Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 5, pp. 2231–2243, DOI 10.12694/scpe.v26i5.4738

ONLINE EDUCATION STUDENT COGNITIVE STATE RECOGNITION BASED ON IMPROVED MULTI-TASK CONVOLUTIONAL NEURAL NETWORK

WEIJUAN AN *, LI SHEN† AND YALI YUAN‡

Abstract. With the widespread application and development of internet technology in public education scenarios in China, the application of deep learning technology on online learning platforms is becoming increasingly widespread. This study aims to address the difficulty in determining students' cognitive state under the current network learning mode, and proposes a face recognition algorithm for student cognitive state detection using multitask convolutional neural network image recognition technology. At the same time, research is conducted on extracting two-dimensional feature points of facial color images through cascading regression tree localization methods. In the practical application experiments of the method, the research method can effectively detect images with facial offset angles greater than 15° for students, and the cognitive state of online learning can be analyzed from the frequency detection of students' blinking and yawning. From the results of image simulation experiments, it can be seen that this study proposes a cascaded regression tree localization optimization multitask convolutional neural network face recognition method, which has the highest image recognition accuracy of 86%, a recall rate of 0.85, and an f1 value of 0.855. The experimental results show that online learning state recognition based on image analysis can effectively monitor abnormal states during students' learning process, improve students' online learning efficiency, and provide necessary student state information support for teachers, promoting the improvement of online teaching quality.

Key words: Multi-task convolutional neural network; Cascaded regression tree; Network teaching; Cognitive load; Teaching link

1. Introduction. The profound impact of the epidemic on social, economic and cultural aspects has also had an impact on the education cause. In the past few years, online teaching situations have been less applied and gradually normalized. However, it is difficult to identify and improve students' learning status through face-to-face supervision and communication. Therefore, scholars at home and abroad have begun to apply machine algorithms and video image recognition technology in online teaching situations [1]. Based on this, the research uses the Multi-task convolutional neural network (MTCNN) algorithm to detect and recognize the face state of students. In this study, a depth-separable convolution structure is introduced for the case that the parameters of the traditional convolution layer are too large, and the median filtering method is used to filter the image [2]. In the recognition of students' cognitive load state, the research analyzes the degree of deviation of the face, the blinking frequency of the students' eyes, and the degree of opening and closing of the mouth. At the same time, the Ensemble of Regression Trees (ERT) algorithm is used to locate the key feature points of the face [3]. In the feature extraction stage, in order to integrate the camera applications in different scenarios in network teaching, the RGB-D image acquisition method of the depth camera is used to analyze the three-dimensional pixel relationship between the depth image and the color image. Different from the current conventional online video communication technology in facial expression recognition, in order to increase the adaptability of image recognition technology to classroom education, research has added indicators such as facial offset and blink frequency to image expression recognition to identify students' classroom learning situation [4]. The method proposed in the study is also different from traditional resource recommendation and learning needs identification models that are built around user data. This method is basically applied to learning scenarios in classroom teaching [5]. The innovation of this experiment lies in the combination of image recognition technology and the characteristics of students' cognitive state, so as to analyze the students' learning cognitive state. It is expected that the method proposed by the research can effectively improve the cognitive

*Shijiazhuang Information Engineering Vocational College, China (anweijuan80@126.com)
†Shijiazhuang Information Engineering Vocational College, China (Corresponding author, anweijuan80@126.com)
‡Shijiazhuang Information Engineering Vocational College, China (yuanyali2023@126.com)

load state of students and optimize the cognitive behavior of online teaching scientifically and effectively.

2. Literature review. In the process of technology driving the development of education, CNN methods that have been applied to image and video monitoring have gradually become more abundant. Minghui et al. used convolutional neural networks to perform privacy and confidentiality work on data communication in outsourcing environments, and optimized the pooling layer by linear averaging. Experiments show that this method can not only improve the information security of outsourced data, but also reduce the risk of privacy leakage [6]. After determining the parameter structure of the convolution layer, pooling layer and classification layer of the CNN network, Atik, I applied it to the classification of electronic components. After the simulation experiment, the research shows that the method has the highest performance of electronic component classification accuracy, and its accuracy rate is 98.99% [7]. Kiki et al. constructed a neurobiologically inspired CNN model and applied the model to the auditory spatial localization of human voice. Experiments show that this method is an effective method for human spatial hearing [8]. El-Shafai et al. proposed a spectrum sensing model for colleges and universities based on CNN network. Compared with the traditional SS method, the optimized spectrum sensing model improves the detection accuracy by up to 17%, and reduces the sensing time by 16.6ms [9]. Thanammal et al. designed a CNN network model based on cross-wind-driven optimization, and applied the model to the detection of tomato pests. Experiments show that the recognition accuracy of this method is 99.86%, and the iteration time is only 12.3s. The data proves that this method can help tomato agricultural production and planting for real-time detection of pests and diseases [10]. Ben et al. proposed a randomly initialized network architecture of RND-CNN and applied the method to detect COVID-19 in chest X-ray images. The experimental results show that the detection accuracy of the optimized method in the COVIDx dataset reaches 94% [11]. Rahimilarki et al. combined time series analysis technology with CNN network and applied it to fault identification and detection of wind turbines. Experiments show that the method is used in simulation experiments, and its fault detection classification accuracy is higher than 95%, and the detection performance is stable [12].

