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

Traitement du Signal
Vol. 41, No. 1, February, 2024, pp. 391-399
Journal homepage: http://iieta.org/journals/ts

The Application and Optimization of Deep Learning in Recognizing Student Learning Emotions

Zheng Wu[1], Dandan Pan[2*]

[1] School of Marxism, Fuyang Normal University, Fuyang 236037, China
[2] School of Foreign Languages, Nanjing Xiaozhuang University, Nanjing 211171, China

Corresponding Author Email: 2022091@njxzc.edu.cn

https://doi.org/10.18280/ts.410133

Received: 13 September 2023
Revised: 28 December 2023
Accepted: 9 January 2024
Available online: 29 February 2024

Keywords:
deep learning, learning emotion recognition, facial expression image preprocessing, temporal expression recognition, attention mechanism

ABSTRACT

With the widespread application of deep learning technology across various fields, its potential value in educational technology, particularly in recognizing student learning emotions, has begun to gain attention. The real-time and accurate identification of learning emotions is crucial for facilitating personalized teaching and enhancing learning efficiency. This paper focuses on the automatic recognition of student learning emotions based on deep learning technology, aiming to improve the accuracy and practicality of recognition by optimizing the preprocessing of facial expression images and the temporal expression recognition model. The research starts with facial detection using *Haar-like* features and the *Adaboost* cascade method, followed by normalization of the detected facial images in scale, angle, and grayscale to enhance the system's robustness to facial image variations. Subsequently, a temporal expression recognition model based on a multi-attention fusion network is proposed. This model utilizes both shallow and deep features of deep learning, along with the prior knowledge of Facial Action Coding System (FACS), to capture the dynamic changes in facial expressions more intricately. Finally, by introducing three different attention mechanisms, this study significantly improved the efficiency and accuracy of emotion feature recognition in sequential data. The findings of this paper not only advance the technology of learning emotion recognition but also provide valuable insights for educational practice.

1. INTRODUCTION

With the rapid development of artificial intelligence technology, deep learning has been widely applied in numerous fields such as image processing, natural language understanding, and speech recognition [1-3]. In the field of education, deep learning technology also shows its immense potential, especially in understanding student learning emotions, providing new avenues for achieving personalized teaching and optimizing learning experiences [4, 5]. Student learning emotions are closely related to their academic performance, learning motivation, and mental health, making it crucial to accurately identify and respond to students' emotional states [6, 7]. Traditional methods of emotion recognition rely on students' self-reports or teachers' subjective judgments, which often suffer from delays and inaccuracies. Deep learning technology, particularly facial expression recognition, offers the possibility of providing a real-time, objective tool for emotion analysis.

The automatic recognition of student learning emotions is essential for creating supportive and adaptive learning environments [8-10]. By monitoring students' emotional states, teachers can adjust teaching strategies in a timely manner to better meet students' personalized needs [11, 12]. At the same time, a student emotion recognition system can also help students self-regulate their learning processes, enhancing learning efficiency [13, 14]. Moreover, the application of this technology can promote the development of distance education and intelligent educational platforms, ensuring that students' emotions receive timely and appropriate attention in any teaching environment.

Despite significant progress in emotion recognition through deep learning, existing research methods still have some shortcomings. Firstly, student facial expression data are often affected by lighting, posture, and individual differences, which can reduce recognition accuracy [15-17]. Secondly, traditional deep learning models lack effective utilization of temporal information, making it difficult to accurately capture the dynamic changes in emotions [18]. Furthermore, existing methods are not mature enough in the application of attention mechanisms, failing to fully exploit emotional features in sequential data.

The main content of this paper is to optimize the method for recognizing student learning emotions. Firstly, this study improves the effect of facial expression image preprocessing through facial detection based on *Haar-like* features and *Adaboost* cascade method, followed by subsequent normalization processing, including scale normalization, angle normalization, and grayscale normalization. Secondly, a novel student temporal expression recognition model based on a multi-attention fusion network is proposed, effectively combining *FACS* prior knowledge, shallow features, and deep features to more accurately capture the dynamic sequence of facial expressions. Lastly, the paper introduces three different attention mechanisms, optimizing the model's ability to integrate features in processing sequential data, thereby significantly enhancing the accuracy and efficiency of student emotion recognition. Through these studies, this paper not only contributes new theories and practices to the technological development of learning emotion recognition but also provides powerful tools for optimizing the educational process and enhancing the learning experience.

2. STUDENT FACIAL EXPRESSION IMAGE PREPROCESSING

In the research on recognizing student learning emotions, facial expression image preprocessing is a crucial step as it directly affects the training effectiveness and recognition accuracy of subsequent models. Facial images in practical applications are often disturbed by various factors, such as changes in lighting conditions, different postures of students' faces, and differences in the resolution of image capture devices. These factors can result in varying quality of the original image data, affecting the deep learning model's ability to extract facial expression features. To address these issues, this paper uses *Haar-like* features and the *Adaboost* cascade method for face detection, ensuring the model focuses on the key areas of facial expressions. Scale normalization, angle normalization, and grayscale normalization are employed to ensure a consistent reference standard across different images, thus reducing the negative impact of variation factors on model performance, and enhancing the robustness and generalization ability of the recognition system.

