Hindawi
Scientific Programming
Volume 2021, Article ID 9944176, 11 pages
https://doi.org/10.1155/2021/9944176

Research Article

Research on Learning State Based on Students' Attitude and Emotion in Class Learning

Dong Huang and WeiXin Zhang

The College of Education, Yunnan Normal University, Kunming 650000, Yunnan, China

Correspondence should be addressed to Dong Huang; 2011010118@st.btbu.edu.cn

Received 28 October 2021; Revised 8 November 2021; Accepted 11 November 2021; Published 7 December 2021

Academic Editor: Bai Yuan Ding

In basic education, timely and accurate grasp of students' classroom learning status can provide real-time information reference and overall evaluation for teachers and managers, which has a very important educational application value. At present, a lot of information technology is applied in the analysis of classroom student behavior state, and the state analysis technology based on a classroom video has the characteristics of strong timeliness, wide dimension, and large capacity, which is especially suitable for the analysis and acquisition of students' classroom state, and attracts the attention of major educational technology companies. However, the current student state acquisition technology based on video analysis lacks large scenes and has low practicability, and finally, the video-based student classroom behavior state analysis technology mainly focuses on a single behavior feature, which cannot fully reflect the student's classroom behavior state. In view of the above problems, this study introduces the face recognition algorithm based on a student classroom video and its implementation process, improves the hybrid face detection model based on a traditional model, and proposes the neural network algorithm of student expression recognition based on a visual transformer. The experimental results show that the proposed algorithm based on students' classroom videos can effectively detect students' attention and emotional state in class.

1. Introduction

It has always been difficult for teachers and administrators to keep track of all students' classroom learning. In traditional education, in order to better educate students, teachers conduct after-class analysis through traditional methods such as teaching diaries, teaching files, watching videos, and homework analysis, and then provide solutions according to the results [1]. However, the traditional after-school analysis method not only increases the burden of teachers but also makes it difficult to ensure the comprehensiveness, objectivity, and real time of the analysis results. With the rapid development of economy and the continuous promotion of education informatization in all countries around the world, teachers are in urgent need of an intelligent classroom learning state analysis tool to help teachers get the learning state of all students in class, deal with and display the overall state of the classroom after class, and then reflect on and improve their teaching process [2]. Intelligent analysis means have important practical significance for improving teachers' professional level and students' learning effect.

In recent years, education in most countries has shifted from elite education to mass education. The widespread popularization of education aims at improving the quality of education in the whole society. The evaluation system of classroom learning has always been accompanied by mass education, but students' test scores are often taken as the absolute criterion for evaluation. As a result, we only know the quality of the results, but do not know the causes of the results, and we cannot find appropriate adjustment schemes to improve the quality of education [3]. The students took notes seriously in class and actively communicated with the teacher, which reflected that the teacher's teaching content was attractive. On the contrary, students bow their heads or look out of the window for a long time with negative facial expressions. These states indicate that they do not understand the content of the class or do not concentrate on the class, which indirectly indicates that the way of teaching cannot attract students well. Usually, school administrators do not consider students' performance in class and only rely on students' scores and leaders to check the classroom situation. Such unilateral evaluation of teachers' teaching quality is not accurate, nor can it help teachers to understand the real learning situation of students. Therefore, in this case, the analysis of the classroom teaching process of students listening to the status of education is crucial [4].

There are many reasons why students' learning efficiency is not good in the process of classroom teaching. As the subject of receiving knowledge in the whole process, students' learning status in class can be studied as an important evaluation index of students' learning efficiency in class. At the same time, it is also the key factor to realize the effective teaching of teachers. Teachers hope to master each student's classroom learning situation comprehensively and accurately in the classroom teaching process, so as to make corresponding adjustments to the teaching content and teaching progress, so as to achieve more efficient teaching purpose. Generally speaking, if a student listens to the teacher carefully, most of his attention is directed toward the teacher during class; that is, he looks up and listens attentively to the teacher, and his emotional state is concentrated.