Aiming at the problems of low resolution and poor sharpness of lens less cameras, Zhang et al. proposed an image reconstruction technique that fuses text proposal network and Convolutional Recurrent Neural Network(CRNN). Experiments have confirmed its applicability [13]. Fernandes et al. proposed a variable convolution backbone-based cascaded architecture to address text extraction difficulties in electronic documents. Experiments show that the extraction accuracy of the method in the two datasets is 99.3% and 95.1% respectively [14]. Shuai TENG et al. proposed a method based on CNN image detection for the structural stability of steel frame buildings. By fusing feature extraction and classification blocks into an intelligent learning system, the structural state of steel frames is detected. Experiments show that the classification accuracy of this method is as high as 99%, while the time used is only 19% of the traditional BP method [15]. MohsinNadia combines Canny and Prewitt edge detection techniques to construct a method for improving image embedding capacity in edge regions. Experiments show that this method has a high embedding capacity, which means that the imperceptibility of steganographic images is maintained [16]. Kamil et al. proposed a face detection method for attendance scenarios, which utilizes SVM algorithm and OpenCV software for image processing. In the experiment, this method can achieve an accuracy of about 81.8%. Compared to the proposed method, the research constructed method can achieve an accuracy of approximately 86% [17]. Andrejevic et al. applied facial recognition technology in campus environments, with the main purpose of addressing issues such as offline campus safety, automatic registration, and student emotion detection. This study mainly considers the feasibility and reasons for the application of technology. The research and construction methods are mainly used for student state identification in online classrooms [18].

To sum up, the CNN network is widely used in image recognition, indicating that the image detection technology has become mature. However, it is rare to combine image detection technology with cognitive state detection of students. Therefore, the research will use the optimized MTCNN image recognition technology to locate and feature extraction of key points such as students' faces, mouths, and eyes to distinguish students' cognitive behaviors and cognitive load states. The innovative contribution of the research lies in the analysis of details such as changes in students' eyes, mouth, and head displacement through image detection technology, and the construction of a detection system for cognitive fatigue in online education for students. The system uses mouth opening and closing, blink frequency, and head offset angle as quantitative indicators to measure

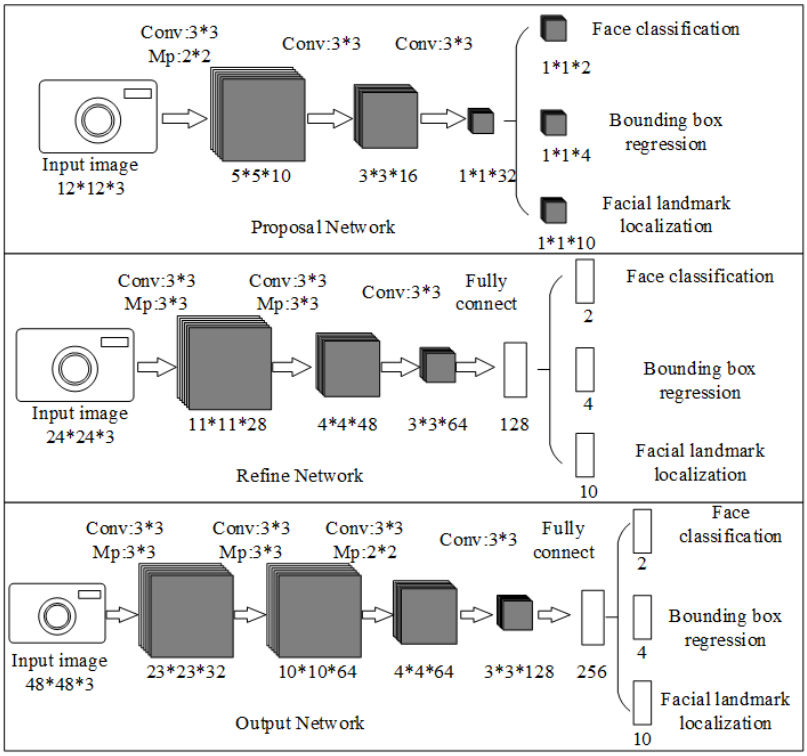

Fig. 3.1: MTCNN's three-layer structure

students' cognitive fatigue status, thereby improving the data sparsity problem in current cognitive fatigue detection. The purpose of the research is to provide a scientific and effective monitoring method of student status for the gradually normalized network teaching.

3. Research on the state recognition algorithm of students in online teaching.

3.1. Construction of online teaching face detection method based on convolutional neural network. Multi-task convolutional neural network (MTCNN) can combine face region detection and face key point detection. At the same time, the algorithm has the idea of candidate frame and classifier, which can detect faces quickly and accurately [19]. Therefore, the core algorithm of cognitive state detection of students in online teaching proposed in this study is MTCNN. MTCNN consists of three layers of network structure, among which the Proposal Network structure is responsible for quickly generating candidate boxes that may contain faces. The Refine Network network structure performs high-precision filtering and selection on the generated candidate boxes to improve the accuracy of detection. The final output network structure is responsible for generating the final facial bounding boxes and feature regression of keypoints, that is, determining the specific position of the face and the precise coordinates of keypoints. The three-layer structure of MTCNN is shown in Figure 3.1.

As can be seen from Figure 3.1, in the above MTCNN standard convolution layer, after obtaining the size of the input feature image, the number of channels and the number of convolution layers, the parameters of the convolution layer can be calculated. The calculation formula is as follows Formula (3.1) is shown.

$$n_p = W_F \times H_F \times D_{im} \times N \tag{3.1}$$

In formula (3.1), it n_p represents the amount of convolution parameters, W_F, , $H_F$and , D_{im}represent the width, height, and number of channels of the feature map of the convolution layer, which are the number $N$of layers of the convolution layer. It can be seen from the above calculation formula that when the size of the image

