32	88.95	51.23
With Spatiotemporal Feature Disentanglement Module	86.34	89.32	51.64

Table 6. Analysis of covariance for the effect of teacher's different facial expressions and action expressions on student engagement

Variable	Effect	df	Mean Square	F	p	η^2
Classroom Performance	Pre-test Engagement Score (Covariate)	1	2.78	1.81	0.184	0.02
	Facial Expression	1	325.26	205.69	<0.001	0.68
	Body Expression	1	0.98	0.62	0.423	0.01
	Facial Expression × Body Expression	1	0.01	0.01	0.985	0.001
Classroom Evaluation Feedback	Pre-test Engagement Score (Covariate)	1	0.87	0.21	0.635	0.002
	Facial Expression	1	2.65	0.65	0.425	0.01
	Body Expression	1	0.01	0.001	0.987	<0.001
	Facial Expression × Body Expression	1	5.89	1.48	0.236	0.02

Table 7. Analysis of covariance for the effect of teacher's different facial and action expressions on student emotional states

Variable	Effect	df	Mean Square	F	p	η^2
Positive Emotion	Pre-test Engagement Score (Covariate)	1	142.26	3.45	0.066	0.04
	Facial Expression	1	321.26	7.54	0.007	0.07
	Body Expression	1	42.56	1.02	0.321	0.01
	Facial Expression × Body Expression	1	101.25	2.36	0.124	0.03
Negative Emotion	Pre-test Engagement Score (Covariate)	1	41.59	3.56	0.058	0.04
	Facial Expression	1	10.36	0.94	0.325	0.01
	Body Expression	1	6.89	0.57	0.445	0.01
	Facial Expression × Body Expression	1	19.87	1.77	0.189	0.02

Table 8. Analysis of covariance for the effect of teacher's different facial and action expressions on student learning motivation

Variable	Effect	df	Mean Square	F	p	η^2
Learning Motivation	Pre-test Engagement Score (Covariate)	1	0.22	0.18	0.678	0.002
	Facial Expression	1	11.23	9.36	0.000	0.09
	Body Expression	1	2.89	2.34	0.123	0.02
	Facial Expression \times Body Expression	1	0.23	0.18	0.638	0.01

Table 7 shows the analysis of covariance results for the effect of different facial and action expressions of teachers on student emotional states. In terms of positive emotions, facial expressions have a significant impact on students' positive emotions, with an F-value of 7.54, p-value of 0.007, and an effect size (η^2) of 0.07, indicating that facial expressions can significantly enhance students' positive emotions. However, the impact of body expressions is not significant, with an F-value of 1.02, p-value of 0.321, and a very small effect size (η^2 of 0.01). The interaction effect between facial expressions and body expressions is also not significant, with an F-value of 2.36 and a p-value of 0.124. In terms of negative emotions, none of the variables, including facial expressions, body expressions, and their interaction effect, have a significant impact on students' negative emotions, with all p-values greater than 0.05. These results indicate that teachers' facial expressions play a significant role in regulating positive emotions, whereas body expressions and their interaction have a minimal effect. The analysis of experimental results shows that facial expressions have a significant impact on promoting students' positive emotions, while the impact of body expressions is relatively weak. This result validates the effectiveness of the teacher facial expression recognition method based on facial action units and temporal attention. By more accurately capturing and analyzing the dynamic changes of teachers' facial expressions, it can significantly improve the accuracy of facial expression recognition, thereby better understanding and analyzing its impact on students' emotional states. In the context of the significant impact on positive emotions, accurate facial expression recognition is especially important.

Table 8 shows the analysis of covariance results for the effect of different facial and action expressions of teachers on student learning motivation. Firstly, the pre-test learning motivation score (covariate) does not have a significant impact in this analysis, with an F-value of 0.18, a p-value of 0.678, and an effect size (η^2) of 0.002. Facial expressions significantly affect student learning motivation, with an F-value of 9.36, a p-value of 0.000, and an effect size (η^2) of 0.09, indicating that teachers' facial expressions can significantly enhance students' learning motivation. In contrast, body expressions do not have a significant impact, with an F-value of 2.34, a p-value of 0.123, and an effect size (η^2) of 0.02. The interaction effect between facial expressions and body expressions is also not significant, with an F-value of 0.18, a p-value of 0.638, and an effect size (η^2) of 0.01. These results indicate that teachers' facial expressions have a significant effect on increasing students' learning motivation, while body expressions and their interaction effects do not show a significant impact. Through the analysis of experimental results, it can be concluded that facial expressions play a significant role in enhancing students' learning motivation. This validates the effectiveness of the teacher facial expression recognition method based on facial action units and temporal attention, which can more accurately capture the dynamic changes of teachers' facial expressions,

thereby better understanding and analyzing their impact on students' learning motivation. In the context of significant impact on learning motivation, accurate recognition of facial expressions is especially important, highlighting the crucial role of facial expressions in the teaching process.