The existing classroom surveillance cameras are basically installed in the front and back of the classroom. By analyzing the images obtained from the cameras in front of the classroom, students' listening status can be judged. The detection of students' learning state in the classroom scenario is divided into two steps. First, the head position of the students is detected, and then, the head state of the students is further identified to see whether the head is looking up to listen to the teacher or looking down at the mobile phone or doing other things [5]. However, at present, many teachers understand students' situation in classroom learning through classroom observation and questioning, which often leads to the lag and one-sidedness of classroom information transmission and feedback. In particular, with the popularity of electronic devices such as smartphones and tablet computers, a large number of "phubbers" have emerged in classroom teaching [6]. Therefore, the combination of statistics and analysis of students' "head-up rate" in class and intelligent algorithm analysis of students' emotional state can judge students' class concentration to a certain extent, thus helping teachers effectively improve classroom teaching efficiency [7].

2. Related Work

Student's state of learning is an important index of students' classroom learning efficiency. The state of students' classroom learning generally refers to whether students look up at the teaching content for a long time and actively communicate with the teacher, whether they take notes carefully, and whether their facial expressions are in a positive or negative state [8]. Wearable devices are invasive to some extent and will inevitably have a certain influence on the subjects. There is a gap between the data obtained and the real state of students' classroom behavior. In addition, wearable devices are expensive, large in size, and complicated in the process of wearing, so it is difficult to popularize them in practical classroom education [9]. With the rapid development of smart devices, another method to collect video, image, voice, and other digital signals through cameras, microphones, and other devices has become widely popular. This method extracts information from these digital signals, such as students' facial expressions, natural language, and body posture, and finally processes, analyzes, and integrates this information to get students' classroom behavior state. The classroom student state analysis technology based on video images only needs to use the classroom camera system, which is of low cost and less invasive and has almost no influence on the learning process of students [10, 11]. Through the artificial intelligence algorithm, students' learning status can be analyzed in real time, comprehensively and multidimensionally. Through the above analysis, students' classroom behavior, emotional state, learning state, and other situational information in the class can be captured by intelligent monitoring equipment and mobile learning devices. Therefore, the current intelligent classroom obtains students' physiological signals and behavioral state data through various devices, and then collects and analyzes these data to get the current students' classroom behavior state, so as to better grasp the classroom situation and timely adjust the teaching strategy to improve the teaching effect. [12].

With the deepening of the research, the scholars have analyzed the research status of the analysis of students' classroom behavior state from two aspects, physiological signals and visual images, according to the different ways of collecting students' characteristics. Physiological signals refer to when people's inner emotions change, the body or brain will send out one or more physiological signals; through the collection and analysis of these physiological signals, such as EEG signals, EMG signals, skin temperature, and eye movement, one can know the current student's mood and psychological state [13]. Nourbakhsh et al. [14] proposed to detect the cognitive load level of students by analyzing the skin signals in the time domain and frequency domain. They collected the skin signals generated by learners in the process of completing learning tasks of different difficulties in experiments and then analyzed these skin signals in the time domain and frequency domain. By comparing the spectral features of skin signals in different difficult learning tasks, we found that the frequency domain features of skin signals had better identification ability for emotional cognition categories. Zhan [15] combined pupil size, blink frequency, and blink frequency with facial expression to construct the recognition framework of learners' emotional state. An intelligent teaching agent evaluated students according to the arousal dimension, interest dimension, and pleasure dimension in the framework and then made corresponding cognitive feedback, such as knowledge point proposal and learning suggestion. This combination of learners' eye movement tracking and facial expression recognition can enable the intelligent teaching agent to more accurately identify the emotional state and cognitive state of distance learners.

Sinha et al. [12] proposed the use of brainwaves and other physiological signals to track and detect learners' cognitive and emotional states during learning. This method uses electroencephalogram (EEG) wave signal to estimate learners' difficulty in understanding the learning content and uses heart rate variability and RGS signals collected by skin electrodes to estimate learners' emotional state, which is compared with learners' current academic performance. Zhu et al. [16] use smart wristbands of wearable devices to extract physiological signals of learners, collect and analyze students' handwriting status and heart rate activity through smart wristbands, and then obtain learners' current cognitive status. This method adopts the method of multisignal synthetic judgment, so the result is relatively accurate. All of the above studies need to obtain the physiological data of learners through wearable collection devices and analyze the classroom behavior state of students by using different physiological performances of people in specific states. Due to the accuracy and specificity of physiological signals, very accurate analysis results can be obtained. However, due to the use of complex wearable devices, the testers will establish a psychological preset and know that they are in a tested state, which will affect the objectivity of the results of physiological signal analysis [17]. Moreover, the cost, size, and deployment requirements of wearable acquisition devices make it difficult for such classroom behavior state analysis methods to be widely used.