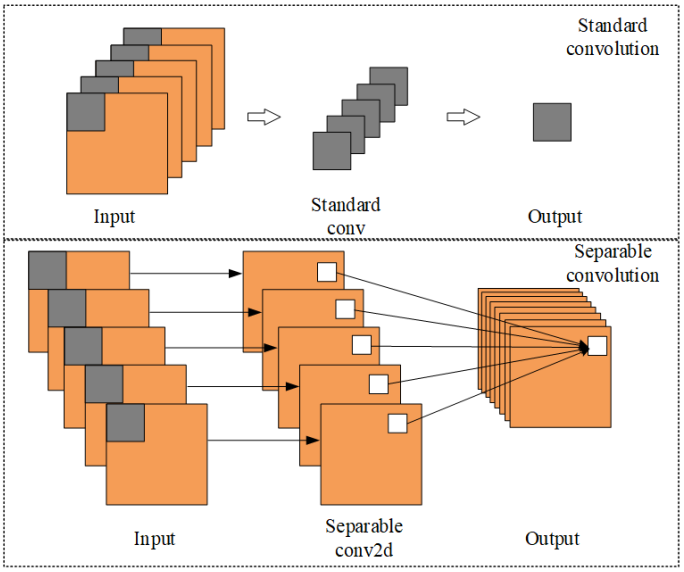

Fig. 3.2: MTCNN Standard Convolution Process and Depth Separable Convolution Process

is large, the total amount of parameter calculation data of the convolution layer is huge, so the research will optimize the standard convolution of traditional MTCNN through depth wise separable convolution, aiming to reduce the volume The amount of product parameters increases the processing efficiency of feature maps [20]. The standard convolution process of MTCNN and the depth wise separable convolution process are shown in Figure 3.2.

It can be seen from Figure 3.2 that the calculation of the parameter amount of the depth wise separable convolution is expressed as $n_p = W_F \times H_F \times D_{im} + D_{im} \times N$, comparing the parameter amount of the standard convolution and the depth wise separable convolution, the result is shown in formula (3.2).

$$\frac{W_F \times H_F \times D_{im} + D_{im} \times N}{W_F \times H_F \times D_{im} \times N} = \frac{1}{N} + \frac{1}{W_F \times H_F} \tag{3.2}$$

It can be seen from equation (3.2) that the parameter amount of the depth wise separable convolution is smaller than the standard convolution parameter amount. In the online teaching environment, not all images have faces displayed in the image detection stage. Therefore, the study introduces the Focal Loss loss function for uneven sample classification, and its optimization principle is to increase the weight according to the difficulty of sample classification. First, the loss function of the MTCNN algorithm in face detection consists of classification, boundary regression and key point feature regression loss functions. Therefore, in the image detection and classification, the face classification and other classifications are expressed as a two-class cross entropy loss function, and its formula is shown in formula (3.3).

$$L^c = -(y^c \log(p) + (1 - y^c)(1 - \log(p))) \tag{3.3}$$

In formula (3.3), it L^crepresents the loss function of face classification, which is $y^c \in \{0, 1\}$the face classification of the image, 0 represents the non-face category, and 1 represents the face category. pRepresents the probability that the image sample is a face. In the traditional cross-entropy loss function, the classification does not consider the difficulty of image classification. Once non-face images account for a large proportion of image samples, it is easy to cause imbalance of sample categories. Therefore, the weight and modulation factor of the Focal Loss loss function are introduced to optimize the image classification loss function as formula (3.4).

$$L^c = -\sigma(1 - p)^\rho \log(p) \tag{3.4}$$

In formula (3.4), $\sigma \in [0,1]$represents the weight factor, which is used to measure the proportion of positive and negative samples, $(1-p)^{\rho}$and represents the modulation factor, which is used to distinguish the difficulty of sample classification. The face frame regression loss function represents the distance between the predicted result of the face regression frame and the actual face frame, so the loss function calculation is completed by the Euclidean distance calculation formula. In the same way, the loss function of the key point of the face feature represents the difference between the pre-positioning of the key point and the actual position, so it is the same as the frame regression loss, which is represented by the Euclidean distance. The specific loss functions of the two are shown in formula (3.5).

$$\begin{cases} L^b = \left\| \widehat{y}^b - y^b \right\|_2^2 \\ L^l = \left\| \widehat{y}^l - y^l \right\|_2^2 \end{cases} \tag{3.5}$$

In formula (3.5), $L^b$and L^lrepresent the face frame regression loss function and the face key point feature regression loss function, respectively. $\widehat{y}^b$ If it is the predicted coordinates of $y^b$the frame, it is the actual coordinates of the frame; if it is the $\widehat{y}^l$ pre-positioning coordinates $y^l$of the key points of the face, it is the actual coordinates of the key points. Combining the classification, boundary regression loss function and face key point regression loss function by introducing weights, the overall loss function of MTCNN face detection is obtained as shown in formula (3.6).

$$L^t = \min \sum_{i=1}^{N} \sum_{j \in \{c,b,l\}} \alpha_j \beta_i^j L_i^j \tag{3.6}$$

In formula (3.6), $L^t$is the overall loss function, $N$is the total number of image samples, $i$represents the number of the image samples, is the number of $j$the three tasks of classification, frame regression, and key point regression, $\alpha_j$represents the task weight, $\beta_i^j \in \{0,1\}$is the sample label, and $L_i^j$is the loss function of the task . When the image enters the convolutional layer for calculation, due to the different image shooting environments of online learning, there are a lot of interference factors in the image, so the research will optimize the image filtering operation. The research and application method are the median filter method for smoothing. The pixels in the image window are sorted according to the gray value, and the median value is taken to replace the pixel value. The formula is expressed as formula (3.7).

$$g(x,y) = mid\{f(x-a, y-b)\}; (a,b) \in S \tag{3.7}$$

In formula (3.7), it $g(x,y)$represents the planting filter pixel, which $f(x,y)$is the original noisy pixel, which $f(x-a, y-b)$represents the domain pixel of the original pixel, which Srepresents the filter. Finally, MTCNN detects the facial situation of online learning students, including image denoising, generating image pyramid structure; determining face candidate frame; deleting unselected candidate frame; using sink to delete candidate frame with high degree of overlap [21]. The applied parameters include the image scaling factor, the minimum detectable face image size, the selection threshold of the human frame, and the screening and intersection ratio (Intersection over Union, IoU) of the Non-Maximum Suppression (NMS) algorithm.) threshold. Its calculation formula is shown in formula (3.8).

$$IoU = \frac{[(Cx2 - Cx1) \times (Cy2 - Cy1)] \cap [(Gx2 - Gx1) \times (Gy2 - Gy1)]}{[(Cx2 - Cx1) \times (Cy2 - Cy1)] \cup [(Gx2 - Gx1) \times (Gy2 - Gy1)]} \tag{3.8}$$

In formula (3.8), $[(Cx2 - Cx1) \times (Cy2 - Cy1)]$the area of the region C is expressed, and the area of $[(Gx2 - Gx1) \times (Gy2 - Gy1)]$the region G is expressed.

