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

Traitement du Signal
Vol. 40, No. 2, April, 2023, pp. 751-758
Journal homepage: http://iieta.org/journals/ts

Emotion Recognition in Learning Scenes Supported by Smart Classroom and Its Application

Zhen Zhu[1], Xiaoqing Zheng[2], Tongping Ke[3], Guofei Chai[4*]

[1] Digital Campus Construction Center, Quzhou College of Technology, Quzhou 324000, China
[2] College of Information Engineering, Quzhou College of Technology, Quzhou 324000, China
[3] Modern Business Research Center, Zhejiang Gongshang University, Hangzhou 310018, China
[4] College of Electrical and Information Engineering, Quzhou University, Quzhou 324000, China

Corresponding Author Email: chaig@qzc.edu.cn

https://doi.org/10.18280/ts.400235

Received: 12 January 2023
Accepted: 26 March 2023

Keywords:
smart classroom, learning scene, emotion recognition

ABSTRACT

Emotion recognition technology is one of the important applications of artificial intelligence and machine learning in the field of education. By recognizing the emotions of students in learning scenes, teachers can better understand the learning status of students and provide them with personalized learning resources and help. Current emotion recognition methods are mainly based on static facial emotions, neglecting the temporal features of facial emotions, which may lead to inaccurate recognition results. In order to overcome these challenges, this study conducts research on emotion recognition and its application in learning scenes supported by smart classrooms. The Transformer encoder is used to extract the temporal features of student facial emotions based on learning scenes, i.e., the self-attention module of the encoder is used to extract the temporal features of facial emotions in learning scenes. Residual attention networks, Transformers, and non-local neural networks are used to extract facial emotion features from different perspectives and levels. The combination of Vision-Transformer (ViT) and NetVLAD enables the model to learn the features of data from multiple perspectives, thereby improving the generalization ability of the model. The experimental results verify the effectiveness of the constructed model.

1. INTRODUCTION

In today's society, student-centered teaching methods are becoming mainstream, and personalized and contextualized learning are receiving increasing attention [1-5]. Therefore, it is necessary for teachers to understand and adjust students' learning emotions in a timely manner to achieve optimal learning outcomes [6-11]. The smart classroom is an important product of new education technology, which uses the latest computer, communication, and Internet technologies to achieve deep perception, intelligent analysis, and accurate delivery of the education process [12-17]. Among them, emotion recognition technology is one of the important applications of artificial intelligence and machine learning in the field of education [18-21]. By recognizing the emotions of students in learning scenes, teachers can better understand the learning status of students and provide them with personalized learning resources and help. Emotions are closely related to learning outcomes. By recognizing and adjusting students' emotions in real-time, their learning enthusiasm and effectiveness can be improved.

Emotion recognition of learners in the classroom plays an important role in improving classroom efficiency. At present, recognition methods based on traditional image processing generally have problems such as low recognition accuracy and difficulty in feature extraction. To effectively solve these problems, Su and Wang [22] proposed a deep learning-based learner emotion recognition method. This method first introduces the convolutional structure of MobileNet to replace the DarkNet-53 of YOLO v3, making the model lighter and reducing the number of parameters. The GIoU loss is then used to improve the loss function of the model. Experimental results show that the improved model's mAP is increased by 4%, the F1 score is increased by 3.2%, and the detection time is reduced by 1/3. Liang et al. [23] focused on teacher's voice signals and designs an emotion detection audio processing system. Teachers' speeches are used to determine their emotions. A speech emotion recognition classification model is built using the recursive neural network (RNN) algorithm. By improving the traditional Mel-frequency cepstral coefficient (MFCC) feature extraction process and adding a second-order differentiation process to eliminate MFCC convolution noise, one-dimensional energy features are added to the 39-dimensional MFCC coefficients for experiments. The experimental results show that the improved MFCC feature values and neural networks are more effective in improving the recognition rate of speech emotions than traditional speech emotion recognition methods and can be used for speech emotion recognition in classroom teaching. Putra and Arifin [24] constructs a real-time facial emotion recognition system, which allows teachers to monitor students' emotions through classroom activities. When running on mid-range computer specifications, the system should have sufficient reliability. Students will receive a questionnaire to measure their stress. The results of the questionnaire survey will be used to analyze whether the use of the system can alleviate students' stress. The results of the questionnaire survey show that the system can detect students' emotions

early, allowing teachers to minimize students' stress as much as possible.

Facial emotions are very complex and subtle signals, and people may display multiple facial emotions at the same time. In addition, different individuals may express the same emotion in different ways, making facial emotion recognition very difficult. Moreover, facial emotions are dynamic, and the initiation, persistence, and termination of expressions all contain rich emotional information. However, current emotion recognition methods are mainly based on static facial emotions, neglecting the temporal features of facial emotions, which may lead to inaccurate recognition results. In order to overcome these challenges, more advanced and accurate emotion recognition methods need to be developed. To this end, this study conducts research on emotion recognition and its application in learning scenes supported by smart classrooms.

2. EXTRACTION OF TEMPORAL FACIAL EMOTION FEATURES BASED ON LEARNING SCENES FOR STUDENTS

Traditional Recurrent Neural Networks (RNN) encounter the problem of gradient vanishing and explosion when processing long sequences, whereas the Transformer model leverages its self-attention mechanism to capture long-term dependencies in sequences, enabling the model to better understand and parse the dynamic changes in facial emotions. Moreover, compared to RNNs, which need to process each time step sequentially, Transformer models can handle all time steps simultaneously, greatly improving computational efficiency, which is crucial for processing a large amount of learning scene data. Based on the self-attention mechanism, the Transformer model can assign different weights according to the importance of facial expressions at different time points, which helps the model capture key emotional changes. Therefore, this study utilizes the Transformer encoder part to extract temporal facial emotion features of students based on learning scenes, that is, extracting temporal features of student facial emotions in learning scenes based on the encoder's self-attention module.