The development of image recognition depends on the progress of image equipment (the progress of intelligent equipment such as HIGH-DEFINITION cameras has promoted the progress of image recognition research). The whole process has a lot of work and many links, so the final results are often biased or even wrong. With the deepening of research, intelligent analysis methods based on visual images are increasing. In this method, video images are first captured by the camera, and then, the data are input into the algorithm to identify, record, and analyze the students' expressions, head posture, and other explicit actions, and finally, the current classroom behavior state of the students is given. According to the different behavior characteristics of students, there are mainly four methods based on face detection, head posture estimation, facial expression recognition, and multiple action recognition, which are discussed next.

2.1. Methods Based on Face Detection. Fujisawa and Aihara [13] estimated learners' interest in learning by detecting the transformation of face size. In the experiment, the face detection algorithm uses the front face detector in the OpenCv [19] open-source vision library to conduct face detection through the camera directly above the computer. The experiment proves that the number of times the face is close to the screen and the entertainment of the material are closely related to the learner's interest. OpenCv visual open-source library contains a large number of tools for computer vision, image processing, behavior recognition, and other related fields. Hou et al. [20] proposed the application of face detection technology to the quality assessment of students' lectures in 2016. Haar-like face features that have been trained in OpenCv open source visual library were selected in the experiment, and these features were applied to the AdaBoost cascade algorithm for face detection, and the classroom head-up rate of students was calculated by detecting the number of faces, and the average classroom head-up rate of students in a fixed time was obtained.

2.2. Method Based on Head Posture Estimation. Rahman et al. [21] proposed to track the learning state of learners according to their head posture and the distance between learners and Kinect. In the experiment, a Kinect motion camera was used to obtain the information of the learner's head posture angle and distance depth, and then interest expression function was constructed. This method used a lot of physical knowledge to calculate and then tracked the learner's interest.

2.3. Methods Based on Facial Expression Recognition. Psychological research shows that positive emotions promote cognitive activities during learning, whereas negative emotions hinder cognitive activities. The research results of psychologist Mehrabian [22] show that emotional information consists of 7% language, 38% voice, and 55% facial expression, so students' emotional states can be obtained through the recognition and analysis of facial expression. Feng et al. [23] used 16 Haar-like features to extract face features, learned and trained classifiers with the AdaBoost algorithm, and cascaded strong classifiers to form the final expression classifier. Facial expression recognition is carried out through the facial expression classifier, and facial expression recognition technology is brought into the remote classroom, which realizes the facial expression recognition and emotion judgment system under the network environment for the first time, and improves the efficiency of online teaching and user satisfaction. Cheng et al. [24] selected 34 feature points to define facial geometric features. After marking feature points, Gabor wavelet was used to extract facial feature information, and SVM (support vector machine) was used to classify expressions to obtain expression classifier. The structure model of the intelligent teaching system based on expression recognition and sight tracking technology is proposed. Sun et al. [25] obtained facial expression classification by combining facial AU unit and third-order tensor. In the experiment, AU facial unit was used to eliminate the influence of individual differences on facial expression recognition effect and improve the accuracy of facial expression recognition. By separating facial features from personal facial features, the function of facial recognition and emotional intervention can be realized with high precision. Jiang et al. [26] used a variety of algorithms to identify and study the "confused" expressions of students in the learning process and concluded that the random forest algorithm has the best effect on identifying the confused expressions of students.

