Research Article

Wei Zhao and Liguo Qiu*

Emotion recognition and interaction of smart education environment screen based on deep learning networks

https://doi.org/10.1515/jisys-2024-0082
received January 27, 2024; accepted May 27, 2024

Abstract: Smart education environments combine technologies such as big data, cloud computing, and artificial intelligence to optimize and personalize the teaching and learning process, thereby improving the efficiency and quality of education. This article proposes a dual-stream-coded image sentiment analysis method based on both facial expressions and background actions to monitor and analyze learners' behaviors in real time. By integrating human facial expressions and scene backgrounds, the method can effectively address the occlusion problem in uncontrolled environments. To enhance the accuracy and efficiency of emotion recognition, a multi-task convolutional network is employed for face extraction, while 3D convolutional neural networks optimize the extraction process of facial features. Additionally, the adaptive learning screen adjustment system proposed in this article dynamically adjusts the presentation of learning content to optimize the learning environment and enhance learning efficiency by monitoring learners' expressions and reactions in real time. By analyzing the experimental results on the Emotic dataset, the emotion recognition model in this article shows high accuracy, especially in the recognition of specific emotion categories. This research significantly contributes to the field of smart education environments by providing an effective solution for real-time emotion recognition.

Keywords: deep neural network, MTCNN, 3D-CNN, intelligent education, emotion recognition

1 Introduction

With the rapid development of information technology, smart education has become an important part of contemporary education reform. Smart education environments represent an evolution from traditional educational technologies, distinguished by their integration of cutting-edge technologies such as big data, cloud computing, and artificial intelligence to create highly adaptive and personalized learning experiences. Unlike conventional educational methods that often employ a one-size-fits-all approach, smart education leverages real-time monitoring and analysis of learner behavior to dynamically adjust teaching content and strategies [1]. In a smart education environment, through real-time monitoring and analysis of learner behavior, the education system can dynamically adjust the teaching content and strategies to accommodate the learning needs and preferences of different students [2]. Furthermore, smart education environments are characterized by their ability to integrate various technological tools and platforms, providing a seamless and interactive learning experience that extends beyond the physical classroom [3]. By providing a seamless and interactive learning experience that extends beyond the physical classroom, smart education environments improve the efficiency and quality of education and foster a more engaging and effective learning environment for students.

* **Corresponding author: Liguo Qiu,** Department of Information Technology, Hunan College of Information, Changsha, 410200, China, e-mail: Liguo_Qiu@163.com
Wei Zhao: Department of Information Technology, Hunan College of Information, Changsha, 410200, China, e-mail: 3347785789@qq.com

Learning screen images mostly belong to computer-generated or synthetic images, and the steps of traditional image emotion recognition methods are as follows: first, extract the low-level visual features of the image such as color and its distribution, texture and lines [4], shape and its spatial layout [5], and then use the training samples to train an image emotion classifier, and then finally, use the trained classifiers to recognize the emotion and intensity of the image. Because human emotion perception of images comes from multiple factors, some of which are implicit and difficult to express and extract, the efficiency and accuracy of traditional image emotion recognition methods are low [6]. Traditional learner expression recognition algorithms mainly include the processes of image preprocessing [7], face detection [8], feature extraction [9], feature selection [10], and classifier construction [11], and the visual features of facial expressions need to be explicitly expressed and extracted. This undoubtedly increases the difficulty of recognition and may loss the key feature information of the original image. With the continuous deepening of research, the convolutional neural network (CNN) has also been proposed by scholars. Relative to the traditional feature extraction methods, CNNs omit the complex image preprocessing and feature extraction process in the early stage [12], so that they no longer rely on manually crafted explicit feature extraction methods, which improves the efficiency and accuracy, but also improves the robustness of the recognition algorithm. However, the CNN ResNet is weak in perceiving the positional relationship between the parts of an image target [13] and is poor at recognizing side-view expressions. To address the shortcomings of CNNs, in the literature, proposed Capsule Network (CapsNet), which can extract facial expression features in more depth and evaluate whether the positional relationships between facial expression features are consistent with the distribution of the features on the facial expression image was proposed. The E2-capsule neural network for FER, which combines the VGG16, the VGG16, and the VGG16 with the VGG16 was also proposed. This network increased the attention of action unit perception to extract key features of facial expressions such as eyes, mouth, and nose. CapsNet, which uses a shallow small convolutional kernel network layer to reduce the parameters of CapsNet was also proposed. Although CapsNet and the related methods mentioned earlier have shown their effectiveness in FER, their accuracy is hindered by the structure of CapsNet.

Despite advancements in smart education technologies, there remain significant gaps in our ability to optimize and personalize learning experiences [14]. Traditional emotion recognition methods in educational settings have struggled with low efficiency and accuracy due to their reliance on explicit feature extraction and the complex nature of human emotion perception [15]. Furthermore, existing neural network approaches often fail to fully capture the nuances of emotional expression [16], particularly in side-view expressions. To achieve adaptive interaction at the affective level of intelligent learning environments and to facilitate learners to learn easily, engagingly, and effectively, this study proposes an image sentiment analysis method based on dual-stream coding of facial and background actions. This method utilizes multi-task convolutional networks (MTCNN), 3D convolutional neural networks (3D-CNN), and improved algorithms to accurately extract and analyze both facial features and background action information. Additionally, the study involves designing an adaptive learning screen adjustment system that dynamically adjusts learning content according to the learner’s emotional responses and visual–emotional preferences, thereby providing a more personalized and engaging learning experience. By leveraging the proposed dual-stream coding approach, this system can better accommodate the varied emotional states of learners in real time, ensuring that the content delivery is always optimal for their current needs. Experimental validation conducted on the Emotic dataset demonstrates that the proposed method significantly enhances the accuracy and robustness of emotion recognition.

2 Neural networks based on dual-stream coding of facial and background actions

In this article, the overall construction of the proposed algorithmic model will be elaborated in detail, which is mainly divided into the following parts: face facial extraction and face coding module, background context-aware coding, attention mechanism, fusion network, and classification.

2.1 Facial and background action-based dual stream coding network framework

As shown in Figure 1, since the first part of Caps Net, Conv1, is shallowly convolved [11], the extracted features are not sufficient to effectively provide the required semantic information of expressions for the second part, the deep extraction layer, Primary Caps [17]. As a result, the capability of the second part is not fully utilized, which leads to a lower capability of the third part, dynamic routing layer dynamic routing [18], to decide the effective capsule vectors.