5. CONCLUSION

This paper proposed two innovative methods to address the shortcomings of existing research: a teacher facial expression recognition method based on facial action units and temporal attention, and a teacher action expression recognition method based on spatiotemporal feature disentanglement. Experimental results show that these methods significantly improve the accuracy of teacher expression recognition. The RoI division rule, temporal attention module, and label correction module all have a significant impact on the accuracy of teacher facial expression recognition; the regional layer network and spatiotemporal feature disentanglement module also significantly improve the recognition rate of teacher action expressions. Additionally, covariance analysis was used to study the impact of different facial and action expressions of teachers on student engagement, emotional states, and learning motivation. It was found that teachers' facial expressions have a significant positive impact on students' learning motivation, while body expressions and their interaction effects did not show a significant impact.

The research content and experimental results of this paper demonstrate that the teacher facial expression recognition method based on facial action units and temporal attention, as well as the action expression recognition method based on spatiotemporal feature disentanglement, significantly improve the accuracy of teacher expression recognition in online learning videos. This not only validates the critical role of facial expressions in enhancing students' learning motivation but also provides important insights into understanding the potential impact of teacher expressions on student engagement and emotional states through the method of spatiotemporal feature disentanglement. The research in this paper has significant value for improving teaching effectiveness based on the accurate recognition of teacher expressions. However, there are certain limitations, such as the insignificant independent impact of body expressions, which may be limited by the current experimental design or the complexity of the dataset. Future research should further expand the sample size and cover more teaching scenarios to verify and extend the current conclusions. Additionally, combining more multimodal data (such as audio and text) for a more comprehensive analysis of teacher expressions and teaching effectiveness will also be an important direction for future research. Through these improvements, the research is expected to provide more comprehensive and in-depth support for enhancing the quality of online teaching and student learning experiences.

ACKNOWLEDGMENT

This paper was supported by: (1) Research on Industry-University Cooperation and Collaborative Education in Jilin Province Higher Education Teaching Reform (Research on Enhancing the Engagement of Online First-Class Course Learners through University-Enterprise Collaboration), (Grant No.: 20213F2MS1G00EE). (2) General Subject of Higher Education and Scientific Research of Jilin Province in 2022 (Research on the Promotion Strategy of 'New Normal' in Higher Education Online Education, (Grant No.: JGJX2022D44). (3) Research Project on Higher Education Teaching Reform by the National Ethnic Affairs Commission of the People's Republic of China (Grant No.: 21124).