2.4. Multifeature-Based Analysis Method. With the development of deep learning, Whitehill et al. [27] marked the degree of participation of students as four levels and collected AU units, hand movements, and head posture information of students' faces by a gaussian wave filter and a support vector machine. Through the continuous training of the audit network, it is concluded that the movements of head lowering, side head, mouth, and eyes have great weight in judging students' participation. Han et al. [28] used AAM (active appearance model) to mark face feature points and then marked key points in the training set and utilized principal component analysis (PCA), extracting average shape by dimensionality reduction as shape model. The researchers studied the tilt of the head and the position of the lip and eye features during class and obtains the data of students "listening," "understanding," "doubt," "resistance," and "disdain." To examine each state of the head posture and to validate a specific analysis of students' expression, a classroom assessment of analysis of facial expression and head posture was conducted. Chen et al. [29] established a random forest model to identify students' head posture and facial expressions and used the teacher-student interaction platform to record the learning interaction between teachers and students in class. Although scholars have conducted a lot of studies, most of them are still based on traditional networks, and there are few recent applications of deep learning neural networks. Based on this, this study integrates traditional methods with the latest methods.

3. Face Recognition Based on Hybrid Architecture

Face detection is a mature aspect in the field of image processing. Scholars have proposed various algorithms for different data sets. Therefore, this study proposes a face detection algorithm under the hybrid architecture based on the characteristics of the row and column distribution of the classroom. The algorithm uses an algorithm with high detection accuracy as the fine detection algorithm and an algorithm with high detection speed as the rough detection algorithm. The algorithm calls the fine detection algorithm and the rough detection algorithm according to different face conditions.

3.1. Description of Algorithm. Face detection algorithm is to use the detection window on the image-intensive multiscale sliding and then determine whether the image in the detection window is a face. The goal of face detection is to find the corresponding positions of all the faces in the image. The output of the algorithm is the coordinate of the outer rectangle of the face in the image and may also include posture information such as tilt angle. The face detection algorithm should first have a large number of samples, then extract face features in the positive samples for learning, and then put into the model for training and finally through the verification results.

3.2. Data Preparation. Before face detection, we need a lot of data to preset rules, telling the machine that images with certain features are human faces, whereas those with other features are nonhuman faces. The diversity of positive samples can make the algorithm have correct answers in different scenarios, whereas negative samples can make the algorithm more accurate to exclude other nonface answers that are very close to human faces. The diversity of datasets can make the algorithm be used in different scenarios and ensure the robustness of the algorithm. The widely used ones are shown in Table 1.

3.3. Feature Extraction. The early template matching algorithm is based on the geometric features of the face to determine whether it is a face; with the improvement of the algorithm, more and more facial feature methods are proposed. The Haar features proposed by Papageorgiou et al. [34] are trained as face features by traversing images with different Haar rectangular frames. Haar features have good modeling ability in uneven lighting scenarios. Zhu et al. [35] proposed the histogram of oriented gradient (HOG), which is resistant to light changes by calculating and counting the histogram of gradient direction in local areas of the image to constitute features. Local binary pattern (LBP) proposed by T. Ojala et al. describes the local texture features of images through operators; it has significant advantages such as rotation invariance and gray invariance. In addition, features such as scale-invariant feature transformation and integral channel feature are used for face detection.

3.4. Model Training. Model training is an important part of face detection, is through the algorithm to input the face features of the training model, and then can directly call the model for face detection. At present, the algorithm commonly used for face detection is the support vector machine algorithm, and its main principle is the sample vector through function transformation mapping to the high-dimensional space and then in the high-dimensional space mapping to find the maximum interval of the interface. Freund proposed the AdaBoost algorithm, which is a classical iterative optimization algorithm. The core idea of its application in target detection is to take the target feature as a weak classifier, combine several weak classifiers into a strong classifier according to certain rules, and finally connect the strong classifier in series to carry out target detection and classification. Then, the classical convolutional neural network was proposed. The convolutional neural network could learn the features of the detected target independently and classify the detected target in the output layer after passing through the convolutional layer and pooling layer.

Table 1: Face recognition dataset.

Data set	Basic components
LFW [30]	The dataset includes the name of the image, the location information of the boundary box, and the location information of the key feature points of the face in the image
FDDB [31]	It is a sample picture of a face rotated and occluded by different lighting and resolution. The 2,845 images included 5,171 annotated faces.
AFLW [32]	It is a large-scale face database including multipose and multiview
COCO [33]	COCO dataset is a large, rich object detection, segmentation, and subtitle dataset, mainly taken from complex everyday scenes.