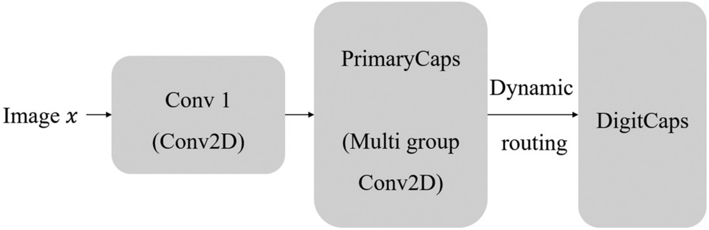

Figure 1: Caps Net network structure. Source: Created by the authors.

To solve this problem, this article proposes an image sentiment analysis method based on dual-stream coding of facial and background actions. The MTCNN model offers superior performance in detecting and aligning facial features due to its hierarchical structure that processes tasks at different scales [19]. This capability is crucial for accurately capturing the subtle facial expressions of learners in diverse educational contexts, where variations in lighting, pose, and background can significantly affect recognition accuracy. The overall architecture of the model, which is designed as a two-channel structure [20], including two feature coding streams: facial coding and background coding streams. The facial coding part first performs face cropping on the input frame-level images, after which the processed face images are fed into the facial coding module, which is based on the 3D-CNN structure [21], and two 2D convolutions are used instead of the original 3D convolution to improve the computational efficiency and obtain the face module features. The background coding stream first takes the cropped image as an input to another 3D-CNN, after which an attention coding module is used to find the attentional region of the contextual information in the background image, enabling the background coding stream to focus on the background context that is more helpful for emotion recognition, which allows the network to reduce ambiguity and improve the accuracy and robustness of emotion recognition. Finally, the two parts of the features are fused and fed into a classifier to predict the emotion category.

The advantage of the dual-stream coding method over traditional single-stream approaches lies in its comprehensive analysis capability. While single-stream methods typically focus solely on facial expressions, they often overlook the context, which can provide essential cues about the emotional state. For example, the same facial expression may convey different emotions depending on the context. By integrating both facial and background information, our dual-stream method provides a more nuanced and accurate emotion recognition, particularly in complex and dynamic learning environments where contextual cues play a significant role.

2.2 Human facial feature extraction

The input image data has to be segmented into face parts separately to obtain a face image before it can be analyzed by the face coding network; therefore, the face region in the image is extracted first; in this article, MTCNN is used to extract the face part of the face. This approach not only improves the efficiency and accuracy, but also reduces the performance loss of the traditional approach of using sliding windows and classifiers. MTCNN constructs three different network models, P-Net, R-Net, and O-Net, through the idea of recursive execution, which makes the face facial detection with higher efficiency and speed. First, the image pyramid produced transforms the image input to the input layer at various scales so that it can detect facial images at different scales. Then, the P-Net network produces a large number of region boxes as candidate target boxes, and then, the R-Net network carries out the initial fine selection and border reduction of these candidate region boxes and removes all the negative examples, and then, the O-Net network, which is a more complex network with higher accuracy, carries out the second discrimination and border reduction of all the candidate region boxes. The specific construction of the three networks is shown in Figure 2. In this article, we will utilize the three-layer network of MTCNN for face extraction and cropping and use the face photo as the input of the subsequent face coding system. This method not only improves the efficiency and accuracy of face extraction, but also provides a high-quality data base for subsequent sentiment analysis.

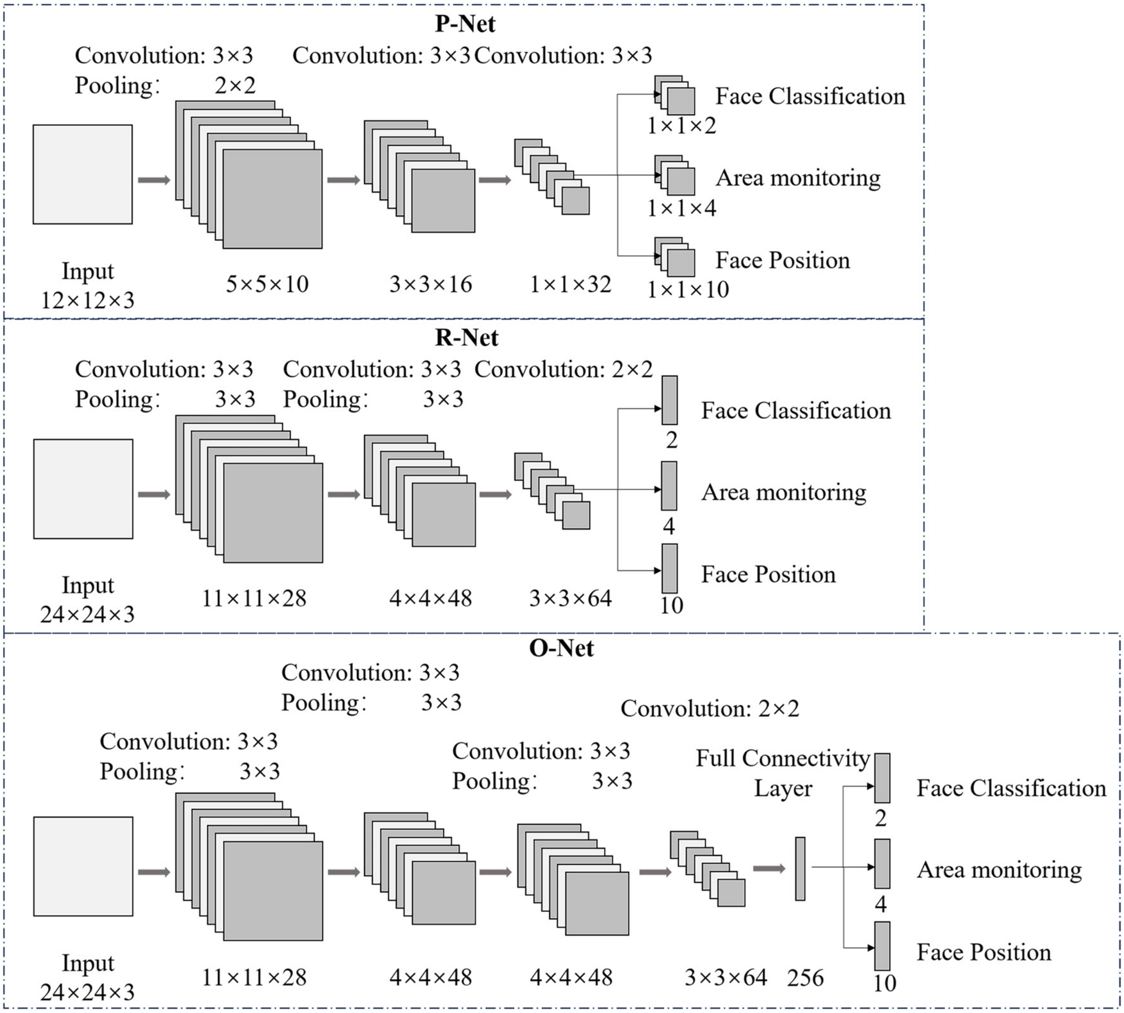

Figure 2: Network architecture of P-Net, R-Net, and O-Net. Source: Created by the authors.