2.1 Face detection

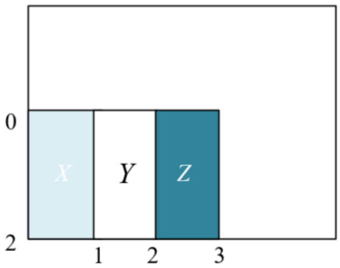

Figure 1. Example of *Haar-like* rectangular feature calculation

This paper first utilizes integral images to obtain *Haar-like* features of student faces. Specifically, it iterates over every pixel point on the student's face, calculating the cumulative sum of all pixel values within the rectangular area defined from the origin of the image to that point. This process is essentially a dynamic programming operation, where the integral value of each point is quickly calculated based on its adjacent values to the left and above. Such a calculation strategy significantly reduces redundant operations since the value of each pixel point needs to be calculated only once. Afterwards, these cumulative sums are stored in a numerical matrix of the same size as the original image, corresponding to their coordinates in the original image. This numerical matrix is the integral image, allowing any new area sum to be quickly computed through table lookup without the need to reiterate over the pixel points.

Further, utilizing the constructed integral image to quickly calculate *Haar-like* feature values, i.e., the sum of pixels within a rectangular area in the image, can be simply achieved through the cumulative sums of the four corners in the integral image. Specifically, this feature value can be obtained by performing a small number of addition and subtraction operations on the integral image values of the four corners of the rectangle, thus avoiding direct summation of every pixel value in the area and significantly improving the efficiency of feature extraction. Taking the calculation of *Haar-like* rectangular features shown in Figure 1 as an example, suppose the coordinate of any point in the image is represented by (u,k), and the sum of pixels in the rectangular area formed from the origin to (u,k) is represented by $SUM(u,k)$, then the sum of pixel values in area X can be represented as $SUM(1,2)$, the sum of pixel values in area Y can be represented as $SUM(2,2)$-$SUM(1,2)$, and the sum of pixel values in area Z can be represented as $SUM(3,2)$-$SUM(1,2)$. The formula for calculating the *Haar-like* rectangular feature value DS is given below:

$$DS=-2SUM\left(1,2\right)+2SUM\left(2,2\right)-SUM\left(3,2\right) \tag{1}$$

This paper further carries out student face detection based on the *Adaboost* cascade, as shown in Figure 2, which includes the following five steps:

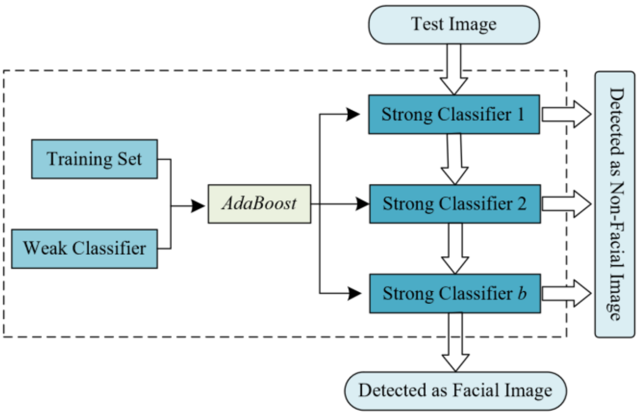

Figure 2. Flowchart of student face detection based on *Adaboost*

(1) Initially, the *Haar-like* feature values of facial expression images are calculated using the integral image method. This process simplifies the feature calculation for each window by converting the original image into an integral image, allowing for the rapid calculation of *Haar-like* features at every possible position and size in the image.

(2) To handle the large number of calculated *Haar-like* feature values, feature selection techniques are employed to reduce the dimensionality of features. In this step, the most informative features that best represent the key areas of the face are selected from thousands of features.

(3) The selected *Haar-like* features are used to train a large number of weak classifiers. In this step, each weak classifier is trained to recognize a specific feature of the face. Although each classifier performs modestly, they provide the foundation for building the final strong classifier.

(4) Through the *Adaboost* iterative algorithm, the best-performing weak classifiers are filtered out and combined

using a linear weighting method to form a strong classifier. Within the framework of *Adaboost*, each weak classifier is assigned a weight value, reflecting its importance in the classification process.

(5) The constructed strong classifiers are cascaded to form a more complex classifier network. In the cascading process, different strong classifiers are arranged in a certain order, with each classifier responsible for filtering out a certain percentage of non-facial areas, and the remaining areas are further detected by the next classifier.

When classifying student faces, it is particularly important to correctly set the threshold for each layer of strong classifiers to effectively filter out non-facial areas and accurately retain facial features within the cascaded structure. This strategy reduces computation by quickly eliminating most obvious non-facial areas in the early layers, while gradually raising the detection standards with increasing cascade depth, facing fewer but more likely facial candidate areas in subsequent layers, thus enhancing detection efficiency and reducing false positives. To achieve this goal, threshold settings must be based on precise performance evaluation to ensure that real faces are not missed due to low thresholds, nor is unnecessary computational burden added due to high thresholds. Furthermore, in practical applications, it may also be necessary to consider changes in the detection environment (such as lighting, facial obstructions) and the diversity of facial features among different students, making appropriate adjustments and optimizations to ensure the robustness and high accuracy of the face detection system in practical applications, thereby providing efficient and reliable input for deep learning models to recognize student learning emotions.

In this paper, for recognizing and optimizing student learning emotions, we opt to use the *CascadeObjectDetector* function in *MATLAB*, which encapsulates a face detection model based on *Haar-like* features and *Adaboost* cascade algorithm. By invoking this function, face detection can be directly performed on the *JAFFE* and *CK+* datasets, which are specifically designed to include facial expressions. The advantage of this method lies in eliminating the complexity and time consumption of training a detection model from scratch, while leveraging *MATLAB*'s powerful computing capabilities and mature algorithm library. In the preprocessing stage, images from these two datasets are inputted into the *CascadeObjectDetector*, which automatically locates the facial regions in the images, serving as input for subsequent deep learning model analysis of expressions.