The hybrid face detection algorithm relies on the steadiness of the student's position and calls the fine and rough face detection algorithms according to different face conditions. At the beginning of the class, the students' face location information was collected for the first time through the rough detection algorithm, and using this information, students' seat, that is, the static position of students, is drawn. In the following detection process, we use the face rough detection algorithm for the first face detection, the detection of the face position information, and the student static position area for comparison. If there is no face detected in the student static position area, the second layer of face fine detection algorithm is called for face detection in the student static position area. If a student static position area uses the fine detection algorithm to detect no face many times, it is considered that the position of the student disappeared and the student static position coordinates are deleted. After that, students' basic state was judged by gesture recognition, and the algorithm structure is shown in Figure 1.

3.5. Classroom Behavior Status Assessment. The facial expression and head posture are combined to analyze the classroom behavior state of the students. The students' expressions obtained by using the expression recognition model of the convolutional neural network are divided into positive emotions and negative emotions. The head posture estimation algorithm was used to divide the recognized head posture of the students into nine directions, and the attention of the students was judged by the difference between the head posture of the target student and the head posture of the surrounding students. In the process of classroom teaching, students need to read books, take notes, and answer questions constantly. Therefore, head posture cannot completely represent students' learning status and can only assist other algorithms to make more detailed judgments.

Head pose estimation (HPE) usually refers to the identification of head position and direction parameters in a spatial coordinate system. The direction parameter refers to the degree of rotation in the three coordinate axes of the spatial coordinate system. The three direction parameters are yaw, pitch, and roll. Head pose estimation is to calculate the head direction parameters by comparing the face feature points in the digital image with the corresponding feature points in the general 3D model. The face feature points in the digital image are obtained by using the Dlib68 feature point detector on the basis of face positioning, and the flow is given next:

3.5.1. Data Acquisition and Processing. The classroom teaching videos of natural environment are obtained through the camera (model), and the images are extracted by OpenCv. The width and height of images are W and H. Then, camera calibration is carried out, where the internal parameter matrix is to transform 3D camera coordinates to 2D homogeneous image coordinates. OpenCv camera calibration function is used to calibrate the camera. The internal parameters of the camera are as follows:

$$A = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{1}$$

where f_x, f_y is the focal length of the camera. Generally, the focal length of the camera selects the width and height of the image, which c_x, c_y represents the offset of the camera's optical axis in the image coordinate system. Generally, the center point of the image is selected.

3.5.2. Feature Detection and Head Pose Estimation. The mixed face detection was used to detect the face, and then, the public Dlib68 feature point detector was used to obtain the coordinate information of six feature points of the face: the outer corner of the left eye and the outer corner of the right eye, the tip of the nose, the left lip angle, the right lip angle, and the tip of the chin. The 2D/3D mapping was obtained by solving the PnP (perspective-n-point) problem, and the rotation and translation vectors of the head pose were output. Then, the student's head pose value (X, Y, Z) is obtained by converting the flip vector into the Euler angle.

3.6. Attention Judgment. People's visual attention refers to the object or gaze direction of people's eyes, and students' attention can be judged by the difference in their eyes. First, the nose tip coordinates obtained by the feature point detector were used as the starting point, and the nose tip coordinates in the 3D coordinates of the face were used as the ending point to draw the students' line of sight. Although the students' line of sight is different, but all the students are

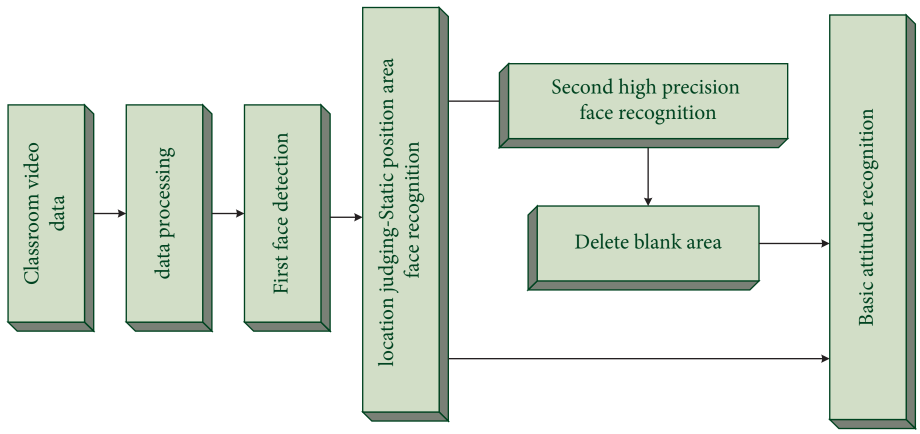

Figure 1: Face recognition algorithm structure diagram.