In the self-attention module, it is assumed that the input student facial emotion feature sequence is mapped through *InputEmbedding*, resulting in s_1 and s_2, which are then transformed through Q^w, Q^j, and Q^c matrices to obtain w^u, j^u, and c^u.

where, w represents the query, j represents the *key*, and c represents the information extracted from s. When the input is assumed to be s_1 and s_2, and the matrix parameter is Q^w, then:

$$w^1 = s_1 Q^w, w^2 = s_2 Q^w \tag{1}$$

Since the model can be parallelized, there is also:

$$\begin{pmatrix} w^1 \\ w^2 \end{pmatrix} = \begin{pmatrix} s_1 Q^w \\ s_2 Q^w \end{pmatrix} \tag{2}$$

Similarly, (j^1 j^2) and (c^1 c^2) can be obtained. In the self-attention mechanism, W is the obtained (w^1 w^2), (j^1 j^2) is J, and (c^1 c^2) is C. To further match w1 with each j, dot product operations are performed. Since the dot product will cause the gradient to become very small after processing by the *softmax* function, it needs to be scaled to obtain the corresponding β.

The calculation process is as follows:

$$\beta_{1,1} = \frac{w^1 j^1}{\sqrt{f}}, \beta_{1,2} = \frac{w^2 j^2}{\sqrt{f}} \tag{3}$$

Likewise, by matching w^2 with all j, $\beta_{2,u}$ can be obtained:

$$\begin{pmatrix} \beta_{1,1} & \beta_{1,2} \\ \beta_{2,1} & \beta_{2,2} \end{pmatrix} = \frac{\begin{pmatrix} w^1 \\ w^2 \end{pmatrix} \begin{pmatrix} j^1 \\ j^2 \end{pmatrix}^Y}{\sqrt{f}} \tag{4}$$

After applying the *softmax* function to each row of the matrix in the above equation, ($\hat{\beta}_{1,1}$,$\hat{\beta}_{2,1}$) and ($\hat{\beta}_{2,1}$,$\hat{\beta}_{2,2}$) can be obtained. By using $\hat{\beta}$ to weight each C, the output result of the self-attention mechanism can be obtained, that is:

$$ATT(W, J, C) = \Omega\left(\frac{WJ^Y}{\sqrt{f_j}}\right)C \tag{5}$$

Similarly, the multi-head self-attention module also transforms β_u through Q^w, Q^j, and Q^c to obtain w^u, j^u, and c^u. Figure 1 shows the architecture of the multi-head attention mechanism. The specific calculation is shown in the following formula.

$$ATT(W_u, J_u, C_u) = \Omega\left(\frac{W_u J_u^Y}{\sqrt{f_j}}\right)C_u \tag{6}$$

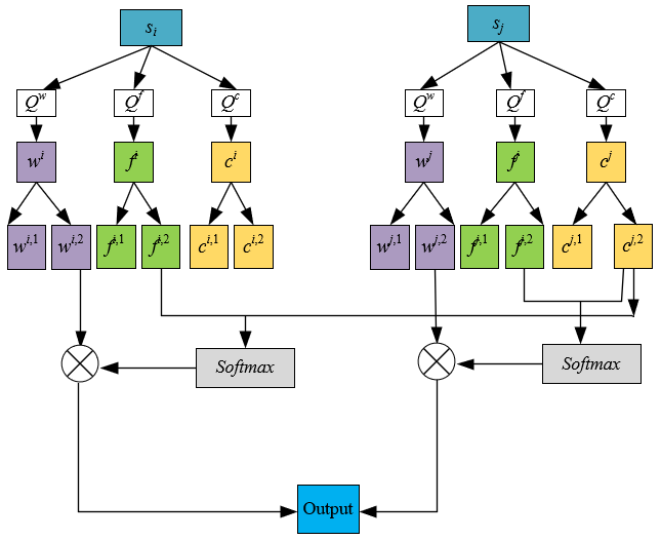

Figure 1. Multi-head attention mechanism architecture

When using this module to extract student facial emotion temporal sequence, it is assumed that the feature sequence learned by the *Vision-Transformer* is represented by λ^k_u. Input λ^k_u into the *Transformer* module to extract the temporal features of student facial emotions in the learning scene sequence. Then, the vector scores of several categories are obtained through the fully connected layer, represented by z_u.

Assuming that the score of the u-th class output by the Transformer module is represented by z_u, the probability vector output by the *softmax* function is represented by o_u, the number of labeled samples is represented by B, the

corresponding true labels are represented by t_u, and the cross-entropy loss of the probability vector o_u to t_u is represented by M_Y. Finally, input z_u into the *softmax* function and the cross-entropy loss function formula to obtain the loss value of the model.

$$\begin{cases} o_u = \dfrac{r^{z_u}}{\sum_{j=1}^{B} r^{z_j}} \\ M_Y = -\sum_{u=1}^{B} t_u LO(o_u) \end{cases} \tag{7}$$

3. FACE EMOTION RECOGNITION OF STUDENTS BASED ON LEARNING SCENARIOS

Existing facial emotion recognition algorithms have many limitations in both data collection and suppression of non-emotional features. Residual attention networks can extract deep-level features and capture subtle changes in images, which is essential for facial emotion recognition. *Transformers* can process time series data and capture dynamic changes in facial emotions. Non-local neural networks can capture long-range dependencies in images, which is particularly useful for understanding complex facial emotions. By integrating these three networks, facial emotion features can be extracted from different perspectives and levels. Figure 2 shows the overall architecture of the network model.