3.2. Facial feature analysis of students' cognitive load status. The cognitive load state of students refers to the physical exhaustion caused by students investing their own cognition and energy resources to receive information and process them in the learning process. The measurement of cognitive load in academics is difficult to achieve through network technology. Therefore, the study summarizes the facial performance in the academic theory of cognitive load state, and uses the specific facial key state features to detect the MTCNN algorithm to identify the academic fatigue state. and cognitive load [22]. In this study, the two-dimensional feature points of the color image of the face are extracted by the feature point location method of the cascade regression tree (Ensemble of Regression Trees, ERT) algorithm. The principle of the ERT algorithm to extract feature points is to connect the strong regressors in series through two-level cascaded regression to build a feature mathematical model. The iterative formula is shown in Equation (3.9).

$$\begin{cases} \hat{S}^{t-1} = \hat{S}^t + \psi_t(I, \hat{S}^t) \\ \Delta S_i^t = S_i + \hat{S}^t \end{cases} \tag{3.9}$$

In formula (3.9), it $\hat{S}^t$represents $t$the predicted shape vector of the stage regressor, which $\psi_t$represents the $t$stage regressor, $I$is the input image, ΔS^trepresents $t$the difference between the prediction of the stage regressor and the actual result, and $i$represents the data sample number. The shape vector of facial feature points is updated by an iterative formula, and the number of regression cascade layers in the first layer is the number of regressors. In the regression calculation of the second layer, the regressor is trained by the method of enhancing the gradient, so the initialization mathematical model of the regressor is shown in formula (3.10).

$$f_0(I, \hat{S}_i^t) = \arg\min \sum_{i=1}^{N} \left\| \Delta S_i^t - \psi \right\|^2 \tag{3.10}$$

In Equation (3.10), it $f_0(I, \hat{S}_i^t)$represents the fitting of the residual regressor of the initial regressor, and the value of the gradient update is set r_{ik}, and its calculation formula is shown in Equation (3.11).

$$r_{ik} = \Delta S_i^t - f_{k-1}(I_i, \hat{S}_i^t) \tag{3.11}$$

In formula (3.11), it $f_{k-1}(I_i, \hat{S}_i^t)$represents $k-1$the fitting of the first-level strong regressor to the residual regressor. $k$represents the number of regressors. Finally, a strong regressor is constructed through a weak regressor, and the final output of the iterative convergence is expressed as Eq. (3.12).

$$\begin{cases} f_k(I, \hat{S}^t) = f_{k-1}(I, \hat{S}^t) + v g_k(I, \hat{S}^t) \\ R(I, \hat{S}^t) = f_K(I, \hat{S}^t) \end{cases} \tag{3.12}$$

In Equation (3.12), $v \in [0, 1]$denotes the learning rate, and $g_k$denotes the weak regressors that constitute the strong regressor, Kdenotes the number of weak regressors. After the two-dimensional positioning of the feature points by the ERT algorithm, the study analyzes and models the cognitive load characteristics. Firstly, students' learning cognitive load status is divided into mouth features and eye features [23]. After obtaining the two-dimensional positioning of the mouth and eyes through the ERT algorithm, the distance change between the two points is calculated to detect the frequency of students blinking and yawning. Finally, in the extraction of 3D feature points of the face, the RGB-D image acquisition method of the depth camera is studied. Corresponding the information of the color image and the depth image, the relationship change formula is shown in formula (3.13).

$$\begin{cases} p_{de} = H_{de} P_{de} \\ p_r = H_r P_r \end{cases} \tag{3.13}$$

Formula (3.13), $p_{de}$is the projection coordinate of the object on the depth image, is the H_{de}internal parameter of the depth camera, and $P_{de}$is the spatial coordinate of the object in the depth image. $p_r$is the projected coordinate of the object on the color image, is the $H_r$internal parameter of the color camera, and $P_r$is the

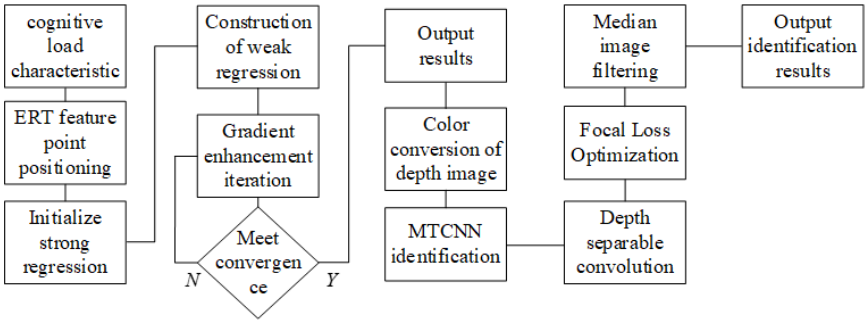

Fig. 3.3: MTCNN algorithm identifies the flow of academic cognitive load status

spatial coordinate of the object on the color camera. Therefore, for the point coordinates of the object in space P, there is a relational expression as shown in Equation (3.14).

$$\begin{cases} P_{de} = R_{de}P + T_{de} \\ P_r = R_r P + T_r \end{cases} \tag{3.14}$$

In formula (3.14), it (P, R, T)represents the three-dimensional coordinates of the object, and the coordinates of the depth image can be converted into the coordinates of the color image through the formula. Finally, the calculation formulas such as precision, recall, and accuracy verify the image detection performance of the MTCNN algorithm for students' online classrooms. The specific formulas are shown in formula (3.15).

$$\begin{cases} pre = \frac{TP}{TP+FP} \\ Re = \frac{TP}{TP+FN} \\ Acc = \frac{TP+TN}{TP+TN+FP+FN} \\ f1 = \frac{2pre \cdot Re}{pre+Re} \end{cases} \tag{3.15}$$

In formula (3.15), it TPrepresents the number of correctly identified samples, the number of FPincorrectly recognized samples, the FNundetected wrong samples, and $TN$the undetected correct samples. The core idea of this research is to firstly summarize the common facial manifestations such as blink frequency, eye closure degree, yawn frequency, etc. under students' cognitive load state. Secondly, the ERT algorithm is used to determine the two-dimensional positioning of key points such as the mouth or eyes. Thirdly, the spatial coordinates of the depth image and color image are converted by the depth camera RGB-D image acquisition method, and finally the image is recognized by the MTCNN algorithm model. Classification. Therefore, the overall flow of the algorithm is shown in Figure 3.3.