P-Net is an important component in MTCNN, which is a face delineation boundary suggestion network based on CNN. The network accepts raw image features and processes them through three convolutional layers, then uses a face classifier to detect and delimit initially conforming face regions through border region reduction and a face keypoint localizer. This step generates a number of target candidates that may contain faces, and these candidates are fed into R-Net for subsequent detection.

R-Net is the second layer network in MTCNN face detection, which is used to further process the possible face regions output from P-Net and filter out the more credible face regions for subsequent O-Net. In the R-Net network, the input candidate face regions are selected more carefully, most of the wrong face regions are removed, and the border reduction and facial keypoint localizer are used again to restore the border and localize the keypoints of the remaining face candidate regions. Compared with the P-Net network, this layer uses 1 × 1 × 32 features generated by the full convolutional network and introduces a 128-dimensional fully connected layer after the final convolutional layer, which is not only more accurate than the P-Net, but also obtains more original image features.

O-Net is the most complex part of MTCNN face detection, which is richer in features, and a 256-dimensional fully connected layer is added at the end of the structure, which also preserves more image features on the basis of R-Net. O-Net obtains the upper-left and lower-right coordinate values and five feature points of the face region through the process of detecting the face, restoring the region's edge, and localizing the key points of the face: coordinate values and five feature points. Compared with the first layer of P-Net, O-Net not only has more feature inputs and more complex network structure, but also shows more superior performance. Therefore, the output of the O-Net network is also used as the output of the MTCNN detection algorithm model.

2.3 3D-CNN convolution

In a normal 2D CNN, the 2D convolution operation takes the local nearest neighbor region on the feature maps acquired in the previous layer to acquire the feature, subsequently adds an additional bias (bias), and finally inputs this result to a tanh function[n] (activation function) [22]. The value of the *j*th feature map of the *i*th layer at position (*x*, *y*) is shown in the following equation:

$$v_{ij}^{xy} = \tanh\left(b_{ij} + \sum_{m}^{P_i-1}\sum_{m}^{Q_i-1} w_{ijm}^{pq} v_{(i-1)^m}^{(x+p)(y+q)}\right), \tag{1}$$

where tanh (∗) is the activation function, b_{ij} is the bias of this feature map, m denotes the set index of the feature maps of the (*i*–1)th layer connected to the current feature map, *w* is the value of the convolution kernel connected to the *k*th feature map at the position of (*p*, *q*) position, and. are the height and width of the convolution kernel. In the downsampling layer, the resolution of the feature maps is reduced by performing pooling operations on the local nearest neighbors of the feature maps in the previous layer. A CNN is a network structure formed by stacking together such alternations of convolution and pooling. The parameters of a CNN, e.g., b_{ij} and kernel weights w_{ijm}^{pq}, are learned in a supervised or unsupervised manner.

In 2D CNNs, convolution is performed on a flat feature map, and 2D convolution is unable to obtain the corresponding motion coding information from multiple consecutive frames when higher-dimensional aspects of the problem need to be solved. In order to deal with this problem, this article adopts 3D convolution operation, which was motivated by its proficiency in analyzing temporal information, which is vital for understanding the dynamic nature of emotional expressions over time. The method is to construct a 3D convolution kernel and convolve it with multiple consecutive frames, so that the feature maps of the convolutional layers are connected into multiple consecutive frames on the previous layers, and thus, the motion information is captured. Finally, the value at position (*x*, *y*, *z*) of the *j*th feature map in the *i*th layer [23] is shown in the following equation:

$$v_{ij}^{xyz} = \tanh\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1} w_{ijm}^{pqr} v_{(i-1)m}^{(x+p)(y+q)(z+r)}\right), \tag{2}$$

where R_i is the scale of the 3D convolution kernel in the time dimension, and the value of (*p*, *q*, *r*) connected to the *m*th feature map convolution kernel of the previous layer is *w*. The difference about 2D convolution and 3D convolution is shown in Figure 3.

Specifically, the face coding network consists of five 3 × 3 × 3 convolutional layers, followed by a BN layer, a ReLU activation function, and five maximum pooling layers, where the first pooling layer has a pooling kernel size of 1 × 2 × 2 in order to avoid premature merging of temporal information. The remaining pooling

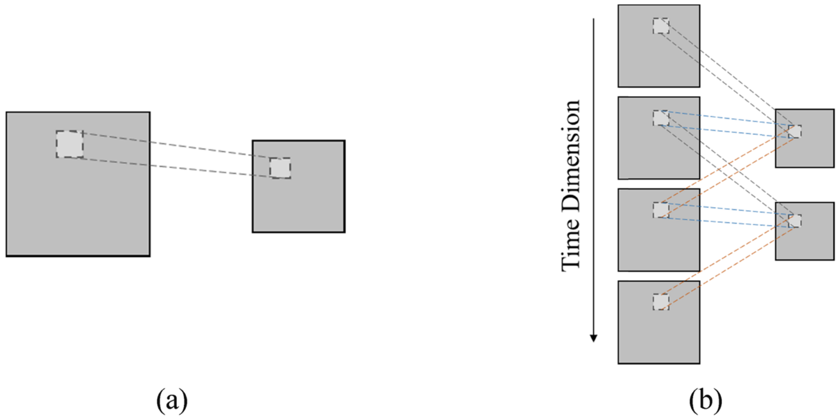

Figure 3: Comparison of 2D and 3D convolutions: (a) 2D convolution and (b) 3D convolution. Source: Created by the authors.

layers are maximum pooling layers of size 2 × 2 × 2, and the kernel counts of the five convolutional layers are 32, 64, 128, 256, and 256, respectively, and finally, the output features are spatially averaged by an average pooling layer to obtain the final output of the face coding network.