REFERENCES

- [1] Hao, Z., Bin Yahya, M.Y., Lu, J. (2023). Influence of blockchain technology application in education on online teaching resources sharing. *International Journal of Emerging Technologies in Learning (Online)*, 18(11): 25-37. <https://doi.org/10.3991/ijet.v18i11.39361>
- [2] Shen, Y., Yang, X., Wang, L., Zheng, R. (2024). Research on the integration of online teaching resources in higher education institutions under the perspective of industry-education integration. *Applied Mathematics and Nonlinear Sciences*, 9(1). <https://doi.org/10.2478/amns.2023.2.00255>
- [3] Xue, Y., Li, N. (2024). Research and application of multimedia compression technology in online physical education teaching task. *Signal, Image and Video Processing*, 18(4): 3459-3470. <https://link.springer.com/article/10.1007/s11760-024-03012-8>.
- [4] Qiao, H. (2024). Personalized recommendations of online English teaching resources for higher vocational education. *Applied Mathematics and Nonlinear Sciences*, 9(1). <https://doi.org/10.2478/amns-2024-0723>
- [5] Gordillo, A., López-Fernández, D., Tovar, E. (2022). Comparing the effectiveness of video-based learning and game-based learning using teacher-authored video games for online software engineering education. *IEEE Transactions on Education*, 65(4): 524-532. <https://doi.org/10.1109/TE.2022.3142688>
- [6] Li, W., Hou, M. (2024). Evaluation method of teaching effect of online physical education based on fuzzy AHP. In *International Conference on E-Learning, E-Education, and Online Training*. Cham: Springer Nature Switzerland, pp. 207-226. https://doi.org/10.1007/978-3-031-51503-3_14.
- [7] He, Y. (2024). Evaluation and improvement of online teaching effectiveness of civil law in the context of data analysis. *Applied Mathematics and Nonlinear Sciences*, 9(1): 1-10. <https://doi.org/10.2478/amns.2023.2.00934>
- [8] Yu, H., Li, X. (2021). An evaluation model of English teaching effectiveness based on online education. *International Journal of Continuing Engineering Education and Life Long Learning*, 31(2): 218-233. <https://doi.org/10.1504/IJCEELL.2021.114389>
- [9] Guo, R. (2024). Analysis of artificial intelligence technology and its application in improving the effectiveness of physical education teaching. *International Journal of Web-Based Learning and Teaching Technologies (IJWLTT)*, 19(1): 1-15. <https://doi.org/10.4018/IJWLTT.335115>
- [10] Haq, H.B.U., Akram, W., Irshad, M.N., Kosar, A., Abid, M. (2024). Enhanced real-time facial expression recognition using deep learning. *Acadlore Transactions on AI and Machine Learning*, 3(1): 24-35. <https://doi.org/10.56578/ataiml030103>
- [11] Abdullah, M.U., Alkan, A. (2022). A comparative approach for facial expression recognition in higher education using hybrid-deep learning from students' facial images. *Traitement du Signal*, 39(6): 1929-1941. <https://doi.org/10.18280/ts.390605>
- [12] Lu, R., Ji, F. (2024). Design and implementation of a virtual teacher teaching system algorithm based on facial expression recognition in the era of big data. *Applied Mathematics and Nonlinear Sciences*, 9(1): 1-10. <https://doi.org/10.2478/amns.2023.2.00053a>
- [13] Zhong, H., Han, T., Xia, W., Tian, Y., Wu, L. (2023). Research on real-time teachers' facial expression recognition based on YOLOv5 and attention mechanisms. *EURASIP Journal on Advances in Signal Processing*, 2023(1): 55. <https://link.springer.com/article/10.1186/s13634-023-01019-w>.
- [14] Zhang, C. (2024). Research on the path of English translation teaching and external communication based on the multimodal analysis method. *Applied Mathematics and Nonlinear Sciences*, 9(1). <https://doi.org/10.2478/amns-2024-1143>
- [15] Al-Atroshi, S.J.A., Ali, A.M. (2023). Improving facial expression recognition using HOG with SVM and modified datasets classified by Alexnet. *Traitement du Signal*, 40(4): 1611-1619. <https://doi.org/10.18280/ts.400429>
- [16] Wu, S., Zhang, B., Cao, T. (2023). Relation between teacher autonomy support, student self-efficacy, and behavioral engagement: A moderated mediation model in project-based team learning. *International Journal of Engineering Education*, 39(6): 1464-1477.
- [17] Zhao, J., Li, J., Jia, J. (2021). A study on posture-based teacher-student behavioral engagement pattern. *Sustainable Cities and Society*, 67: 102749. <https://doi.org/10.1016/j.scs.2021.102749>
- [18] Durga, B.K., Rajesh, V., Jagannadham, S., Kumar, P.S., Rashed, A.N.Z., Saikumar, K. (2023). Deep learning-based micro facial expression recognition using an adaptive Tiefs FCNN model. *Traitement du Signal*, 40(3): 1035-1043. <https://doi.org/10.18280/ts.400319>
- [19] Deore, S.P. (2023). Enriching song recommendation through facial expression using deep learning. *Ingénierie des Systèmes d'Information*, 28(1): 225-229. <https://doi.org/10.18280/isi.280126>
- [20] Verma, N., Getenet, S., Dann, C., Shaik, T. (2023). Designing an artificial intelligence tool to understand student engagement based on teacher's behaviors and movements in video conferencing. *Computers and Education: Artificial Intelligence*, 5: 100187. <https://doi.org/10.1016/j.caeai.2023.100187>
- [21] Ladia, A.M., Landman, D., Peri, M., Jasim, M., Mahyar, N. (2023). Reimagining the virtual classroom: Enhancing engagement and student-teacher interaction in the digital age. In *ACM International Conference Proceeding Series*, pp. 384-386. <https://doi.org/10.1145/3591196.3596617>
- [22] Reda, N.H., Abbas, H.H. (2024). 3D human facial traits'

analysis for ethnicity recognition using deep learning. *Ingénierie des Systèmes d'Information*, 29(2): 501-514. <https://doi.org/10.18280/isi.290211>

[23] Sundaram, S.M., Narayanan, R. (2023). Human face and

facial expression recognition using deep learning and SNet architecture integrated with BottleNeck attention module. *Traitement du Signal*, 40(2): 647-655. <https://doi.org/10.18280/ts.400223>