As can be seen from Figure 3.3, the optimization of the algorithm model in this experiment includes the depth wise separable convolution of the MTCNN model, which is used to reduce the number of parameters in the standard convolution. At the same time, the Focal Loss method is also introduced to optimize the model loss function, which is expected to increase the performance of MTCNN in student cognitive state monitoring through optimization.

4. Application Analysis of MTCNN State Recognition Algorithm for Cognitive Load in Online Teaching.

4.1. Simulation Experiment of MTCNN Image Recognition. In order to verify the image detection performance of the MTCNN network teaching student state recognition algorithm constructed in the research, the research will carry out simulation experiments from the open network face image data set. This simulation training experiment is carried out on the TensorFlow platform. The three-layer structure parameters of the MTCNN algorithm optimized by filtering optimization and depth wise separable convolutional network parameters are shown in Table 4.1.

Table 4.1: Three-layer structure parameters for optimizing MTCNN algorithm

P-Net		R-Net		O-Net	
base-Ir	0.001	base-Ir	0.001	base-Ir	0.001
batch-size	384	batch-size	384	batch-size	384
epochs	30	epochs	twenty-two	epochs	twenty-two
momentum	0.9	momentum	0.9	momentum	0.9
Lr - factor	0.1	Lr - factor	0.1	Lr - factor	0.1
Lr - epoch	6	Lr - epoch	14	Lr - epoch	20

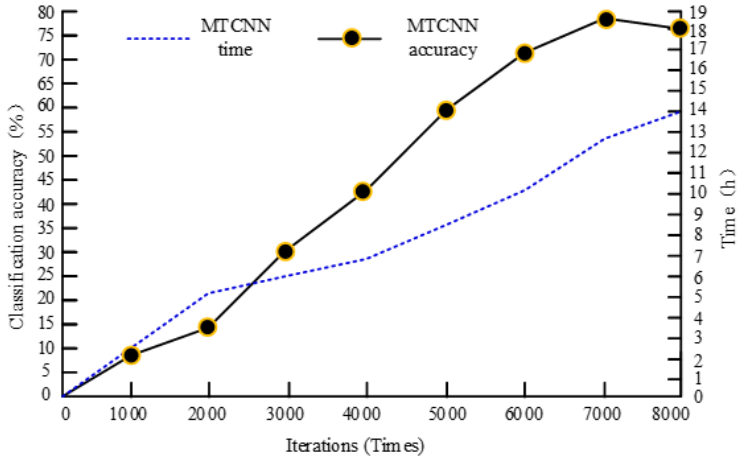

Fig. 4.1: Optimizing the overall training recognition accuracy and time consumption of MTCNN model

In Table 4.1, the size of the training image is 12*12*3, and its main task is to classify the image as containing or not containing faces, so its classification loss function is used for R-Net 's bounding box regression loss and O-Net The key point regression loss is not required, so the weight parameters of the overall loss function of the α_jP-Net simulation training are set to 1, 0.5, and 0.5 in the three tasks, respectively. At the same time, the maximum number of training samples is set to 8000 times, and the overall recognition accuracy and time-consuming of the model are shown in Figure 4.1.

It can be seen from Figure 4.1 that when the image size is 12*12*3, the overall recognition accuracy of the MTCNN algorithm in this simulation increases with the number of iterations. After the number of iterations exceeds 7 000, the classification accuracy increases. The trend turned to decline. In the simulation training, the highest classification accuracy in the MTCNN algorithm model is 77.8%, and the time to complete 8000 trainings is 14.1h. The experimental results show that the MTCNN algorithm has high classification accuracy performance in face image recognition, and at the same time, due to the optimization of the depth wise separable convolution, the simulation training takes less time. After the simulation training, the research will validate the MTCNN algorithm after training in the Emotional State Dataset for Online Education (DAiSEE) and the 2018 Emotion Recognition Challenge Dataset (EmotiW2018). The sample size distribution of the dataset and the performance analysis of the MTCNN algorithm are shown in Table 4.2.

It can be seen from the table that there is a large amount of data in the emotional state data set of online education. Among the 1813 sample identifications in the validation set, the precision, recall rate and f1 value of the MTCNN algorithm are 0.9076, 0.6923, and 0.7855, respectively. In the 48 data sets of the 2018 Emotion Recognition Challenge dataset, the precision, recall and f1 value of the MTCNN algorithm were 0.69, 0.89, and 0.78, respectively. Experiments show that the recognition accuracy performance of MTCNN is higher than 75% in data sets with large and small data samples, indicating that the recognition ability of MTCNN algorithm

Table 4.2: Sample size distribution of data sets and performance analysis of MTCNN algorithm

Data set	Sample distribution		MTCNN effect	
DAiSEE	train	5441	pre _	0.9076
	verification	1813	Re	0.6923
	test	1814	f 1	0.7855
EmotiW2018	train	147	pre _	0.69
	verification	48	Re	0.89
	test	64	f 1	0.78