2.2 Normalization processing

(1) Angle Normalization

Angle normalization is aimed at eliminating the impact of facial angle changes caused by the tilt of a student's head on emotion recognition. In real environments, students may have various head postures, such as tilting, looking down, or looking up, which can cause the location of facial features in the images to change, affecting the accuracy of feature extraction. Angle normalization means adjusting the detected face to a standard frontal position through rotation and transformation, ensuring that the features extracted correspond to a unified reference frame regardless of how the head is tilted in the original image.

Specifically, the process begins with the detection of eye positions using the *Viola-Jones* algorithm, which is the starting point for pupil localization, utilizing *Haar-like* features and a cascade classifier to quickly locate the eye area in the image. Further, the *Canny* edge detection algorithm is applied to the detected eye area to highlight the edge information of the eye area. Edge detection generates a binary image that prominently displays the contour of the pupil, preparing suitable input data for the next step, the *Hough* transform. The *Hough* transform provides the center position and radius of the pupil, which is crucial for subsequent angle normalization operations. By accurately locating the center of the pupil, it is ensured that the facial image can be rotated to face forward based on the position of the eyes, regardless of the original image's angle, providing a standardized and aligned input image for emotion recognition. Assuming the coordinates of the left and right eye pupils' centers are represented by (a_m,b_m) and (a_e,b_e) respectively, and the angle required for rotation is denoted by ϕ, the calculation formula is as follows:

$$\varphi = ARCTAN\frac{b_e - b_m}{a_e - a_m} \tag{2}$$

Assuming the coordinates of any point in the original image are (a,b), the coordinates after rotation can be calculated using the following formula:

$$\left(a',b'\right)^S = \begin{pmatrix} COS\varphi & SIN\varphi \\ -SIN\varphi & COS\varphi \end{pmatrix}\left(a,b\right)^S \tag{3}$$

(2) Scale Normalization

The purpose of scale normalization is to ensure that the detected facial images have a uniform size and proportion, regardless of their actual size in the original images. Since images from different sources may vary in size due to shooting distance or camera parameters, using unnormalized images directly for feature extraction would make it difficult for the learning model to correctly understand and compare the same features across different images. By adjusting each detected facial image to a predetermined size, the features learned by the network are unaffected by the original image size, further improving the consistency of feature extraction and the accuracy of subsequent emotion classification.

Specifically, the distance between the left and right pupils, obtained using pupil localization technology, is calculated first. This distance is a key parameter in the subsequent cropping and scale normalization process, providing a standardized measure to determine the size and position of the facial area based on this distance. The purpose of calculating the distance between pupils is to obtain a reference scale, ensuring that faces in different images maintain a consistent size ratio after scaling. Further, based on the calculated distance between the pupils, the boundaries for cropping the facial area are determined. Vertically, based on the line connecting the left and right pupils, 0.5 times the distance is taken as the upper boundary, and 1.5 times the distance as the lower boundary for cropping; horizontally, centered on the perpendicular bisector of the line connecting the left and right pupils, the distance between the pupils is taken as the left and right boundaries for cropping. This cropping method, based on the general proportional rules of facial structure, can well include key feature areas of the face, such as the eyes, nose, and mouth, while excluding unnecessary background information. Finally, the cropped facial images are scaled to a uniform size, selected to be 96×96 pixels in this study. Scaling is done through interpolation algorithms to maintain image quality and ensure facial features are not distorted due to size adjustments.

(3) Grayscale Normalization

Grayscale normalization is aimed at reducing the interference caused by changes in lighting conditions and color information. Color information is usually not essential for emotion recognition, and different lighting conditions and color components may mislead emotion recognition. By converting images to grayscale, it eliminates the interference of colors while also reducing the amount of data the algorithm needs to process.

First, calculate the probability density function of the grayscale levels of the original image, specifically by counting the frequency of each grayscale level and dividing by the total number of pixels in the image. Assuming the original grayscale levels are represented by $e_j \in [0,1]$, since e_j is a discrete random variable, the original grayscale levels can be characterized by the probability density function $O_e(e_j)$, with the calculation formula being:

$$O_e\left(e_j\right)=\frac{v_j}{V} \tag{4}$$

Next, calculate the new grayscale levels through the cumulative probability density function, which is the cumulative sum of the probabilities of each grayscale level. Assuming the new grayscale levels are represented by $t_j \in [0,1]$, the calculation formula is:

$$t_j=S\left(e_j\right)=\sum_{k=0}^{j}O_e\left(e_j\right)=\sum_{k=0}^{j}\frac{v_k}{V},\ j\in\left[0,255\right] \tag{5}$$

Finally, apply these new grayscale levels by modifying the original image's grayscale values through histogram equalization. In this step, each pixel value of the original image is updated according to the newly calculated grayscale levels, resulting in an image that is more uniformly distributed over the grayscale levels. In the context of emotion recognition, this means that the facial expression images of students will have better contrast, allowing the deep learning model to more accurately recognize and analyze emotional states. Assuming the grayscale levels after histogram equalization are represented by T, then the calculation formula is:

$$T_j=255\cdot t_j \tag{6}$$

3. STUDENT TEMPORAL EXPRESSION RECOGNITION BASED ON MULTI-ATTENTION FUSION NETWORK

In recognizing student learning states, facial expression changes exhibit high temporality and subtle spatial features, requiring models to capture fine temporal continuity and the importance of different facial regions. To address the issue that single images or traditional methods struggle to capture continuous emotional changes and complex emotional expressions, this research conducted a study on student temporal expression recognition based on a multi-attention fusion network. The model introduces shallow attention networks and deep attention networks, incorporating three types of attention mechanisms: self-attention, weight attention, and convolutional attention. The multi-attention mechanism can focus on multiple key areas of facial expressions simultaneously, and the fusion network helps integrate these areas' dynamic changes over time series, thus providing richer and more precise emotional recognition information for deep learning models.