2.4 Background context coding

Compared with the face coding network, the background context coding layer includes a context coding module and an attention inference module. In order to extract contextual information other than facial expressions, specifically, the face part is first masked to obtain the remaining background image, which is fed into the coding network, which has the same design of convolutional and pooling layers as the face coding network, and also employs two convolutions to improve the computation rate to obtain the extracted background contextual features, which are then fed into an attentional inference module, through which the context cataloging can be made to focus on different contexts, extracting the part of the background context that is helpful for sentiment analysis.

The core idea can be expressed by the idea of mathematical level: $X = [x_1, x_2,..., x_N]$ is taken as N inputs, and at the same time, in order to reduce the cost of computing, it is not necessary to perform any processing on these N inputs by neural network, and only some task-related information from these N inputs need to be selected for computing. The soft attention mechanism is a method of selecting input information that, unlike hard attention, instead of selecting only 1 out of N information, is processed by weighted averaging the N inputs and then feeding them into the residual network for computation. The computation of the attention value usually consists of two aspects, i.e., computing the attention distribution of the input data and weighted averaging the input data by the attention distribution obtained from the computation.

Specifically, the background image that has been pooled by 3D convolution is taken as input X_c, and since 3D convolution has already feature extracted the input in the temporal dimension, the attention weights are denormalized only in the spatial dimension. In the attention inference module, 128 and 1 feature channels are generated by two convolutional layers with a convolutional kernel size of 3 × 3 × 3 with a normalization layer, and the attention $A \in R^{H \times W}$ is obtained from the input X_c, where $H \times W$ is the spatial resolution of X_c. Afterward, in order to make the sum of the attention of each pixel to be 1, the output is passed through a spatial Softmax normalization layer. The output is shown in the following equation:

$$\hat{A}_\imath = \frac{\exp(A_\imath)}{\sum_j \exp(A_\imath)}, \tag{3}$$

where $\hat{A}_\imath$ denotes the attention to the context at i, j. Afterward, the new background context feature with attention optimized is obtained by combining the weights of the attention part of the output on the input feature X_c, and finally, the output is obtained through a spatially averaged pooling layer [24] as well, which is formulated as follows:

$$\bar{X}_C = \hat{A} \odot X_C, \tag{4}$$

where $\odot$ denotes the element-by-element multiplication operator.

The algorithm model proposed in this article takes the cross-entropy of the predicted data and the actual results as the loss function of the model, and optimizes the model parameters by obtaining the minimum value of the loss function, and at the same time, adopts the method of random deactivation (Dropout) and the L_2 regularization operation to avoid the overfitting phenomenon during the training, and the formula of the cross-entropy loss function is expressed as

$$\text{Loss} = -\sum_i \sum_j \hat{y}_i^j \log y_i^j + \lambda \sum_{\theta \in \varphi} \theta^2, \tag{5}$$

where y_i^j denotes the prediction data generated by the ith sentence; $\hat{y}_i^j$ denotes the corresponding actual result; j denotes the specific number of classifications, which is set according to the sentiment category of the dataset for division; λ is the coefficient of the L_2 regularization term; and φ denotes all the parameters to be learned.

3 Learning screen adaptive adjustment system design

In the smart learning environment, the learning screen is a key medium for transferring information, which directly affects the emotional state and learning effect of learners. The technical backbone of this system comprises sophisticated algorithms capable of continuous emotional and engagement monitoring, employing real-time analysis to facilitate swift adjustments to the learning content. This ensures that the learning environment is consistently optimized for each learner's emotional state and learning pace. In this article, we use artificial intelligence, affective computing, and other technologies to analyze the learners' emotional response to the learning screen, and adjust the key visual features of the learning screen by combining the emotions of the learning screen and the learners' visual–emotional preferences. This process not only realizes the continuous improvement of the visual–emotional preference model of the learning screen, but also enhances the learners' emotional engagement and learning experience through adaptive interaction. By monitoring learners' expressions and reactions in real time, the system is able to dynamically adjust the presentation of learning content, thus optimizing the learning environment and improving learning efficiency. The adaptive adjustment model of the learning screen based on learners' expressions in the smart learning environment is shown in Figure 4.

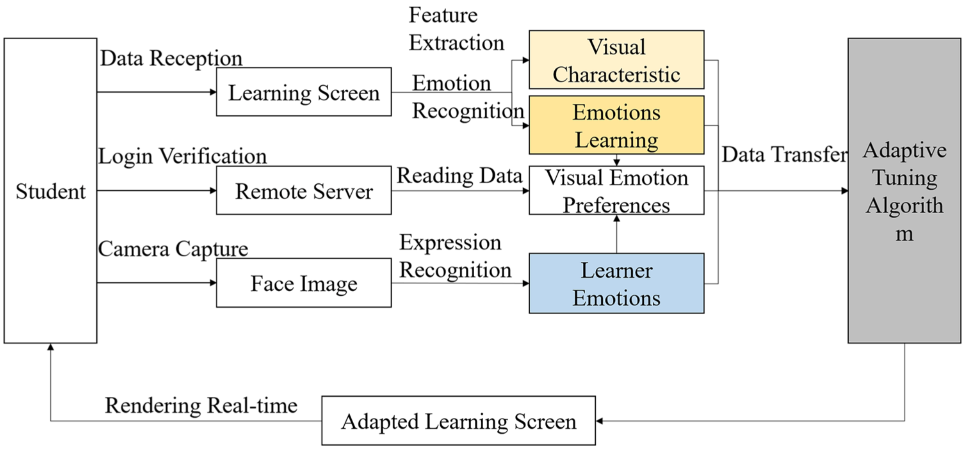

Figure 4: Framework of adaptive adjustment system for learning screens. Source: Created by the authors.

In this article, we propose a multi-module-integrated system designed to optimize the learning experience in smart learning environments, consisting of five key components, each focusing on a different function and analysis process. According to the system business architecture, it can be divided into five layers, as shown in Figure 5.