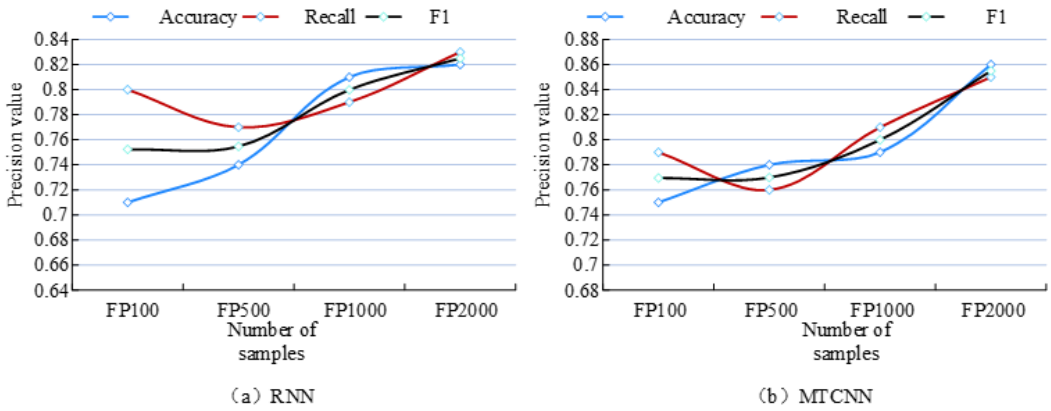

Fig. 4.2: Image recognition performance of RNN and MTCNN algorithms with different dataset sizes

is stable. Finally, the research will compare the image recognition performance of the recurrent convolutional neural network and the multi-task convolutional neural network. The number of samples in this test set is 100, 500, 1000, and 2000 four stages to compare the recognition performance of the two algorithms. The result is shown in Figure 4.2.

The RNN in the figure represents a recurrent convolutional neural network, while MTCNN is a multi-task convolutional neural network optimized for research. In the comparison of the results in Figure 4.2, it can be seen that the overall image recognition performance of the MTCNN algorithm is higher than that of the RNN algorithm. When the number of samples in the dataset increases, the recognition performance of the two algorithms is slightly improved. When the number of samples in the test dataset is 100, the accuracy of the recurrent convolutional neural network is 0.71, while the recognition accuracy of the MTCNN algorithm is 0.75; the recall rate and f1 value of the RNN algorithm are 0.8 and 0.752, respectively, but the recognition accuracy of the MTCNN algorithm is 0.8 and 0.752, respectively. In the test, the recall and f1 value of the optimized algorithm were 0.79 and 0.769, respectively. When the maximum number of samples in the test dataset is 2000, the recognition accuracy of the recurrent convolutional neural network is the highest 0.82, the recall rate is 0.83, and the f1 value is 0.825. The optimized multi-task convolutional neural network achieves the highest recognition accuracy of 0.86, recall of 0.85, and f1 value of 0.855. Experiments show that the performance of the optimized MTCNN algorithm is higher than that of the recurrent convolutional neural network in network image recognition.

4.2. Application Analysis of Cognitive Regulation Strategies in Online Teaching Based on MTCNN. The practical application of the MTCNN optimization algorithm takes a social science course in the MOOC forum as the experimental object. In the online teaching video, the course implementation process is divided into two categories: student-led and teacher-led, and the course content is divided into teaching platforms and other topics. 6 content categories: questioning, discussion, discussion results integration,

Table 4.3: The influence of different links on students' cognitive behavior

link	s1	s2	s3	s4	s5	s6
t1	11.98	2.69	-2.83	-0.52	-0.81	-0.64
t2	1.12	17.48	0.94	-1.46	-2.29	-5.6
t3	-0.94	0.19	61.22	-22.94	-6.42	-39.77
t4	-1.63	0.01	-21.97	67.48	-17.9	-35.44
t5	-0.78	-3.23	-8.04	-16.06	54.36	1.31
t6	0.86	-5.04	-39.88	-35.53	1.03	87.27
link	s1	s2	s3	s4	s5	s6
s1	34.06	-0.21	6.46	-0.02	-0.06	5.1
s2	-0.42	5.29	0.79	-0.17	-0.46	-2.28
s3	-1.93	1.98	-1.07	-0.79	-2.08	-11.61
s4	-1.6	-3.74	-3.44	1.53	-1.73	-9.93
s5	-0.55	0.7	-0.68	-0.22	1.18	-3.46
s6	1.58	-1.81	-2.33	-0.55	1.25	-2.98

solution narrative, and social background narrative. The cognitive concentration degree of different course links is quantified from the students' learning cognitive behavior, and the moderating effect of the arrangement of different course links on students' cognitive load is analyzed. The students' teaching platform and other topics, questioning, discussion, discussion results integration, solution narrative, and social background narrative are classified as s1-s6, and the six teacher-led links are divided into s1-s6. From t1 to t6, the following sequence analysis method is used to calculate the influence of different links dominated by different subjects on students' cognitive behavior. The specific results are shown in Table 4.3.

It can be seen from the table that the teaching platform guided by teachers and other topic links, questioning link, discussion link, discussion result integration link, solution method narrative link, and social background narrative link all mobilize students' cognitive behavior. It has a positive impact, and the effect is most significant in the results integration link and the social background narrative link. At the same time, in the process of student-led communication with teachers, the frequency of students' cognitive behavior is higher than other states. To sum up, the experiments show that teacher-guided discussions and summaries have a scheduling effect on students' cognitive behavior. Teachers can effectively improve students' cognitive load by asking questions, exchanging information, and by expressing questions and showing ideas. Therefore, after the optimization of the implementation sequence through the course links, the effect of improving students' cognitive behavior can be analyzed through the MOOC teaching videos, and 80 videos showing the student's status with a length of 10s and the students after the adjustment of the teaching implementation link are displayed. Ease to perform MTCNN image recognition analysis. The cognitive state of students is analyzed through the eye, mouth and face recognition proposed in the research method. The specific results of the face part are shown in Figure 4.3.