3.1 Shallow attention network

The role of the shallow attention model in the student dynamic sequence facial expression recognition system is to capture the subtle differences of facial action units using human prior knowledge of facial expressions. Facial action units, a coding system describing facial muscle movements, are crucial for expressing specific emotions. By analyzing the relative positions of facial landmarks and the texture features of local areas, the shallow attention model can extract the basic components of facial expressions, such as minor muscle movements and foundational changes in expressions. These shallow features help the system locate and recognize subtle changes in facial expressions, which is particularly important for understanding and predicting changes in students' emotional states during the learning process. For instance, a minor eyebrow raise of a student might indicate confusion or questioning, and the shallow attention model is dedicated to capturing such key non-verbal cues. Figure 3 provides a structural diagram of the shallow attention network.

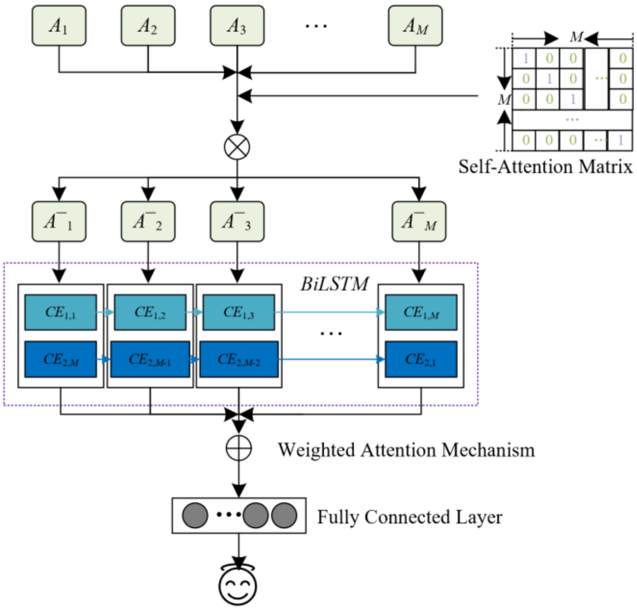

Figure 3. Shallow attention network structure diagram

Then, a mathematical model is used to describe the shallow features of student faces.

First, identify landmark features of the student's face, such as the corners of the eyes, eyebrows, tip of the nose, corners of the mouth, etc. Assuming 70 facial landmark key points detected by $O=[o_1,o_2,\ldots,o_i,\ldots o_n,\ldots,o_{70}]$, with o_i's coordinates being (a_{oi},b_{oi}). The number of line segments is 38, represented as $E=[e_1,e_2,\ldots,e_k,\ldots,e_{38}]$, divided into 7 regions, namely X_1, X_2, X_3, X_4, X_5, X_6, X_7.

The distance between detected key points can provide useful information about facial expressions. For example, the distance between the corners of the mouth increases when smiling. These distances can form a vector representing the face's geometric features. Assuming e_k's two endpoints are o_i and o_n, the calculation formula is:

$$F_{e_k} = \sqrt{\left(a_{o_i} - a_{o_n}\right)^2 + \left(b_{o_i} - b_{o_n}\right)^2} \tag{7}$$

To ensure the features are unaffected by the size and position of the head, normalization is necessary, typically involving the introduction of scale and translation invariance. The normalized distance between key points is as follows:

$$\overline{F_{ek}} = \frac{F_{e_k}}{\sqrt{\left(a_{o28} - a_{o31}\right)^2 + \left(b_{o28} - b_{o31}\right)^2}} \tag{8}$$

Local Binary Patterns (*LBP*) are a powerful texture descriptor capable of capturing the subtle texture information within local facial regions. By calculating the *LBP* value around each key point, texture features of facial expressions can be obtained. *LBP* features are somewhat robust to changes in lighting, making them suitable for describing local texture changes in facial expressions. Assuming the grayscale value of a pixel is represented by *GR*(.), the number of sampling points around a pixel is represented by *l*, and the *LBP* feature value of each pixel can be calculated through the following formula:

$$LBP(z) = \sum_{u=1}^{l} 2^u T\left(GR(u) - GR(c)\right) \tag{9}$$

$$T(s) = \begin{cases} 1, s \geq 0 \\ 0, \text{others} \end{cases} \tag{10}$$

For ease of analysis and classification, the *LBP* features of each local area are converted into statistical histograms. Histograms account for the frequency of different *LBP* patterns, summarizing and representing the texture information of facial regions. This representation method is concise and information-rich, facilitating subsequent machine learning processing. Let the set of pixels contained in region X_j in the image be represented by [oua_1,oua_2,...,oua_q,..], where the number of pixels in X_j is denoted by *v*. Then, the *LBP* feature of X_j is the statistical histogram of *LBP*(oua_q), represented by B_{XJ}, as follows:

$$B_{X_j} = COUNT\left(N == LBP\left(oua_q\right)\right) \tag{11}$$

Finally, the normalized key point distance vector and all regional *LBP* histogram vectors are concatenated to form a comprehensive feature vector *A*, which contains geometric and texture information about the student's facial expressions.