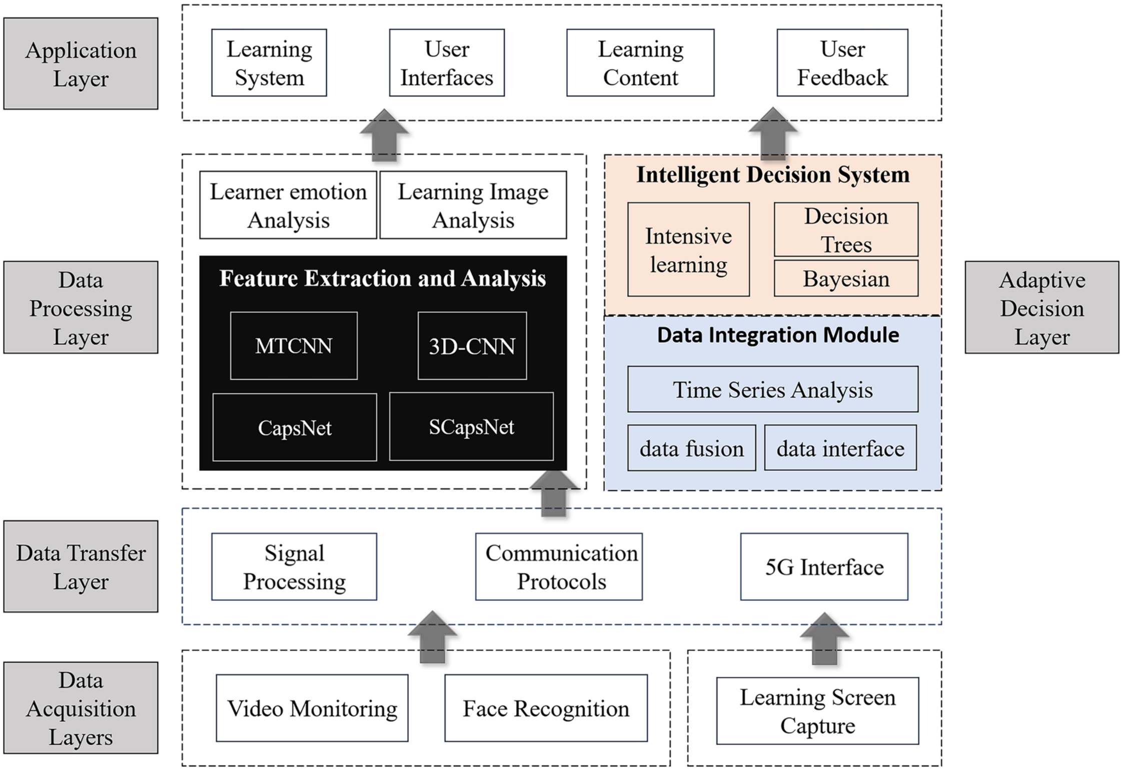

Figure 5: Learning screen adaptive adjustment system architecture diagram. Source: Created by the authors.

Among them, the data acquisition layer is mainly responsible for collecting raw data. The video monitoring module captures the learner's behavior and provides real-time video streaming for behavior analysis. The face recognition module implements the monitoring of learners' facial expressions through advanced image processing techniques, providing support for sentiment analysis and identity verification. The learning screen capture module records learners' screen interactions and captures key information during the learning process.

The data transmission layer undertakes the important task of ensuring the stability and security of data flow. It is responsible for efficiently and securely transferring all kinds of information collected by the data collection layer, optimized by encryption and compression technologies, to the data processing layer for further analysis.

The data processing layer is responsible for processing and analyzing the data incoming through the data transmission layer. It consists of several sub-modules, including MTCNN for high-precision face detection, 3D-CNN for extracting spatial–temporal features from video data, and two capsule networks, Caps Net and SCapsNet, for capturing hierarchical features and spatial relationships in images. These techniques work together to enhance the model's ability to understand dynamic and complex scenes.

The adaptive decision-making layer utilizes the analysis results provided by the data processing layer to implement efficient decision making through a series of algorithms and models such as reinforcement learning, decision trees, and Bayesian methods. The data integration module, on the other hand, is responsible for integrating multi-dimensional data through time-series analysis, data fusion, and data interface techniques to enhance the accuracy and reliability of decision making.

The topmost application layer provides users with direct interactive interfaces and learning resources. The learning system module manages educational content and learning progress, the user interface module provides an intuitive and friendly operating environment, the learning content module provides learners with rich and diverse teaching resources, and the user feedback module collects user feedback and provides data support for continuous improvement and personalization of the system. The user interface design incorporates adaptive feedback mechanisms, visual cues, and personalized learning dashboards, all aimed at enhancing user engagement and learning outcomes. Furthermore, the adoption of this system within smart education frameworks heralds a significant shift toward highly personalized learning experiences. By automating the adaptation of content presentation based on individual learner profiles, it not only enhances learning outcomes but also elevates the overall quality of education.

4 Results and discussion

4.1 Experimental data collection

In this article, experiments were conducted using the Emotic dataset [17], which combines annotations for 26 discrete emotion categories and three continuous emotion dimensions. The image sources for the Emotic dataset are divided into two main categories: one part originates from two public datasets – COCO and Ade20k; the other one part is obtained through Google search engine. Moreover, the inclusion of both discrete and continuous emotion dimensions offers a multifaceted view of emotional expressions, essential for developing a more accurate and sensitive emotion recognition model tailored to smart education. As shown in Figure 6, the selected images are characterized by two distinctive features: the wide diversity of backgrounds and the richness of different locations and environments. These features not only ensure the richness and diversity of the Emotic dataset, but also bring additional complexity and challenges to the emotion recognition task. The choice of the Emotic dataset was driven by its comprehensive coverage of diverse emotions and realistic settings, making it highly relevant for the study's objectives. The Emotic datasets were connected to a computer system equipped with an Intel Core i7 processor, NVIDIA GeForce RTX 2080 Ti GPU, 32GB RAM, running on Ubuntu 18.04 LTS. For data processing and simulation, we utilized PyTorch, a powerful open-source machine learning library, and Python 3.8.

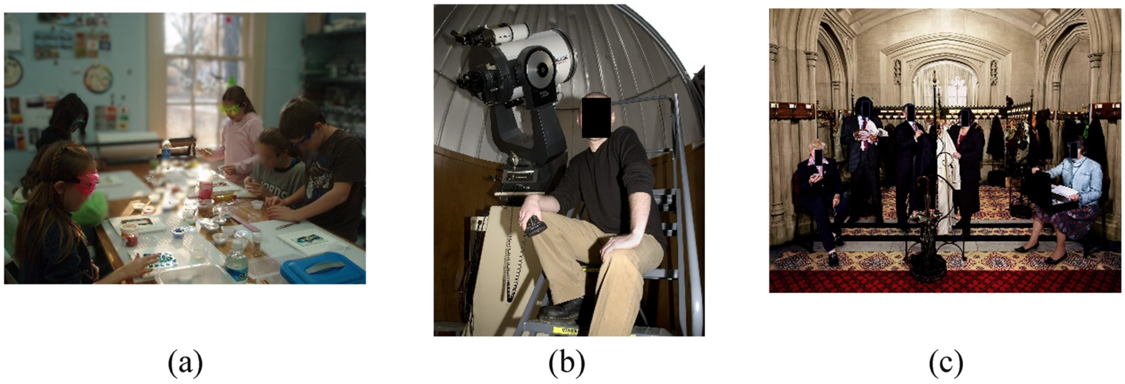

Figure 6: Example of Emotic dataset [17]. (a) Displays an image with multiple people in varying poses, (b) shows an example of an image with a clear background, and (c) includes an image with significant background clutter.