facing the blackboard, and only a very few students' line of sight is toward other positions in the classroom, the students' special behavior can be obviously judged according to the students' line of sight, and the students are in a state of inattention at this time with great probability. The students' head posture was divided into nine directions, and the attention of the students was judged by comparing the line of sight between the target student and the surrounding students.

3.6.1. Partitions. Get the three-dimensional coordinate of student's frontal head posture $[X, Y, Z]$.

3.6.2. Euclidean Distance. The face coordinate of the current student is taken as the starting point, and the face coordinate of other students in the face dataset is taken as the ending point:

$$D_i = \sqrt{(x_0 - x_i)^2 + (y_0 - y_i)^2}. \quad (2)$$

3.6.3. Screening and Judgment. Since all the students in the class are facing the blackboard, the eyes of the students on both sides are not the same, so the eyes of the students on the left side have no reference significance to the right side, and the eyes on the right side are the same. Set the aisle distance as A, delete the faces of the students whose D value is greater than A in the face dataset, sort the faces of the remaining students according to the size of D_i, and get A set whose distance from the target student is from near to far.

To judge the state of a single student, this study selected a group of 9 people and judged the attention of the target student by the difference between the facial orientation of the 8 students and the target student. Eight students are selected from the set of faces in ascending order obtained in Step 3 for comparison. If the value of the target student's face orientation is the same as that of the target student's face direction, then the target student's attention value is increased by one to compare one by one and save the last attention value. It is considered that the attention value of students is greater than 5 and the attention rate of students in the class is obtained by calculating the value of all students and comparing it with the number of students.

3.7. Facial Expression Recognition. Research by psychologist Mehrabiai shows that emotional messages consist of 7 percent of words, 38 percent of voices, and 55 percent of facial expressions. In this study, the convolutional neural network is used to train the facial expression recognition model and the recognized facial expressions are divided into positive emotions and negative emotions according to psychology. The expression recognition model based on the convolutional neural network can get rid of the traditional algorithm to extract the display features of each expression. By combining the extraction of each expression feature with the fuzzy classification of the network, the expression recognition model can improve the performance and generalization ability.

The network is mainly optimized based on the ResNet network structure, because the ResNet network structure can well extract the features of the image, and the optimization of the network structure can make the network better extract the features of the facial region. The overall structure of the model can be divided into three parts: feature extraction, relationship modeling, and expression classification. The optimized ResNet is used as the backbone network to extract features. The weight of the extracted feature is calculated by a layer of self-attention mechanism, and the weight obtained is multiplied by the feature matrix to obtain the final feature matrix. The eigenmatrix is then flattened and projected onto specific dimensions as input to the transformer. Then a transformer encoder is used to model the relationship between face regions. The network eventually calculates the expression of the input image through a simple Softmax function. The facial expression recognition neural network established in this study is shown in Figure 2.