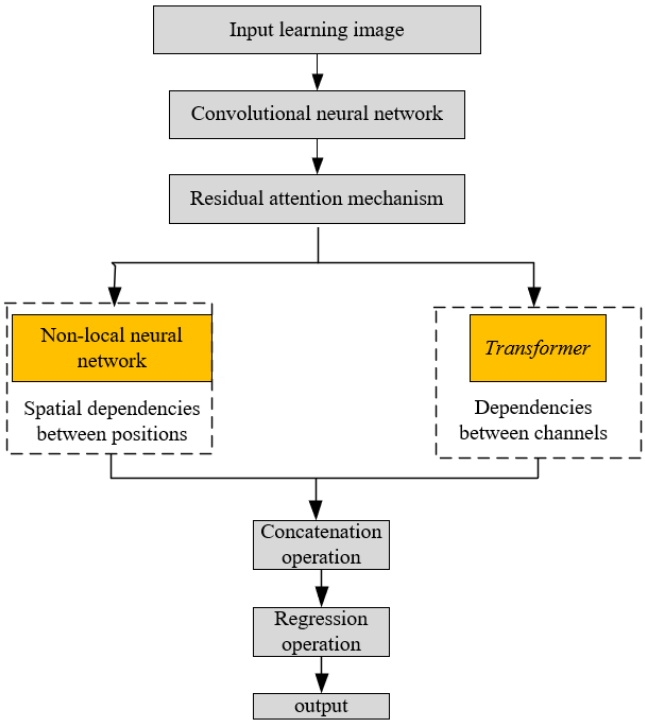

Figure 2. Network model architecture

The residual attention network consists of two main branches: the main branch and the mask branch. These two branches work together to extract useful emotion features from input images and learn attention weights. Figure 3 illustrates the architecture of the residual attention network.

The main branch is primarily responsible for extracting features from input images. It typically comprises a series of convolutional layers, activation functions, and pooling layers, forming a deep neural network structure. The design of the main branch mainly refers to the structure of the residual network (*ResNet*), in which each residual block contains multiple convolutional layers and a *shortcut* connection, effectively extracting deep-level features from images and preventing the vanishing gradient problem. Moreover, the main branch can extract different receptive field image features by changing the convolution kernel size and stride, thereby better understanding local and global information in the image.

A batch normalization layer and *ReLU* linear activation unit are set after each convolutional layer in the main branch. Assuming the output of the m-th convolutional layer is represented by x^m, the output after batch normalization is represented by p^m, the activation function is represented by $d(\cdot)$, and the weight values and biases are represented by Q and n, respectively, the linear activation unit calculation formula is given by:

$$p^m = d\left(x^m\right) = d\left(Qo^{m-1} + n\right) \tag{8}$$

The primary task of the mask branch is to learn the attention weights of input images, guiding the main branch to focus more on key regions in the image. The mask branch generally consists of lightweight convolutional layers and activation functions, and finally outputs attention weights for each pixel through a *sigmoid* function. These weights are applied to the feature map of the main branch to achieve dynamic attention adjustment at the feature level. The design of the mask branch enables the model to adaptively focus on more critical image regions for the task, thus improving the model's performance.

The attention feature map corresponding to the learning scenario can be obtained by performing dot product operations on the features and weights output by the two branches. Assuming that the input of the residual attention network is represented by z, the output of the main branch is represented by $Y_{u,v}(z)$, and the output of the mask branch is represented by $L_{u,v}(z)$, then:

$$G_{u,v}(z) = L_{u,v}(z) * Y_{u,v}(z) \tag{9}$$

To avoid affecting the transmission of extracted student facial emotion features, this study further adopts a residual connection-like approach to fuse the outputs of the two branches. The output expression of the residual attention network model is given by:

$$\left(1 + L_{u,v}(z)\right) * Y_{u,v}(z) \tag{10}$$

In facial emotion recognition tasks, high-dimensional features, such as the three channels of RGB images and feature maps from different convolutional layers, often need to be processed. *Feed-Forward Neural Networks (FFNN)* can effectively handle these high-dimensional features and extract useful information. By combining multi-head attention networks, feature dependencies can be modeled in high-dimensional spaces, thereby better understanding and recognizing facial emotions.

The workflow of the adopted multi-head attention network is given by:

$$ATT\left((J,C),W\right) = ATT\left((J,C),w_1\right) \oplus \cdots \oplus ATT\left((J,C),w_g\right) \tag{11}$$

Assuming the input features of the m-th layer are represented by x^m, the output features of the m-th layer are represented by p^m, the weight matrix of the m-th layer is represented by Q^m, and the bias of the X-th layer is represented by n^m. The iterative process of the feed-forward neural network is given by:

$$x^m = Q^m p^{m-1} + n^m \tag{12}$$

$$p^m = d^m\left(x^m\right) \tag{13}$$

In facial emotion recognition, global feature dependencies (e.g., interactions between eyes and mouth) are crucial for understanding overall facial emotions. Non-local neural networks calculate relationships between any two points, modeling feature dependencies globally, and aiding in extracting more complete emotional information. Since non-local neural networks can capture richer spatial information, they can significantly improve performance in facial emotion recognition tasks. Combined with the previous residual attention network and *Transformer* network, this allows the model to better understand complex facial emotions, thereby improving emotion recognition accuracy. Figure 4 shows the architecture of the non-local neural network. Assuming the input features are represented by z, the output for position u is represented by t_u, the normalization factor is represented by $V(z)$, and all feature positions required to be obtained for correlation calculation with u are represented by k. The definition of non-local operation is given by:

$$t_u = \frac{1}{V(z)}\sum_{\forall_k} d\left(z_u, z_k\right) h\left(z_k\right) \tag{14}$$

The correlation scores between u and all positions can be calculated through d (z_u, z_k), and the input representation of k can be calculated through $h(z_k)$.