4.2 Loss functions and evaluation indicators

The loss function is a weighted combination of two individual losses, $L = \lambda_1 L_1 + \lambda_2 L_2$, L_1 and L_2 are the sum of 26 discrete affective losses and the sum of 3 continuous affective losses, respectively, and λ_1 and λ_2 are the weights of the discrete affective losses and the weights of the continuous affective losses, respectively. The discrete affective loss L_1 uses multi-label focal loss (MFL) [25] and the continuous affective loss L_2 uses Huber loss [26], defined as

$$\mathrm{MFL}_{a,\gamma}(\mathbf{y}^{\mathrm{disc}}, \hat{\mathbf{y}}^{\mathrm{disc}}) = -\sum_{i=1}^{26} \alpha(1 - \hat{\mathbf{y}}_i^{\mathrm{disc}})^{\gamma}\mathbf{y}_i^{\mathrm{disc}}\log(\hat{\mathbf{y}}_i^{\mathrm{disc}}) + (1 - \alpha)(\hat{\mathbf{y}}_i^{\mathrm{disc}})^{\gamma}(1 - y_i^{\mathrm{disc}})\log(1 - \hat{\mathbf{y}}_i^{\mathrm{disc}}), \tag{6}$$

$$\mathrm{Huber}_{\delta}(y^{\mathrm{cont}}, \hat{y}^{\mathrm{cont}}) = \sum_{i=1}^{3} \begin{cases} \frac{1}{2}x^2 \ \text{ for } \ |x| \leq \delta \\ \delta\left(|x| - \frac{1}{2}\delta\right), \text{otherwise} \\ x = y_k^{\mathrm{cont}} - \hat{y}_k^{\mathrm{cont}} \end{cases} \tag{7}$$

In equation (5), $\hat{\mathbf{y}}^{\mathrm{disc}}$ is the predicted value of the ith category, $\hat{\mathbf{y}}^{\mathrm{disc}}$ is the true label of the ith category, and a and y are two hyperparameters. In equation (6), $\hat{y}^{\mathrm{cont}}$ is the predicted value of the kth consecutive sentiment, y^{cont} is the true value of the kth consecutive sentiment, and is an empirical parameter. After many experiments, the recognition effect of the model reaches the best when $\alpha = 0.5$, $\gamma = 3$, and $\delta = 0.5$. The evaluation metrics are Precision (Precision) and mean absolute error (MAE). Furthermore, this article utilized the Adam optimization algorithm with a learning rate of 0.001, selected for its adaptive learning rate properties. The number of epochs was set to 100, to ensure the model adequately learns from the training data without overfitting, with 32 batch size. The choice of these hyperparameters was based on extensive experimentation aimed at balancing model performance and training efficiency.

4.3 Tests results

The sentiment accuracy and sentiment recognition qualitative results obtained by the sentiment recognition model on the Emotic dataset are shown in Figure 7, and the continuous sentiment MAE is shown in Table 1.

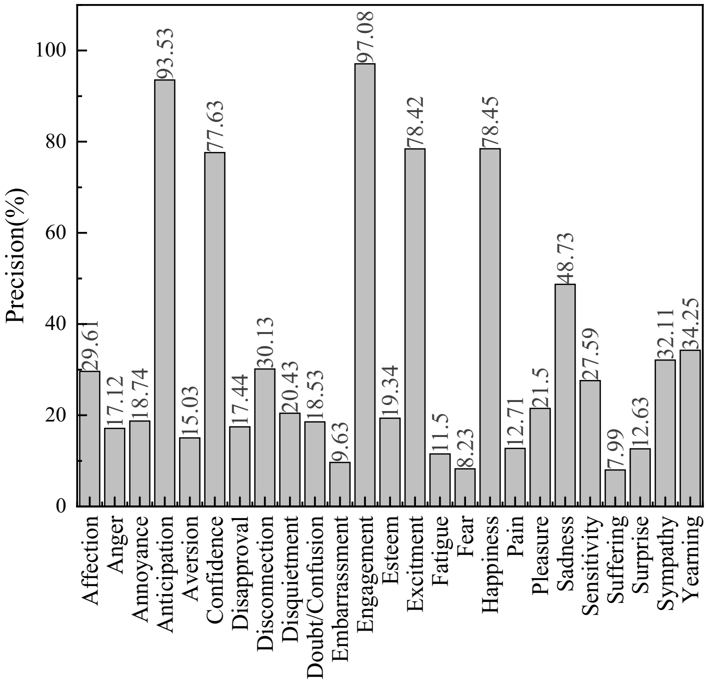

Figure 7: Emotional recognition precision results. Source: Created by the authors.

Table 1: MAE of emotic dataset

Continuous dimension	MAE
Valence	0.0756
Arousal	0.0895
Dominance	0.0879
MAE	0.0843

The analysis of Figure 7 and Table 1 reveals the final average accuracy rate achieved by the sentiment recognition model on the Emotic dataset to be 32.517%. Notably, the model exhibits exceptional performance in accurately classifying emotions associated with anticipation, engagement, confidence, excitement, and happiness, with accuracy rates soaring above 70%. This high level of precision underscores the model's efficacy in discerning these specific emotional states, which are crucial for understanding learner engagement and receptivity in smart educational environments.

4.4 Ablation experiment

Ablation experiments with different branch combinations are performed to ensure that the other parameters of the experiments are consistent. Experiment 1 has character body information, experiment 2 is a combination of character body information and environment semantic information, and experiment 3 is a combination of character body information and depth map information. The ablation experiments with different fusion strategies are carried out to ensure that the other parameters of the experiments are consistent, experiment 1 uses feature-level fusion, experiment 2 uses decision-level fusion, and experiment 3 uses hybrid-level fusion, and the average precision (AP) and mean absolute error (MAE) obtained are shown in Table 2. The AP and mean absolute error (MAE) are shown in Table 2. The experimental results highlight a notable improvement in the performance of the hybrid fusion strategy over its counterparts. Specifically, the hybrid fusion strategy exhibits a 3.1% increase in average accuracy compared to feature-level fusion and a 6.6% increase compared to decision-level fusion. Furthermore, this strategy demonstrates a significant reduction in the mean absolute error, decreasing by 0.0058 in comparison with feature-level fusion and by 0.035 in comparison with decision-level fusion.