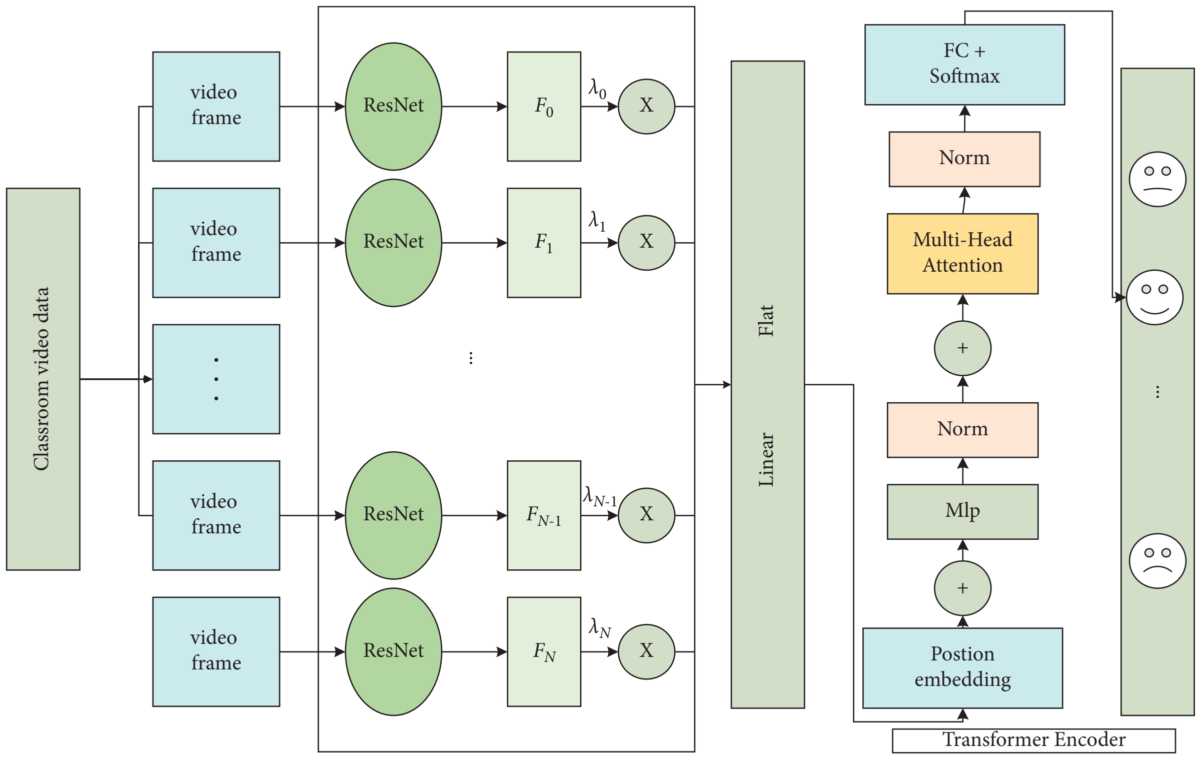

Figure 2: Facial expression recognition neural network.

(1) Image cutting: The MTCNN [21] model is used for face location. According to the output results of the MTCNN model, the face is cropped. In order to model the relationship between nonocclusion face regions, the cropped face region needs to be segmented into uniformly segmented images with a size of 20×20 pixels. These segmented images are sent to the feature extraction network for feature extraction. In order to avoid the loss of boundary information in the segmentation process, the image is divided into overlapping image blocks in this study. Each image block has repeated pixels with the surrounding blocks, increasing the correlation between image blocks. Then, these image blocks are sent into the feature extraction network.

(2) Feature extraction: ResNet introduces identity mapping into the network, which can solve the problem of network model degradation and gradient disappearance with the increase in network depth [18], thus improving network performance. ResNet works as follows: suppose the input is x and a certain network layer is set as H. The original network is learning output $h(x)$. After the identity mapping is introduced, the original input x is transmitted to the output through shortcut connections. At this point, the network only needs to learn the residual $f(x) = h(x) - x$ of input and output, and the problems of previous models can be solved through residual learning. Feature extraction is the small-size image after processing. If the maximum pooling operation is still used, some global features may be lost, so SoftPool [22] is adopted in this study to replace maximum pooling. Compared with other pooling operations, SoftPool can retain both global and local information during pooling. Recognition works better. The calculation method of SoftPool is shown in the following formula:

$$\widetilde{a} = \sum_{i \in R} \frac{e^{a_i} * a_i}{\sum_{j \in R} e^{a_j}}. \tag{3}$$

(3) In order to better study the facial expressions of students, this study does not directly use Softmax for classification, but also needs to send the extracted features into the transformer for further feature extraction. Therefore, the full connection layer of the last layer is modified, and the original full connection layer is changed into two full connection layers, 512 and 100 dimensions. Finally, the obtained 100-dimensional feature vectors are sent into the transformer as tokens for training, which can retain more nonlinear features and effectively reduce the occurrence of overfitting phenomenon. Through experiments, feature extraction using SpResNet can effectively improve the accuracy of recognition.