In this study, the correlation score between u and k in the feature map is calculated by taking Q and z as the inputs of the non-local neural network, denoted as $d(z_u, z_k)=(Q_\phi z_u)^Y(Q_\rho z_k)$. The detailed calculation process of the non-local neural network is described as follows:

Step 1: Perform 1×1 convolution on the feature map, and map the z result to the ϕ, ρ, and $h3$ spaces.

Step 2: Calculate the correlation score between u and k, using the formula $d(z_u, z_k)=(Q_\phi z_u)^Y(Q_\rho z_k)$.

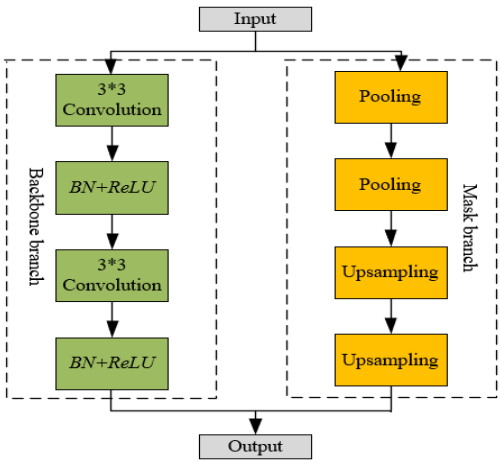

Figure 3. Residual attention network architecture

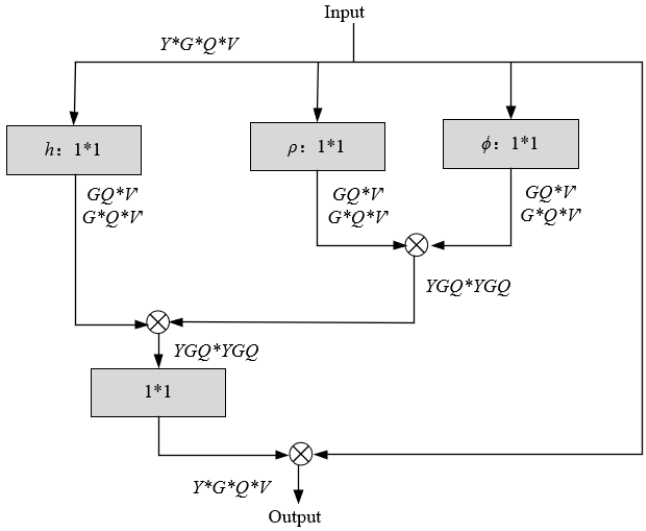

Figure 4. Non-local neural network architecture

Step 3: Perform a weighted summation of the correlation score and the feature representation function $h(z_k)$ of k to obtain the learning result of the dependency relationship between u and k.

4. OVERALL MODEL AND UPDATE ITERATION PROCESS

In this study, *Vision-Transformer* (*ViT*) is used as the main framework. *ViT* is an image processing model designed to extract spatial features from images. The *NetVLAD* module is also introduced to obtain more valuable features from images through clustering methods. Finally, the *Transformer* module is used to learn the temporal features of image sequences, which further extracts features in the time dimension. This allows for comprehensive extraction of spatial-temporal features of the learning screen, resulting in more accurate emotion recognition of students. The combination of *ViT* and *NetVLAD* enables the model to learn the features of the data from multiple perspectives, thereby improving the generalization ability of the model and enabling it to perform well on unseen data. This can be represented by the following formula:

$$M_c = M_B + \eta M_Y + \frac{\eta_2}{2}\|q\| \tag{15}$$

As can be seen from the above formula, the unified loss function M_c includes *ViT* loss M_B and *NetVLAD* loss M_Y. η_1 and η_2 are hyperparameters used to balance M_Y and M_B, respectively.

In order to improve the robustness of the model and maintain high recognition accuracy when facing different environments and individuals, this study uses *StarGAN* to generate images of students in peak frames of learning screen sequences. The images generated by *StarGAN* can increase data diversity, which can be seen as a data augmentation method. Data augmentation can improve the generalization ability of the model, making it perform better on unseen data. Finally, the generated images and peak frame images are mixed together and input into the *ViT* network, allowing the model to understand and learn facial emotions from different perspectives, which helps to improve the performance of the

model in facial emotion recognition tasks.

At this point, the mixed input image is represented by $z^{l}_{u} \in Z$ and the label is represented by $t_u \in T$. Assuming that the probability vector output by the *softmaxSM* function is represented by o_u, the corresponding true label is represented by t_u, the number of labels is represented by B, and the cross-entropy loss M_H of the probability vector o_u with respect to t_u is represented. The main framework network then calculates the loss using the *softmax* function and the cross-entropy function, as follows:

$$M_H = -\sum_{u=1}^{B} t_u \log(o_u) \tag{16}$$

Finally, the overall loss function of the model includes three parts: M_B, M_Y, and M_H, as follows:

$$M = M_H + M_B + \eta_1 M_Y + \frac{\eta_2}{2}\|q\| \tag{17}$$

5. EXPERIMENTAL RESULTS AND ANALYSIS

Figure 5 shows the comparison of emotion recognition accuracy before and after introducing the NetVLAD module. It can be seen that after introducing the NetVLAD module, the recognition accuracy of all emotion categories has improved. For "fear" recognition, the accuracy increased by 0.02 after introducing the NetVLAD module, from 0.71 to 0.73. For "anger" recognition, the accuracy increased by 0.02 after introducing the NetVLAD module, from 0.68 to 0.7. For "disgust" recognition, the accuracy increased by 0.03 after introducing the NetVLAD module, from 0.61 to 0.64. For "happy" recognition, the accuracy increased by 0.08 after introducing the NetVLAD module, from 0.89 to 0.97. This is the largest improvement among all emotion categories. For "sad" recognition, the accuracy increased by 0.06 after introducing the NetVLAD module, from 0.65 to 0.71. For "positive" recognition, the accuracy increased by 0.015 after introducing the NetVLAD module, from 0.695 to 0.71. For "neutral" recognition, the accuracy increased by 0.06 after introducing the NetVLAD module, from 0.84 to 0.9. From these results, it can be seen that after introducing the NetVLAD module, the recognition accuracy of all emotion categories has improved, especially for "happy" recognition, which has the largest improvement. This indicates that the NetVLAD module plays a positive role in feature extraction, effectively enhancing the model's emotion recognition capabilities. This result also validates the advantage of the NetVLAD module in obtaining more valuable features through its clustering method, thus improving the model's recognition capabilities.

Furthermore, this study analyzes the changes in RMSE for the emotion recognition results in terms of emotional arousal and emotional valence as the number of iterations increases. As can be seen from Figure 6, the RMSEs for both emotional arousal and emotional valence decrease during the model's iteration process and stabilize after approximately 30 iterations. For emotional arousal, the RMSE decreases from the initial 0.75 to the final 0.42, indicating that the model's prediction error for emotional arousal is gradually decreasing and the model's performance is improving. For emotional valence, the RMSE decreases from the initial 1.68 to the final 0.49, also indicating that the model's prediction error for emotional valence is gradually decreasing and the model's performance is improving. In addition, the decrease in emotional valence error is larger than that of emotional arousal, indicating that the model performs better in predicting emotional valence than emotional arousal. Overall, as the number of iterations increases, the model's prediction error gradually decreases, demonstrating good learning performance. Particularly after 30 iterations, the error basically stabilizes, indicating that the model has converged and further training cannot significantly improve the model's performance. This result also confirms the effectiveness of our previous strategy of introducing the NetVLAD module and using the StarGAN and ViT networks for facial emotion recognition.

Similar conclusions can be drawn from observing the consistency correlation coefficient change curve. With increasing iterations, the model's prediction consistency for both emotional arousal and emotional valence is improved, indicating good learning performance of the model (Figure 7). However, it should be noted that even though the consistency correlation coefficients have increased, they have not reached 1, indicating a certain gap between the model's prediction results and the actual labels, leaving room for further improvement in the model. These results further prove that the strategy of introducing the NetVLAD module and using StarGAN and ViT networks for facial emotion recognition can effectively improve the model's prediction consistency, thereby enhancing the model's performance.

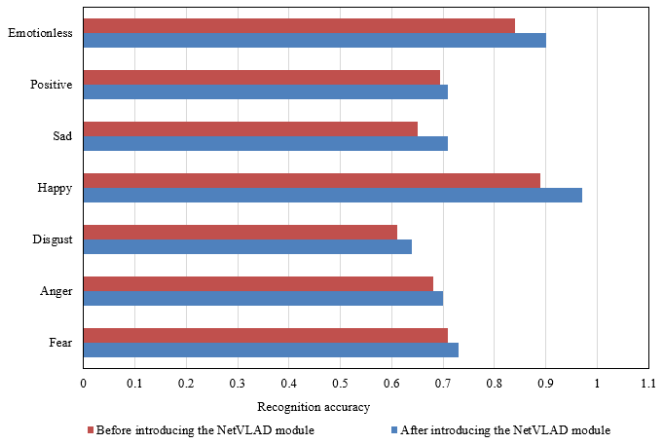

Figure 5. Comparison of emotion recognition accuracy before and after introducing the NetVLAD module

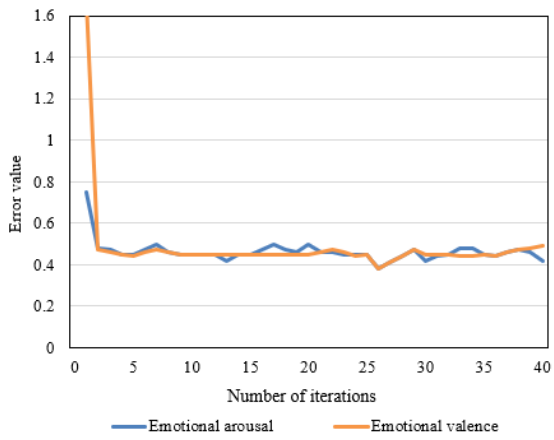

Figure 6. Root Mean Square Error (RMSE) change curve

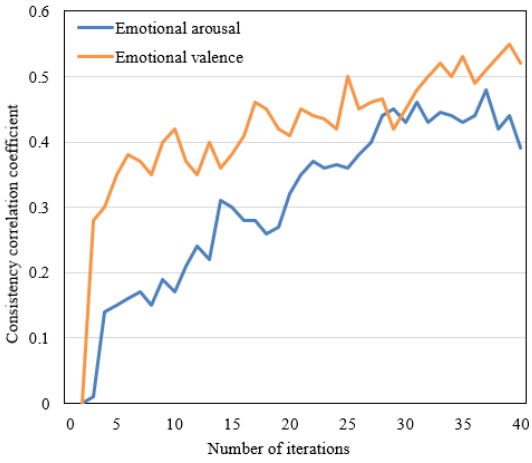

Figure 7. Consistency correlation coefficient change curve

Table 1. Performance comparison of different network models

Network Model	RMSE		Consistency Correlation Coefficient	
	Emotional Arousal	Emotional Valence	Emotional Arousal	Emotional Valence
CNN-GRU	0.501	0.378	0.189	0.587
LBVCNN	0.411	0.389	0.332	0.645
Enhanced ConvLSTM	0.439	0.376	0.335	0.679
FAN	0.424	0.359	0.367	0.733
Ours	0.387	0.356	0.511	0.454