$$A = \left[\overline{F_{e_k}}; B_{X_j}\right] \tag{12}$$

Based on the feature vector *A* extracted through the steps above, a dynamic sequence of images with a length of *M* can be constructed, namely [A_1,A_2,…,A_s,…,A_M]. The facial expressions of students dynamically change over time, which is crucial for recognizing learning emotions. Therefore, this paper introduces a self-attention matrix in the dynamic sequence facial expression recognition system, aiming to capture and enhance the correlation between different frames in sequential data. Through the self-attention matrix, the system can analyze a series of temporally continuous shallow feature vectors to identify which specific facial features are more prominent in expressing emotions. Thus, the model focuses not only on a single static image but achieves more accurate emotion recognition by considering changes in facial features over time. The calculation of the self-attention mechanism is as follows:

$$\overline{A_s} = \sum_{u=1}^{l} SATT\ Lxs_{su} * a_s \tag{13}$$

LSTM, particularly its bidirectional variant *BiLSTM*, is widely used for its capability to capture long-term dependencies in sequential data. *BiLSTM* can learn information from both past and future contexts, but not all information in the sequence is equally important for the current task. The addition of a weight attention mechanism can automatically identify and assign higher weights to those features that are more critical for determining the current emotional state. Specifically, the dynamic sequence images are first encoded through a *BiLSTM* layer to obtain high-level features containing contextual information. Then, the weight attention mechanism calculates the weights of feature vectors at each time step through a well-trained attention network. These weights are used to weight the output of the *BiLSTM*, generating a comprehensive, weighted feature representation that focuses more on time steps that are more helpful for emotion state classification. Finally, this weighted feature representation is fed into a subsequent classifier to complete the emotion recognition task.

Specifically, this paper processes the feature matrix treated with self-attention using bidirectional *BiLSTM*. Assuming the forward *LSTM* unit is represented by $CE_{2,M\text{-}s+1}$, the backward *LSTM* processing is represented by $CE_{1,s}$, the concatenation operation is represented by [CE_1,CE_2], the length of the image sequence is denoted by *L*, and the weight of the features of image s is denoted by Q_s. The weight attention applied to the result of *BiLSTM* is as follows:

$$QX\left(CE_1, CE_2, Q\right) = \frac{1}{M} \sum_{u=1}^{M} Q_u \bullet \left[CE_{1,s}, CE_{2,M-s-1}\right] \tag{14}$$

3.2 Deep attention network

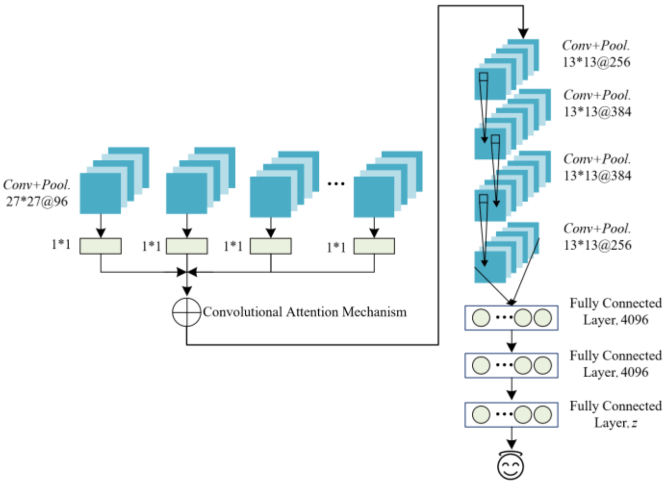

Figure 4. Deep attention network structure diagram

The deep attention model, inspired by the *ALexNet* architecture, is designed to capture high-level semantic features typically associated with the overall and dynamic characteristics of facial expressions. Utilizing the powerful capability of *CNNs* to extract deep features, the deep attention model can identify high-level emotional expressions, such as happiness, sadness, or anger, from sequence images. This model is particularly effective in understanding students' emotional fluctuations over a longer duration, revealing emotional patterns not discernible from a single frame, thus capturing the dynamic changes in students' emotions. Figure 4 provides a structural diagram of the deep attention network.

In the constructed student dynamic sequence facial expression recognition system, this paper chooses to modify the *AlexNet* structure to handle three-dimensional spatiotemporal data and improve the accuracy of emotion recognition. First, the convolution and pooling layers of *AlexNet* are extended to a sequence mode, meaning the same set of parameters is applied to each image in the sequence, capturing the dynamic features of the time series while maintaining high sensitivity to spatial features. Secondly, the introduction of convolutional attention mechanisms and 1×1 convolution kernels effectively integrates features from different time points and eliminates the time dimension. Assuming the convolution operation is represented by ⓧ, the 1×1 convolution kernel by ZX_s, and the output from the previous layer by ZO_s, the channel attention mechanism can be calculated as follows:

$$ZX\left(ZO_s ZX_s\right)=\frac{1}{M}\sum\nolimits_{s=1}^{M} ZX_s \otimes ZO_s \tag{15}$$

3.3 Network fusion

The calculation method for the fusion result D of the output of the deep attention network and the deep attention network is as follows:

$$D\left(P_1,P_2,Q_1,Q_2\right)=\frac{1}{2}\left(Q_1^S P_1+Q_2^S P_2\right) \tag{16}$$

After obtaining the basic facial expression recognition results, this paper chooses to combine the analysis of facial expressions with classroom contextual information, including the nature of classroom activities, students' learning tasks, interactions among individuals, and classroom atmosphere. For example, a student might show a frustrated expression during a challenging learning task, but if this occurs in an environment that encourages innovative thinking, the expression might reflect deep contemplation rather than negative emotions. By transcending a single data source and integrating multimodal data analysis, including facial expressions and the specific teaching environment and context of the students, a more accurate capture and interpretation of students' emotional states is achieved. This provides educators with insights to adjust teaching strategies and enhance student engagement.