Table 2: Experimental results on different fusion strategies

Experiments	AP rate	Mean absolute error
Feature-level fusion	30.042	0.0853
Decision-level fusion	26.547	0.1153
Hybrid-level fusion	33.143	0.0795

Figure 8 presents a comparative analysis of the precision rates achieved through different fusion strategies employed in our emotion recognition model. The strategies include feature-level fusion, decision-level fusion, and hybrid-level fusion. The results indicate that the hybrid-level fusion strategy outperforms the other two, with an AP rate of 33.143% and an mean absolute error (MAE) of 0.0795. In contrast, feature-level fusion and decision-level fusion strategies yielded lower precision rates of 30.042 and 26.547%, respectively, with corresponding increases in MAE. These findings underscore the effectiveness of the hybrid fusion strategy in enhancing emotion recognition performance, offering a robust approach for integrating various data sources, and improving model accuracy.

The proposed emotion recognition model is compared with the models proposed by Kolbadi and Sarvar [21], and the AP and mean absolute error (MAE) obtained are shown in Table 3.

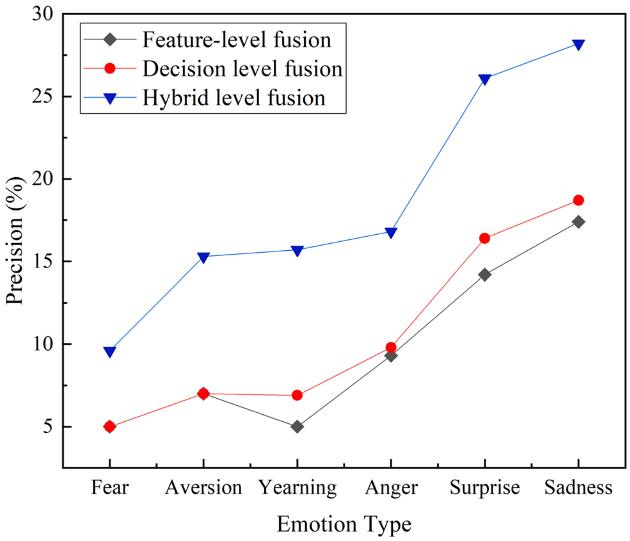

Figure 8: Comparison of precision of different fusion. Source: Created by the authors.

Table 3: Experimental results on different branch combinations

Emotion type	**Precision**		
	Kosti	**Lee**	**FB-DS**
Affection	27.85	17.83	29.61
Anger	9.49	4.57	**17.12**
Annoyance	14.06	5.51	**17.84**
Anticipation	58.64	52.75	**83.53**
Aversion	7.48	4.57	**15.03**
Confidence	**78.35**	64.96	77.63
Disapproval	14.97	7.24	**17.44**
Disconnection	21.32	17.71	**30.13**
Disquietment	16.89	13.97	**20.43**
Doubt/confusion	**29.63**	14.07	18.53
Embarrassment	3.18	2.19	**9.63**
Engagement	87.53	71.05	**89.35**
Esteem	17.73	12.5	**23.61**
Excitement	77.16	42.36	**77.74**
Fatigue	9.7	5.47	**10.8**
Fear	**14.14**	5.6	9.55
Happiness	58.26	49.71	**78.45**
Pain	8.94	3.5	**12.71**
Peace	**21.56**	16.19	21.5
Pleasure	45.46	30.11	**48.73**
Sadness	19.66	10.3	**27.59**
Sensitivity	**9.28**	4.9	7.99
Suffering	**18.84**	4.47	12.63
Surprise	18.81	7.81	**25.75**
Sympathy	14.71	10.64	**27.48**
Yearning	8.34	6.26	**15.37**
AP (%)	27.38	18.7	**32.52**
MAE	**0.057**	0.1	0.08

The bold values indicate the highest precision achieved for each emotion type among the three compared models: Kosti, Lee, and FB-DS. This highlights the superior performance of the FB-DS model in recognizing specific emotions compared to the other two models.

5 Conclusion

In this article, we address the problem of emotion recognition and interaction in smart education environments and propose an advanced approach based on deep learning networks to enhance the personalization and adaptability of the learning experience. A framework for comprehensively analyzing learners' emotional states is proposed by integrating dual-stream coding of facial expressions and scene contexts. The face is accurately extracted using a MTCNN, a deep learning model designed for simultaneous face detection, and the extraction and analysis of facial expression features is further enhanced by 3D-CNNs, which extend traditional convolutional networks by adding a temporal dimension to capture dynamic information over time. This integrated approach demonstrates its unique advantages when dealing with complex emotion data, especially when dealing with occlusion problems in uncontrolled environments. By combining attentional mechanisms and background context coding into the emotion recognition process, the study significantly improves the accuracy and robustness of emotion recognition. This integrated analytical approach allows the model to better understand and adapt to the learner's emotional needs, thus providing a more personalized and adaptive learning experience in smart learning environments. The specific findings are as follows:

This study successfully developed and validated an image sentiment analysis method based on dual stream coding of facial and background actions. The method combines MTCNN, 3D-CNN structure as well as CapsNet and its improved E2-CapsNet and SCapsNet, which effectively improves the accuracy and efficiency of emotion recognition. This method is able to comprehensively analyze human facial expressions and scene backgrounds, providing a new perspective to understand and analyze human emotions in videos.

By monitoring learners' expressions and reactions in real time, the proposed adaptive learning screen adjustment system is able to dynamically adjust the learning content and optimize the intelligent learning environment. Such a system not only provides a personalized learning experience, but also adapts itself to learners' visual–emotional preferences. This finding is of great significance for enhancing learning efficiency and improving the learning experience, and provides effective technical support for adaptive interaction at the emotional level in smart education environments.

Experimental results on the Emotic dataset show the excellent performance of this research method. The model shows high accuracy rate in the recognition of specific emotion categories, especially in the recognition of emotions such as anticipation, engagement, confidence, excitement, and happiness. In addition, through ablation experiments, this study also demonstrates the significant effects of different network components and fusion strategies on model performance, with the hybrid-level fusion strategy performing the most effective in improving model performance.