Table 2. Recognition accuracy comparison of different models

Network Model	Accuracy	Input Method
CNN-GRU	88.62%	Learning Frame Sequence
LBVCNN	88.72%	Learning Frame Sequence
Enhanced ConvLSTM	84.23%	Learning Frame Sequence
FAN	85.28%	Learning Frame Sequence
Our Model	91.93%	Learning Frame Sequence

Table 1 presents the performance comparison results of different network models. It can be seen how each model performs in terms of RMSE and consistency correlation coefficient for emotional arousal and emotional valence. Overall, our model (Ours) outperforms or is equal to the other four models in terms of these four indicators. In terms of RMSE for emotional arousal, our model (0.387) is lower than the other models (0.501, 0.411, 0.439, 0.424), indicating higher accuracy in predicting emotional arousal for our model. In terms of RMSE for emotional valence, our model (0.356) is also the lowest, further highlighting its superior accuracy. In terms of consistency correlation coefficient for emotional arousal, our model (0.511) is significantly higher than the other models (0.189, 0.332, 0.335, 0.367), indicating stronger consistency in predicting emotional arousal. In terms of consistency correlation coefficient for emotional valence, although our model (0.454) is slightly lower than the FAN model (0.733), it still outperforms the other three models (0.587, 0.645, 0.679), indicating that our model still has a good performance in predicting emotional valence consistency.

Table 2 compares the recognition accuracy of different models. As shown in the table, the recognition accuracy of each model when processing learning scene sequences is presented. Overall, the accuracy of the "Proposed Model" is the highest, reaching 91.93%, which is significantly higher than the other four models. Specifically, the accuracy of the CNN-GRU model is 88.62%, ranking third among the five models. The accuracy of the LBVCNN model is slightly higher at 88.72%, ranking second among the five models. The accuracy of the Enhanced ConvLSTM model is 84.23%, ranking fourth among the five models. The accuracy of the FAN model is 85.28%, ranking fifth among the five models. Based on this data, the following conclusions can be drawn. The accuracy of all models in processing learning scene sequences reaches a high level, indicating that these models can effectively recognize and understand learning scene sequences to some extent. However, the performance of the Proposed Model is the most outstanding, with an accuracy of 91.93%, which is significantly higher than other models. This is because the Proposed Model can better understand and process the complexity and diversity of learning scene sequences, thus providing more accurate recognition results. These results further demonstrate that the Proposed Model has a significant advantage in the emotion recognition task of learning scene sequences.

Three scenarios are set for this experiment: 1) Scenario 1: This is the beginning of a new semester when the teacher introduces a new and interesting topic. The teacher is energetic and explains the topic in an easily understandable way. Students are highly engaged, curious, and excited about the new knowledge. There is frequent interaction between the teacher and students, and the classroom atmosphere is lively. Students actively participate in discussions and activities, demonstrating good learning outcomes. 2) Scenario 2: This is when the teacher teaches relatively difficult or dull knowledge. Although the teacher tries to explain in a lively and interesting way, students' reactions are not as positive as in Scenario 1, because they find it difficult to understand or lack interest. However, overall, students are still trying to keep up, and the classroom atmosphere and teaching effectiveness are still good. 3) Scenario 3: This is in the middle of the course, dealing with repetitive, low-difficulty but memory-demanding knowledge. The teacher appears slightly tired due to the repetitive content, and students become somewhat bored due to the dull or repetitive nature of the content, resulting in decreased engagement. At this point, the classroom atmosphere and teaching effectiveness are relatively average, requiring a change in teaching methods or the addition of interactive elements to boost engagement.

Table 3 provides the emotion recognition results for the 3 scenarios. The changes in the facial activity of teachers and students, as well as the classroom emotional activity and teaching effectiveness, can be observed. In Scenario 1, the teacher's facial activity is the highest, at 0.7583, and the positive facial activity of students 1, 2, and 3 is also relatively high, at 0.6475, 0.5684, and 0.7853, respectively, while their negative facial activity is low, indicating that they are in a relatively positive state during learning. The overall classroom emotional activity is 0.6753, which is quite high, indicating a lively classroom atmosphere, frequent interaction, and a "Good" evaluation of teaching effectiveness. In Scenario 2, the teacher's facial activity decreases to 0.5637, and students' positive facial activity also declines, but their negative facial activity increases, especially for students 1 and 2, with negative facial activity rising significantly to 0.5338 and 0.5642, indicating that students are encountering difficulties or challenges in their learning. The overall classroom emotional

activity decreases to 0.3826, and the classroom atmosphere is somewhat suppressed or the teaching content is more challenging, but the evaluation of teaching effectiveness remains "Good". In Scenario 3, the teacher's emotional expression activity further decreased to 0.3875, and the students' positive emotional expression activity also decreased. The negative emotional expression activity increased, especially for student 3, whose negative emotional expression activity reached 0.7836, which was very high. This indicates that the students encountered significant difficulties or learning challenges. The classroom emotional activity dropped to 0.1846, and the teaching effect evaluation was "Average." This suggests that there were some issues with classroom teaching at this stage, and the teacher needed to adjust teaching methods or strategies promptly. Overall, emotion recognition based on learning scenes effectively reflects teaching situations in the classroom, helping teachers understand students' learning status in real-time, and adjusting teaching strategies to improve teaching results.