The ordinate in FIG. 4.3 is the student's face offset angle, and the student's face offset represents a negative angle to the left and a positive angle to the right. When the student's face offset angle is between -15° and 15°, it indicates that the student is facing the front screen and his cognitive behavior is in a state of concentration, while the student's face is greater than 15° and less than -15°, it indicates that the students are in a state of cognitive load. It can be seen from the dot matrix distribution diagram in Figure 6(a) that 34 of the top 80 students in the teaching process are in a state of cognitive load with excessive facial offset. In Figure (b), only 5 of the 80 students have a facial offset of less than -15° after the optimization and adjustment of the teaching process. Experiments show that the improved regulation of online teaching links proposed in the study has a significant positive impact on students' cognitive behavioral attention. Finally, the changes of the students' mouth and eyes are also compared, and the specific results are shown in Figure 4.4.

Figure 4.4 shows the changes in the eyes and mouth of middle school students during video monitoring. In the online teaching, this study expressed the students' blink rate of more than 25 times and less than 10 times per minute as the cognitive load state. In the change of the mouth, the ratio of the vertical and horizontal distance of the mouth, which represents the opening and closing degree of the mouth, is used to distinguish

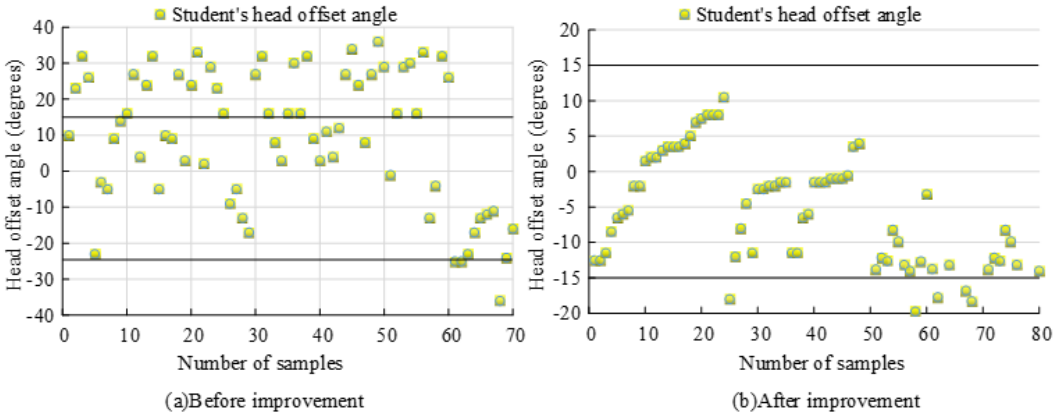

Fig. 4.3: Comparative analysis of students' facial deviation before and after teaching optimization

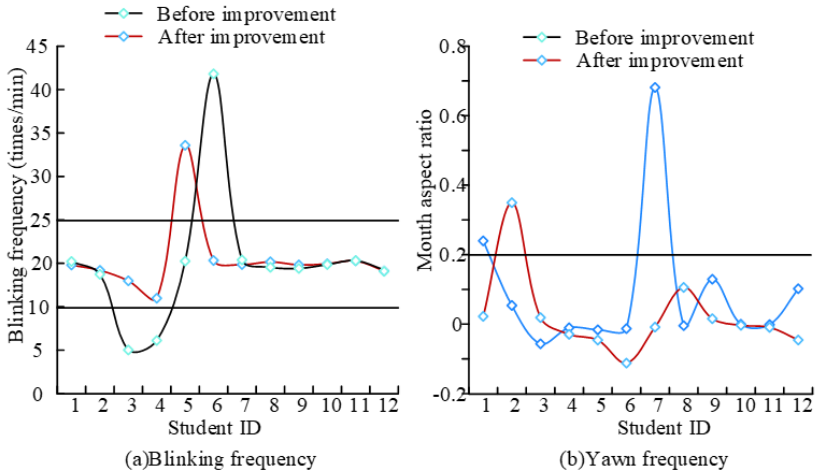

Fig. 4.4: Comparative analysis of changes in students' mouth and eyes before and after teaching optimization

the speaking state and the yawning state of the students. The case where the mouth aspect ratio is greater than 0.2 is classified as yawning, and the case where the ratio is less than 0.2 is classified as speaking or the mouth is tightly closed. It can be seen from the sub-figure (a) that among the blink frequencies in the tired and dazed states, 3 of the 12 students' video samples are in a state of cognitive load, but after the adjustment of the online teaching link, only A classmate's blink rate is 33 times per minute. In sub-figure (b), among the 12 students before the course optimization, there were 2 students whose mouth aspect ratio was greater than 0.2, but after the course optimization, the students' cognitive load status improved, and only one classmate's mouth was Aspect ratio greater than 0.2. Experiments show that after the optimization and adjustment of the course link of online teaching, the blinking frequency of students is gradually in a normal state, rather than a state of fatigue that is too fast and a state of daze that is too slow; at the same time, the situation of students yawning is also reduced.

5. Conclusion. In the context of online teaching, it is difficult for students to understand their cognitive behaviors and cognitive states through communication. Therefore, the research proposes a cognitive feature extraction method combined with MTCNN image recognition, and conducts simulation experiments and prac-

tical applications. In the simulation experiment, the highest classification accuracy of the MTCNN algorithm model is 77.8%, and the time to complete 8000 training sessions is 14.1h. At the same time, when comparing the RNN algorithm, its accuracy and f1 value are respectively 4% and 3% higher. In the practical application experiment, the study proposes to improve the cognitive load state of students and mobilize students' cognitive behavior by adjusting the links of online teaching courses. Through sample case analysis, it is found that 34 of the first 80 students have facial deviations. The angle is greater than 15° or less than -15°, and after adjustment, only 5 students' faces are not concentrated in front. In the video recognition of blink frequency and yawn state, 3 students in the video samples of the first 12 students are in a state of cognitive load, but after the adjustment of the online teaching link, only 1 student has a blink frequency of 33 times per class. minute. Before the course optimization, 2 of the 12 students had a mouth aspect ratio greater than 0.2, but after the course optimization, the students' cognitive load status improved, and only one student had a mouth aspect ratio greater than 0.2. Experiments show that the optimized method of MTTCNN network identification and students' cognitive feature extraction can help optimize and adjust courses in network teaching and improve students' learning status. The shortcomings of the research are that the detection system has high hardware requirements and the production cost is relatively expensive. In future research, efforts will be made to adjust the adaptability of the detection model to other hardware, thereby reducing device costs and improving the applicability of the method.