4. EXPERIMENTAL RESULTS AND ANALYSIS

The correspondence between facial expressions and learning emotions can be varied. This paper adopts the following seven correspondences for the experiment: 1) Happy: Student facial expression is smiling, eyes curving, corresponding to emotions of interest or satisfaction, indicating students enjoy the learning process. 2) Sad: Student facial expression is drooping, eyes revealing disappointment, indicating students experience frustration or disappointment. 3) Surprise: Student eyes are wide, mouth open, representing curiosity or shock, sometimes due to suddenly understanding a difficult problem. 4) Fear: Student facial expression is tense, eyebrows raised, indicating anxiety or nervousness, especially when facing exams or difficult problems. 5) Disgust: Student nasolabial folds are wrinkled, expression appears displeased, corresponding to dissatisfaction or boredom, a reaction to learning materials or certain teaching methods. 6) Anger: Student's eyebrows are tightly locked, mouth closed or chin protruding, indicating frustration or hostility, a direct reaction to learning challenges. 7) Confusion: Student's eyebrows are tightly furrowed, corners of the mouth downturned, indicating confusion or thinking, occurring when students try to understand complex concepts.

Table 1. Distribution of sample set in the experiment

Database	Happy	Sad	Surprise	Fear	Disgust	Anger	Confusion	Total
Standardized Expression Database *CK*+	68	27	84	36	58	46	19	338
Natural Classroom Recordings	41	31	41	27	32	31	25	345
Simulated Dataset	81	81	81	81	81	81	81	352

Table 2. Recognition accuracy of different methods for recognizing student facial expressions

Method	Standardized Expression Database *CK*+ (%)	Natural Classroom Recordings (%)	Simulated Dataset (%)
BiLSTM-At	92.31	61.23	71.24
Mul-LSTM-Att	82.25	63.45	55.58
HAN	87.25	58.26	67.26
Conv-Transforme	92.36	77.45	-
Self-Att-BERT	93.25	74.12	75.54
GAT	96.39	70.15	81.34
Trans-XL	93.48	81.23	81.12
XLNet-Att	93.87	84.52	92.38
Cross-Att-BERT	88.12	68.45	71.37
CBAM	93.98	82.36	78.22
The Proposed Method	98.22	89.87	87.32

Table 3. Confusion matrix for recognition results in the standardized expression database *CK*+

Accuracy (%)	Happy	Sad	Surprise	Fear	Disgust	Anger	Confusion
Happy	92.36	6.24	0	0	0	0	0
Sad	0	100	0	0	0	0	0
Surprise	0	0	100	0	0	0	0
Fear	0	0	0	100	0	0	0
Disgust	0	0	0	0	100	0	0
Anger	0	0	0	0	0	100	0
Confusion	0	0	0	0	0	0	100

Table 4. Confusion matrix for recognition results in natural classroom recordings

Accuracy (%)	Happy	Sad	Surprise	Fear	Disgust	Anger	Confusion
Happy	82.36	15.98	0	0	0	0	0
Sad	0	100	0	0	0	0	0
Surprise	0	0	100	0	0	0	0
Fear	0	16.9	0	81.6	0	0	0
Disgust	0	0	0	0	100	0	0
Anger	0	0	25.8	0	0	73.2	0
Confusion	0	0	0	0	0	0	100

Table 5. Confusion matrix for recognition results in the simulated dataset

Accuracy (%)	Happy	Sad	Surprise	Fear	Disgust	Anger	Confusion
Happy	88.2	12.4	0	0	0	0	0
Sad	11.3	88.3	0	0	0	0	0
Surprise	0	0	61.2	12.9	23.6	0	0
Fear	0	0	0	100	0	0	0
Disgust	0	0	0	0	85.3	0	0
Anger	0	0	0	0	0	100	0
Confusion	0	0	0	0	0	0	100

Table 6. Comparison of recognition rates with different attention mechanisms

Method	Standardized Expression Database *CK*+ (%)	Natural Classroom Recordings (%)	Simulated Dataset (%)
+ Self-Attention	82.36	64.12	62.25
+ Weighted Attention	85.25	64.76	63.79
+ Convolutional Attention	88.26	67.35	68.36
Self-Attention + Weighted Attention	89.33	69.46	70.78
Weighted Attention + Convolutional Attention	92 89	80.56	77.24
The Proposed Method	94.46	81.34	79.23

Table 7. Cross-dataset experiment results

Train on Standardized Expression Database *CK*+ (%)		Train on Natural Classroom Recordings (%)		Train on Simulated Dataset (%)	
Natural Classroom Recordings	62.36	Standardized Expression Database *CK*+	65.32	Standardized Expression Database *CK*+	54.26
Simulated Dataset	44.26	Simulated Dataset	32.31	Natural Classroom Recordings	37.26

The experimental sample sets specifically include: 1) Natural Classroom Recordings: Collect student facial expression videos in natural classroom environments, providing authentic learning situations. 2) Standardized Expression Databases: Such as *CK*+ (*Extended Cohn-Kanade*) and *FER*-2013, these databases contain standardized facial expression images for baseline performance assessment. 3) Simulated Dataset: Expression data generated using computer vision technology, which can be used to test the algorithm's performance under extreme or rare emotional expressions. Table 1 presents the distribution of the experimental sample set.

Based on the recognition accuracy rates of different student facial expression recognition methods listed in Table 2, across three datasets (Standardized Expression Database *CK*+, Natural Classroom Recordings, and Simulated Dataset), the following analysis and conclusions can be drawn. For the Standardized Expression Database *CK*+, the proposed method performed the best, with a recognition accuracy of 98.22%. This demonstrates our method's significant advantage under standardized conditions compared to other existing methods. Compared to the second-highest accuracy rate (93.98% of *CBAM*), the proposed method shows an improvement of over 4 percentage points. In the more realistic setting of natural classroom recordings, the proposed method also exhibited superior performance, leading with a recognition accuracy of 89.87%. This result indicates that even in complex and variable real-world environments, the proposed method remains robust and applicable. *XLNet-Att* is in second place with an 84.52% accuracy, and the proposed method has an improvement of over 5 percentage points relative to *XLNet-Att*. On the simulated dataset, the proposed method again leads with an accuracy rate of 87.32%, with *XLNet-Att* closely following at a high accuracy rate of 92.38%. The gap between the proposed method and *XLNet-Att* is small, but the proposed method still maintains the lead.