Table 3. Emotion recognition results in 3 scenarios

		Scenario 1	Scenario 2	Scenario 3
Teacher	Facial Activity	0.7583	0.5637	0.3875
Student 1	Positive Facial Activity	0.6475	0.4826	0.3902
	Negative Facial Activity	0.3246	0.5338	0.6383
Student 2	Positive Facial Activity	0.5684	0.4985	0.3784
	Negative Facial Activity	0.4531	0.5642	0.6284
Student 3	Positive Facial Activity	0.7853	0.5829	0.2946
	Negative Facial Activity	0.2345	0.4726	0.7836
Classroom Emotional Activity		0.6753	0.3826	0.1846
Teaching Effectiveness Evaluation		*Good*	*Good*	*Average*

Table 4 presents the emotional evaluation results for the three scenarios. In these three typical scenarios, we can observe the changes in the emotional evaluations of teachers, students, and the overall classroom, as well as the teaching effect. In Scenario 1, both the teacher and students' emotional evaluations were considered "active," the classroom emotional evaluation was "excellent," and the teaching effect evaluation was also "excellent." This means that, in this environment, teachers were enthusiastic about teaching, students participated actively in learning, the classroom atmosphere was positive and lively, and the teaching effect was outstanding. In Scenario 2, the teacher's emotional evaluation remained "active," but the students' emotional evaluation dropped to "normal," and both the classroom emotional evaluation and teaching effect evaluation were considered "good." This means that although the teacher continued to teach actively, the students' learning status and participation decreased due to the difficulty of the teaching content or other reasons, the classroom atmosphere and teaching effect declined compared to Scenario 1, but overall, they were still good. In Scenario 3, both the teacher and students' emotional evaluations were "normal," and the classroom emotional evaluation and teaching effect evaluation were also "average." This indicates that in this scenario, both the teacher's teaching enthusiasm and the students' learning enthusiasm decreased, and the classroom atmosphere and teaching effect were mediocre. In summary, these three scenarios reveal the close relationship between emotional states and teaching effects. The emotional states of teachers and students directly affect the classroom atmosphere, which in turn affects the teaching effect. Therefore, teachers should pay attention to and manage their own and students' emotional states to improve teaching effects.

Table 4. Emotional evaluation results for 3 scenarios

	Teacher Emotional Evaluation	Student Emotional Evaluation	Classroom Emotional Evaluation	Teaching Effect Evaluation
Scenario 1	Active	Active	Excellent	Excellent
Scenario 2	Active	Normal	Good	Good
Scenario 3	Average	Average	Average	Average

6. CONCLUSION

In conclusion, this study investigates emotion recognition in learning scenes supported by smart classrooms. The Transformer encoder is utilized to extract temporal features of students' facial emotions based on learning scenes, i.e., the self-attention module of the encoder extracts temporal features of students' facial emotions in learning scenes. The combination of residual attention networks, Transformer, and non-local neural networks achieves the extraction of facial emotion features from different perspectives and levels. The combination of Vision-Transformer (ViT) and NetVLAD enables the model to learn data features from multiple angles, thereby improving the model's generalization ability. The study first compared the recognition accuracy of each emotion before and after the introduction of the NetVLAD module, verifying the advantage of the NetVLAD module in obtaining more valuable features through its clustering method and enhancing the model's recognition ability. The root mean square error and consistency correlation coefficient of the emotion arousal and emotion valence indicators in the emotion recognition results were analyzed with the change of iterations. This further confirmed that the introduction of the NetVLAD module and the use of StarGAN and ViT networks for facial emotion recognition could effectively improve the model's predictability and performance. The performance comparison results of different network models were provided, verifying that the proposed model still had good performance in predicting emotion valence consistency. Three scenarios were set for the experiment, and the emotion recognition results of the three scenarios were given, verifying that emotion recognition based on learning scenes could effectively reflect teaching situations in the classroom, helping teachers understand students' learning status in real-time, adjusting teaching strategies, and improving teaching effects.

ACKNOWLEDGMENT

This paper was supported by Guiding Project of Science and Technology Plan in Quzhou (Grant No.: 2023ZD146); and Scientific Research Project of Quzhou College of Technology (Grant No.: QZYY2113).

REFERENCES

[1] Wang, H.B., Zhou, J.T., Hu, C.L., Chen, W.W. (2022). Vehicle lateral stability control based on stability category recognition with improved brain emotional learning network. IEEE Transactions on Vehicular Technology, 71(6): 5930-5943. https://doi.org/10.1109/TVT.2022.3159271

[2] Zhang, H.Z., Yin, J.B., Zhang, X.L. (2020). The study of a five-dimensional emotional model for facial emotion recognition. Mobile Information Systems, 2020: Article ID: 8860608. https://doi.org/10.1155/2020/8860608

[3] ElBedwehy, M.N., Behery, G.M., Elbarougy, R. (2020). Emotional speech recognition based on weighted distance optimization system. International Journal of Pattern Recognition and Artificial Intelligence, 34(11): 2050027. https://doi.org/10.1142/S0218001420500275

[4] Xu, Z.Q., He, J.L., Liu, Y.Z. (2020). Exploration of emotional pattern recognition and emotional model construction based on big data. International Conference on Frontier Computing: Theory, Technologies and Applications, 982-992.