REFERENCES

[1] Xiangyu Liu. (2021). Validation research on the application of depthwise separable convolutional facial expression recognition in non-pharmacological treatment of BPSD. *Clinical Nursing Research*, 5(4), 31–37.

[2] Fu, H., Guan, J., Jing, F., …. (2021). A real-time multi-vehicle tracking framework in intelligent vehicular networks. *China Communications*, 18(6), 89–99.

[3] Liu, J., Xu, X., Shi, Y., Deng, C., Shi, M. (2022). RELAXNet: Residual efficient learning and attention expected fusion network for real-time semantic segmentation. *Neurocomputing*, 474(Feb.14), 115–127.

[4] Ye, F. (2022). Emotion recognition of online education learners by convolutional neural networks. *Computational Intelligence and Neuroscience*, 2022(1), 4316812.

[5] Kaur, R., Gupta, D., Madhukar, M., Singh, A., Abdelhaq, M., Alsaqour, R., Goyal, N. (2022). E-Learning Environment Based Intelligent Profiling System for Enhancing User Adaptation. *Electronics*, 11(20), 3354.

[6] Li, M., Chow, S. M. S., Hu, S., Yan, Y., Shen, C., Wang, Q. (2022). Optimizing privacy-preserving outsourced convolutional neural network predictions. *IEEE Transactions on Dependable and Secure Computing*, 19(3), 1592–1604.

[7] Atik, I. (2022). Classification of electronic components based on convolutional neural network architecture. *Energies*, 15(7), 2347–2361.

[8] Kiki, V., Mehrkanoon, S. (2022). Goal-driven, neurobiological-inspired convolutional neural network models of human spatial hearing. *Neurocomputing*, 470(22), 432–442.

[9] El-Shafai, W., Fawzi, A., Sedik, A., Zekry, A. M., El-Banby, G. M., Khalaf, A. A. M., Abd El-Samie, F. E., Abd-Elnaby, M. (2022). Convolutional neural network model for spectrum sensing in cognitive radio systems. *International Journal of Communication Systems*, 35(6), e5072.1–e5072.22.

[10] Indu, V. T., Priyadharsini, S. (2022). Crossover-based wind-driven optimized convolutional neural network model for tomato leaf disease classification. *Journal of Plant Diseases and Protection*, 129(3), 559–578.

[11] Ben Atitallah, S., Driss, M., Boulila, W., Ben Ghezala, H. (2022). Randomly initialized convolutional neural network for the recognition of COVID-19 using X-ray images. *International Journal of Imaging Systems and Technology*, 32(1), 55–73.

[12] Rahimilarki, R., Gao, Z., Jin, N., Zhang, A. (2022). Convolutional neural network fault classification based on time-series analysis for benchmark wind turbine machine. *Renewable Energy*, 185(Feb.), 916–931.

[13] Zhang, Y., Wu, Z., Lin, P., Wu, Y., Wei, L., Huang, Z., Huangfu, J. (2022). Text detection and recognition based on a lensless imaging system. *Applied Optics*, 61(14), 4177–4186.

[14] Fernandes, J., Simsek, M., Kantarci, B., Khan, S. (2022). TableDet: An end-to-end deep learning approach for table detection and table image classification in data sheet images. *Neurocomputing*, 468(Jan.11), 317–334.

[15] Teng, S., Chen, G., Wang, S., Zhang, J., Sun, X. (2022). Digital image correlation-based structural state detection through deep learning. *Frontiers of Structure and Civil Engineering*, 16(1), 45–56.

[16] Mohsin, N. A. (2021). A hybrid method for payload enhancement in image steganography based on edge area detection. *Cybernetics and Information Technologies*, 21(3), 97–107.

[17] Kamil, M.H.M., Zaini, N., Mazalan, L., Ahamad, A.H. (2023). Online attendance system based on facial recognition with face mask detection. *Multimedia Tools and Applications*, 82(22), 34437–34457.

[18] Andrejevic, M., Selwyn, N. (2020). Facial recognition technology in schools: Critical questions and concerns. *Learning, Media and Technology*, 45(2), 115–128.

[19] Adelson, C., Jordan, M. I., Muller, R. (2022). SOUL: An energy-efficient unsupervised online learning seizure detection classifier. *IEEE Journal of Solid-State Circuits*, 57(8), 2532–2544.

[20] Li, Q., Xu, L., Yang, X. (2022). 2D multi-person pose estimation combined with face detection. *International Journal of*

Pattern Recognition and Artificial Intelligence, 36(2), 2256002.1–2256002.23.
[21] Iqbal, K., Abbas, S., Khan, M. A., Ather, A., Khan, M. S., Fatima, A., Ahmad, G. (2021). Autonomous parking-lots detection with multi-sensor data fusion using machine deep learning techniques. *Computers, Materials and Continuum*, 67(2), 1595–1612.
[22] Aria, M., Agnihotri, V., Rohra, A., Sekhar, R. (2020). Secure online payment with facial recognition using MTCNN. *International Journal of Applied Engineering Research*, 15(3), 249–252.
[23] Sharma, N., Gupta, S., Mohamed, H.G. (2020). Siamese convolutional neural network-based twin structure model for independent offline signature verification. *Sustainability*, 14(18), 11484.

Edited by: Nasrullah Sheikh
Special issue on: Transformative Horizons:
The Role of AI and Computers in Shaping Future Trends of Education
Received: Aug 27, 2024
Accepted: Feb 5, 2025