From the data presented, it is evident that the proposed method significantly outperforms other comparative methods across all datasets. This validates the effectiveness of the optimizations in preprocessing and feature extraction in our study, as well as the efficacy of the multi-attention fusion network in capturing the dynamic changes of facial expressions. Moreover, the ability of the proposed method to maintain high recognition accuracy in both real and simulated environments further indicates its robustness and wide applicability.

After analyzing the confusion matrices in Tables 3, 4, and 5, the following conclusions can be drawn. From Table 3, it is known that except for the "Happy" expression, the recognition accuracy rates for all other expression categories reached 100%. The recognition accuracy for the "Happy" expression was slightly lower than others, at 92.36%, with 6.24% of "Happy" expressions being misidentified as "Sad". The proposed method performed almost perfectly when processing the standardized dataset, demonstrating extremely high recognition capabilities. According to Table 4, most expressions such as "Sad", "Surprise", "Disgust", and "Confusion" were accurately recognized, with recognition rates reaching 100%. The recognition accuracy for "Fear" and "Anger" expressions decreased to 81.6% and 73.2%, respectively. Notably, 25.8% of "Anger" expressions were misidentified as "Surprise". The recognition accuracy for "Happy" expressions also slightly decreased to 82.36%, but still remained higher than most other categories. Even in the complex real-world environment of natural classroom settings, the proposed method demonstrated good recognition capabilities. According to Table 5, in the simulated dataset, recognition rates for "Fear", "Disgust", "Anger", and "Confusion" expressions still maintained at 100%. The recognition accuracies for "Happy" and "Sad" expressions were relatively high, at 88.2% and 88.3%, respectively. The recognition accuracy for "Surprise" expressions significantly dropped to 61.2%, indicating that recognizing this particular expression in a simulated environment is a challenge, with 12.9% of "Surprise" expressions being misidentified as "Fear", and 23.6% as "Disgust". It can be concluded that the proposed method performed almost perfectly in the Standardized Expression Database *CK*+ and, although performance slightly decreased in natural classroom recordings and simulated datasets, it still showed strong recognition accuracy and robustness. In both real and simulated environments, the proposed method demonstrated high accuracy and reliability in recognizing most expression categories, especially in natural settings, proving the practicality of the method.

From Table 6, it is observable that different attention mechanisms vary in their recognition rates across three different datasets. Analyzing these data, the following conclusions can be drawn. The method introduced in this paper, which incorporates three different attention mechanisms, outperforms other methods across all datasets, especially reaching a recognition rate of 94.46% on the Standardized Expression Database *CK*+, 81.34% on Natural Classroom Recordings, and 79.23% on the Simulated Dataset. This indicates that the proposed method can effectively combine multiple attention mechanisms to optimize performance in expression recognition. It can be concluded that the three different attention mechanisms have optimized the model's capability to integrate features in processing sequential data, providing better focus on important information, which is particularly crucial in recognizing sequences of facial expressions. This information again validates that the proposed method demonstrates superior recognition performance across different datasets, particularly showing clear advantages in complex and varied environments like natural classroom recordings and simulated datasets.

Table 7 demonstrates the performance when the model is trained on one dataset and tested on another. From the table, the following conclusions can be drawn. When the model is trained on the Standardized Expression Database *CK*+, the performance on Natural Classroom Recordings is 62.36%, and on the Simulated Dataset, it is 44.26%. This indicates that the Standardized Expression Database *CK*+ is a useful data source for training models with better generalization capability, especially compared to the Simulated Dataset. The model trained on Natural Classroom Recordings tested on the Standardized Expression Database *CK*+ achieved a recognition rate of 65.32%, but only 32.31% on the Simulated Dataset. This suggests that data from Natural Classroom Recordings is more diverse and complex, thereby training more robust models on standardized data, but this diversity might not apply to all types of simulated data. When the model is trained on the Simulated Dataset, its performance on the Standardized Expression Database *CK*+ is 54.26%, and on Natural Classroom Recordings, it is 37.26%. This indicates that simulated data is the least ideal choice for training models with strong generalization capabilities, as it shows lower recognition rates in two different testing environments.

5. CONCLUSION

This paper first detects students' faces using the method based on *Haar-like* features and the *Adaboost* cascade, followed by improving the quality of facial expression image preprocessing through normalization (including scale, angle, and grayscale normalization). This lays a solid foundation for subsequent expression recognition. The study proposes a temporal expression recognition model based on a multi-attention fusion network. This model combines the prior knowledge of the FACS and shallow and deep features extracted from the dynamic sequence of facial expressions to achieve more accurate recognition of students' learning emotions. Three different attention mechanisms are introduced, optimizing the model's capability to process features in sequential data, thereby enhancing the accuracy and efficiency of emotion recognition.

In the experimental section, not only are the recognition accuracies of different methods analyzed, but detailed recognition results in different datasets (Standardized Expression Database *CK*+, Natural Classroom Recordings, Simulated Dataset) are also presented through confusion matrices, assessing the impact of different attention mechanisms. It can be concluded that the method presented in this paper performs excellently in recognizing students' learning emotions in complex natural classroom environments. Face detection and normalization preprocessing ensure consistency and quality of input data. The proposed multi-attention fusion network model, combining various features and prior knowledge, has demonstrated effectiveness in capturing facial expressions through dynamic sequences. The introduced attention mechanisms significantly improved the model's capability to process features in sequential data, thus enhancing the accuracy and efficiency of emotion recognition, which was validated in cross-dataset experiments. The experimental results show that the proposed method has broad applicability in real-world scenarios and can handle different types of datasets.