[5] Yang, H.S., Fan, Y.Y., Lv, G.Y., Liu, S.Y., Guo, Z. (2023). Exploiting emotional concepts for image emotion recognition. The Visual Computer, 39(5): 2177-2190. https://doi.org/10.1007/s00371-022-02472-8

[6] Sheng, W.J., Lu, X.Y., Li, X.D. (2023). Data augmentation by separating identity and emotion representations for emotional gait recognition. Robotica, 41(5): 1452-1465. https://doi.org/10.1017/S0263574722001813

[7] Qasim, M., Habib, T., Urooj, S., Mumtaz, B. (2023). DESCU: Dyadic emotional speech corpus and recognition system for Urdu language. Speech Communication, 148: 40-52. https://doi.org/10.1016/j.specom.2023.02.002

[8] Tropmann-Frick, M. (2023). Sign language recognition by similarity measure with emotional expression specific to signers. Frontiers in Artificial Intelligence and Applications, Information Modelling and Knowledge Bases XXXIV, 364: 21-37. https://doi.org/10.3233/FAIA220490

[9] Banskota, N., Alsadoon, A., Prasad, P.W.C., Dawoud, A., Rashid, T.A., Alsadoon, O.H. (2023). A novel enhanced convolution neural network with extreme learning machine: facial emotional recognition in psychology practices. Multimedia Tools and Applications, 82(5): 6479-6503. https://doi.org/10.1007/s11042-022-13567-8

[10] Zhou, Z.J., Asghar, M.A., Nazir, D., Siddique, K., Shorfuzzaman, M., Mehmood, R.M. (2023). An AI-empowered affect recognition model for healthcare and emotional well-being using physiological signals. Cluster Computing, 26(2): 1253-1266. https://doi.org/10.1007/s10586-022-03705-0

[11] Chen, K.Y., Yang, X., Fan, C.J., Zhang, W., Ding, Y. (2022). Semantic-rich facial emotional expression recognition. IEEE Transactions on Affective Computing, 13(4): 1906-1916. https://doi.org/10.1109/TAFFC.2022.3201290

[12] Sutedja, I., Septhia, J. (2022). Emotional expression recognition: A systematic literature review. In 2022 International Conference on Information Management and Technology (ICIMTech), pp. 184-188. https://doi.org/10.1109/ICIMTech55957.2022.9915210

[13] Ryumina, E., Ivanko, D. (2022). Emotional speech recognition based on lip-reading. In Speech and Computer: 24th International Conference, SPECOM 2022, Gurugram, India, pp. 616-625. https://doi.org/10.1007/978-3-031-20980-2_52

[14] Jiang, H.P., Jia, J.J. (2020). Research on EEG emotional recognition based on LSTM. In Bio-inspired Computing: Theories and Applications: 14th International Conference, BIC-TA 2019, Zhengzhou, China, pp. 409-417. https://doi.org/10.1007/978-981-15-3415-7_34

[15] Deng, H.X., Qian, G.Y., Zhang, Y.F., Hu, C.X., Liu, Y., Li, H.F. (2020). Emotional analysis and recognition based on EEG brain network. In 2020 Eighth International Conference on Advanced Cloud and Big Data (CBD), Taiyuan, China, pp. 56-61. https://doi.org/10.1109/CBD51900.2020.00019

[16] Akçay, M.B., Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication, 116: 56-76. https://doi.org/10.1016/j.specom.2019.12.001

[17] Cui, Z.L., Zhao, Y.J., Guo, J., Du, H.B., Zhang, J.J. (2020). Typical and reverse other-race effect on Tibetan students' emotional face recognition. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, pp. 754-760. https://doi.org/10.1109/FG47880.2020.00031

[18] Saeki, K., Kato, M., Kosaka, T. (2020). Language model adaptation for emotional speech recognition using Tweet data. In 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Auckland, New Zealand, pp. 371-375.

[19] Zhang, F., Li, X.C., Lim, C.P., Hua, Q., Dong, C.R., Zhai, J.H. (2022). Deep emotional arousal network for multimodal sentiment analysis and emotion recognition. Information Fusion, 88: 296-304. https://doi.org/10.1016/j.inffus.2022.07.006

[20] Dindar, M., Järvelä, S., Ahola, S., Huang, X., Zhao, G. (2020). Leaders and followers identified by emotional mimicry during collaborative learning: A facial expression recognition study on emotional valence. IEEE Transactions on Affective Computing, 13(3): 1390-1400. https://doi.org/10.1109/TAFFC.2020.3003243

[21] Fahad, M.S., Singh, S., Ranjan, A., Deepak, A. (2022). Emotion recognition from spontaneous speech using emotional vowel-like regions. Multimedia Tools and Applications, 81(10): 14025-14043. https://doi.org/10.1007/s11042-022-12453-7

[22] Su, C., Wang, G. (2020). Design and application of learner emotion recognition for classroom. In Journal of Physics: Conference Series, 1651(1): 012158. https://doi.org/10.1088/1742-6596/1651/1/012158

[23] Liang, J., Zhao, X.Y., Zhang, Z.H. (2020). Speech emotion recognition of teachers in classroom teaching. In 2020 Chinese Control and Decision Conference (CCDC), Hefei, China, pp. 5045-5050. https://doi.org/10.1109/CCDC49329.2020.9164823

[24] Putra, W.B., Arifin, F. (2019). Real-time emotion recognition system to monitor student's mood in a classroom. In Journal of Physics: Conference Series, 1413(1): 012021. https://doi.org/10.1088/1742-6596/1413/1/012021