Overall, the methodology and findings in this article provide new perspectives in the field of smart education, especially in utilizing deep learning techniques to achieve affective-level adaptive interaction. Future research could explore the integration of multimodal data sources, such as auditory signals and physiological responses, to further enhance the accuracy and sensitivity of emotion recognition systems. Investigating the synergistic effects of combining these diverse data streams presents a promising avenue for developing more sophisticated, responsive, and personalized educational technologies. Additionally, future studies could address limitations related to the diversity of the dataset and the generalizability of the model across different educational contexts.

Funding information: The authors state no funding involved.

Author contributions: Wei Zhao wrote this article, and Liguo Qiu edited and revised the article.

Conflict of interest: The authors state no conflict of interest.

Data availability statement: The data that support the findings of this study are available from the corresponding author upon reasonable request. Figure 6 taken from https://github.com/rkosti/emotic.

References

[1] Zheng W, Tang H, Lin Z, Huang TS. Emotion recognition from arbitrary view facial images. Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part VI 11. Springer Berlin Heidelberg; 2010. p. 490–503. doi: 10.1007/978-3-642-15567-3_36.

[2] Barman A, Paramartha D. Facial expression recognition using distance and texture signature relevant features. Appl Soft Comput. 2019;77:88–105. doi: 10.1016/j.asoc.2019.01.011.

[3] Komalawardhana N, Patcharin P. Trends and development of technology-enhanced personalized learning in science education: a systematic review of publications from 2010 to 2022. J Comput Educ. 2023;11:1–22. doi: 10.1007/s40692-023-00276-w.

[4] Akhand MAH, Roy S, Siddique N, Kamal MAS, Shimamura T. Facial emotion recognition using transfer learning in the deep CNN. Electron. 2021;109:1036. doi: 10.3390/electronics10091036.

[5] Revina IM, Sam Emmanuel WR. A survey on human face expression recognition techniques. J King Saud Univ-Comput Inf Sci. 2021;33(6):619–28. doi: 10.1016/j.jksuci.2018.09.002.

[6] Saxena A, Ashish K, Deepak G. Emotion recognition and detection methods: A comprehensive survey. J Artif Intell Syst. 2020;2 (1):53–79. doi: 10.33969/ais.2020.21005.

[7] Lajevardi SM, Hussain ZM. Automatic facial expression recognition: feature extraction and selection. Signal Image Video Process. 2012;6:159–69. doi: 10.1007/s11760-010-0177-5.

[8] Wang G, Shucheng H, Zhe T. Shallow multi-branch attention convolutional neural network for micro-expression recognition. Multimed Syst. 2023;29:1–14. doi: 10.1007/s00530-023-01080-3.

[9] Li B, Dimas L. Facial expression recognition via ResNet-50. Int J Cognit Comput Eng. 2021;2:57–64. doi: 10.1016/j.ijcce.2021.02.002.

[10] Sabour S, Nicholas F, Hinton GE. Dynamic routing between capsules. Adv Neural Inf Process Syst. 2017;30:25–8. doi: 10.48550/arXiv.1710.09829.

[11] Azari Arani G, Ahmadi A, Azari Arani K. Geo-spatial analysis of the environment affects human health in Tehran. J Remote Sens Geoinf Res. 2023;1(2):127–34. doi: 10.22061/jrsgr.2023.2006.

[12] Tunisian MA. The use of local and local analysis and decision-making systems for urban design and development. 2023 Jun. doi: 10.22061/jrsgr.2023.2007.

[13] Ghasemi KA. Applications of satellite-based geodesy in navigation and earth monitoring. J Remote Sens Geoinf Res. 2023;1(2):143–51. doi: 10.22061/jrsgr.2023.2008.

[14] Pratama MP, Sampelolo R, Lura H. Revolutionizing education: harnessing the power of artificial intelligence for personalized learning. Klasikal: J Educ Lang Teach Sci. 2023;5:350–7. doi: 10.52208/klasikal.v5i2.877.

[15] Gandhi A, Adhvaryu K, Poria S, Cambria E, Hussain A. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Inf Fusion. 2023;91:424–44. doi: 10.1016/j.inffus.2022.09.025.

[16] Xu L, Ding X, Zhao D, Liu AX, Zhang Z. A three-dimensional resnet and transformer-based approach to anomaly detection in multivariate temporal–spatial data. Entropy. 2023;252:180. doi: 10.3390/e25020180.

[17] Kolbadi NM, Sarvar R. Land use investigation and its distribution analysis in various districts of Tehran city according to land use planning standards. J Remote Sens Geoinf Res. 2023;1(2):163–76. doi: 10.3390/e25020180.

[18] Heidarimozaffar M, Hosseini SA. Extracting FaçadePoints of urban buildings from mobile laser scanner point clouds. J Remote Sens Geoinf Res. 2023;1(2):153–62. doi: 10.22061/jrsgr.2023.1990.

[19] Abolali S, Silavi T, Saberian J. Improving location indices in design of oil transmission lines with an economic and environmental protection attitude. J Remote Sens Geoinf Res. 2023 Jun 22;1(2):177–88. doi: 10.22061/jrsgr.2023.2005.

[20] Haddad J, Olivier L, Philippe H. 3d-cnn for facial emotion recognition in videos. Advances in Visual Computing: 15th International Symposium, ISVC 2020, San Diego, CA, USA, October 5–7, 2020, Proceedings, Part II 15. Springer International Publishing; 2020. doi: 10.1007/978-3-030-64559-5_23.

[21] Ku H, Wei D. Face recognition based on mtcnn and convolutional neural network. Front Signal Process. 2020;4(1):37–42. doi: 10.22606/fsp.2020.41006.

[22] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86(11):2278–324. doi: 10.1109/5.726791.

[23] Luvizon DC, David P, Hedi T. Multi-task deep learning for real-time 3D human pose estimation and action recognition. IEEE Trans pattern Anal Mach Intell. 2020;43(8):2752–64. doi: 10.1109/TPAMI.2020.2976014.

[24] Niu Z, Guoqiang Z, Hui Y. A review on the attention mechanism of deep learning. Neurocomputing. 2021;452:48–62. doi: 10.1016/j.neucom.2021.03.091.

[25] Gao Z, Peng Q, Yong D. HAAN: Human action aware network for multi-label temporal action detection. Proceedings of the 31st ACM International Conference on Multimedia; 2023. doi: 10.1145/3581783.3612097.

[26] Chi HG, Lee K, Agarwal N, Xu Y, Ramani K, Choi C. AdamsFormer for spatial action localization in the future. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023. doi: 10.1109/CVPR52729.2023.01715.