Future research could explore further algorithm optimization directions, such as using more advanced deep learning models to improve recognition effects. Attempts could be made to train and test on more diverse datasets to verify the model's generalization ability and extend it to different learning environments and cultural backgrounds.

ACKNOWLEDGMENT

This paper was funded by Youth Fund of the Ministry of Education: Research on the Construction of Ideological Positions in Colleges and Universities in the Omni-Media Era (Grant No.: 20YJC710065).

REFERENCES

[1] Bertsimas, D., Villalobos Carballo, K., Boussioux, L., Li, M.L., Paskov, A., Paskov, I. (2023). Holistic deep learning. Machine Learning, 113(1): 159-183. https://doi.org/10.1007/s10994-023-06482-y

[2] Zhou, X. (2024). Deep learning algorithms in enterprise accounting management analysis. Applied Mathematics and Nonlinear Sciences, 9(1): 1-14. https://doi.org/10.2478/amns.2023.2.00367

[3] Afaq, Y., Manocha, A. (2024). Blockchain and deep learning integration for various application: A review. Journal of Computer Information Systems, 64(1): 92-105. https://doi.org/10.1080/08874417.2023.2173330

[4] Purushotham, P., Kiran, A., Reddy, Y. (2023). Sentiment analysis using deep learning for students' feedback: A survey. In 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, pp. 1-4. https://doi.org/10.1109/ICCCI56745.2023.10128533

[5] Purushottama, R.K., Janet, B. (2023). Detecting academic affective states of learners in online learning environments using deep transfer learning. Scalable Computing, 24(4): 957-970. https://doi.org/10.12694/scpe.v24i4.2470

[6] Gupta, S., Kumar, P., Tekchandani, R. (2023). EDFA: Ensemble deep CNN for assessing student's cognitive state in adaptive online learning environments. International Journal of Cognitive Computing in Engineering, 4: 373-387. https://doi.org/10.1016/j.ijcce.2023.11.001

[7] López, B., Arcas-Túnez, F., Cantabella, M., Terroso-Sáenz, F., Curado, M., Muñoz, A. (2022). EMO-Learning: Towards an intelligent tutoring system to assess online students' emotions. In 2022 18th International Conference on Intelligent Environments (IE), Biarritz, France, pp. 1-4. https://doi.org/10.1109/IE54923.2022.9826770

[8] Xu, Y., Li, Y., Chen, Y., Bao, H., Zheng, Y. (2023). Spontaneous visual database for detecting learning-centered emotions during online learning. Image and Vision Computing, 136: 104739. https://doi.org/10.1016/j.imavis.2023.104739

[9] Gazawy, Q., Buyrukoglu, S., Akbas, A. (2023). Deep learning for enhanced education quality: Assessing student engagement and emotional states. In 2023 Innovations in Intelligent Systems and Applications Conference (ASYU), Sivas, Turkiye, pp. 1-8. https://doi.org/10.1109/ASYU58738.2023.10296748

[10] Ta, C.D.C., Thai, N.T. (2024). Face emotion using deep learning. In Proceedings of the 1st International Conference on Intelligent Systems and Data Science (ISDS 2023), 1950: 174-184.

[11] Yang, W. (2024). Extraction and analysis of factors influencing college students' mental health based on deep learning model. Applied Mathematics and Nonlinear Sciences, 9(1): 1-14. https://doi.org/10.2478/amns.2023.2.00773

[12] Li, X., Yue, R., Jia, W., Wang, H., Zheng, Y. (2021). Recognizing students' emotions based on facial expression analysis. In 2021 11th International Conference on Information Technology in Medicine and Education (ITME), Wuyishan, Fujian, China, pp. 96-100. https://doi.org/10.1109/ITME53901.2021.00030

[13] Harb, A.A., Gad, A., Yaghi, M., Alhalabi, M., Zia, H., Yousaf, J., Ghazal, M. (2023). Diverse distant-students deep emotion recognition and visualization. Computers and Electrical Engineering, 111: 108963. https://doi.org/10.1016/j.compeleceng.2023.108963

[14] Zhao, L. (2023). Use of a deep learning approach for the evaluation of students' online learning cognitive ability. International Journal of Emerging Technologies in Learning, 18(12): 58-74. https://doi.org/10.3991/ijet.v18i12.41093

[15] Khemakhem, F., Ellouzi, H., Ltifi, H. (2022). A novel deep multi-task learning to sensing student engagement in e-learning environments. In 2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA), Abu Dhabi, United Arab Emirates, pp. 1-7. https://doi.org/10.1109/AICCSA56895.2022.10017756

[16] Trabelsi, Z., Alnajjar, F., Parambil, M.M.A., Gochoo, M., Ali, L. (2023). Real-time attention monitoring system for classroom: A deep learning approach for student's behavior recognition. Big Data and Cognitive Computing, 7(1): 48. https://doi.org/10.3390/bdcc7010048

[17] Wang, C. (2022). Emotion recognition of college students' online learning engagement based on deep learning. International Journal of Emerging Technologies in Learning, 17(6): 110-112. https://doi.org/10.3991/ijet.v17i06.30019

[18] Chen, Q., Lee, B.G. (2023). Deep learning models for stress analysis in university students: A sudoku-based study. Sensors, 23(13): 6099. https://doi.org/10.3390/s23136099