(4) Vision Transformer is a model proposed by Google in 2017. Originally used in natural language processing tasks, a transformer relies on the attention mechanism and can make the network pay attention to certain words selectively. Later, Carion [13] introduced a transformer into the field of computer vision and proposed an end-to-end target detection

model, DETR. By combining CNN and transformer, the predicted results are finally output. Google proposed a new Vision Transformer(ViT) [19], which migrates the transformer originally used for NLP task into visual classification task out of the box, using transformer instead of CNN. Finally, excellent results have been achieved in large-scale datasets.

Unlike a traditional transformer, which receives serialized tokens as input, Vision Transformer's input is a 3D image, Therefore, the original 3D image data $x \in R^{H\times W\times C}$ need to be divided into image blocks and then the picture is expanded into a one-dimensional vector $x_p \in R^{N\times (p^2 c)}$, where (H, W) is the resolution of the original image, C is the number of channels of the image, and $(p^2 c)$ is the size of each image sequence. Finally, these vectors are flattened to the model size and the output $x_p e$ is embedded, This is followed by the addition of an additional classification header to the sequence, which is a learnable embedding vector through which classification is ultimately performed. Since each image block has a certain position in the uncropped image, location coding needs to be added to the sequence to retain location information. The calculation method is shown in the following formula:

$$z_0 = \left[x_0, x_p^1 e, x_p^2 e, \ldots, x_p^n e\right] + e_{\text{pos}}, \tag{4}$$

where e_{pos} is position embedding and z_0 is the initial input to the transformer. The transformer consists of multiheaded self-attention mechanisms and MLP blocks, each of which is followed by a LayerNorm (LN) layer. The calculation method for the transformer is shown as follows:

$$\begin{aligned} z'_\ell &= \text{MSA}\left(\text{LN}\left(z_{\ell-1}\right)\right) + z_{\ell-1}, \\ z_l &= \text{MLP}\left(\text{LN}\left(z'_\ell\right)\right) + z'_\ell, \\ y &= \text{LN}\left(z_L^0\right), \end{aligned} \tag{5}$$

where L is the number of image blocks and $\ell = (1, \ldots, L)$. Multiheaded self-attention (MSA) is the core mechanism of transformer. It is composed of single-headed attention mechanism, namely, self-attention (SA). The calculation method of single-headed attention is shown in the following formula:

$$\text{Attention}\,(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{6}$$

where Q is a query, K is a key, and V is a value. They are linear variations of input tokens. d_k is the dimension of K. By calculating the dot product, the similarity between different tokens can be calculated, thus obtaining the global long-term concern, which is conducive to modeling the relationship between clipping regions. Multiheaded attention mechanism is a series of K single-headed attention output, and the calculation method is as follows:

$$\text{MSA}\,(z) = \left[SA_1(z), \ldots, SA_k(z)\right]U_{\text{msa}}, \tag{7}$$

where $U_{\text{msa}} \in R^{kd_k\times D}$ and $SA_1(z)$ is the single-headed attention mechanism.

4. Experimental Analyses

4.1. Experimental Environment. In this study, experiments were carried out under the configuration of artificial intelligence computer, and the model proposed was trained and tested using NVIDIA Tesla V100 GPU. In the experiment, the MTCNN model was used to conduct face alignment and face region cropping for all images in the dataset, and then, they were adjusted to 224×224 size, mainly using ResNet18 as the baseline experiment. Adam was used to optimize the model, and the initial learning rate was set at 0.001. The latest face rough detection algorithm and fine detection algorithm, as well as the hybrid face detection algorithm of this article, were selected to detect 90 seconds of real classroom teaching video from the detection speed and accuracy analysis. The number, accuracy, and speed of faces detected by each algorithm in the detection process of 15 seconds, 30 seconds, 45 seconds, 60 seconds, 75 seconds, and 90 seconds are selected for analysis and comparison. A represents the model of this study, B represents the TRADITIONAL FINE detection model of CNN, and C represents the traditional AdaBoost coarse detection model.

4.2. Analysis of Experimental Results. From Figure 3, as time goes by, the total number of faces detected by all models does not change linearly, and the maximum number of faces detected by all models decreases first and then increases slowly. Because we randomly selected the time period, it proved the objectivity of the experiment from the side, and this phenomenon was consistent with real life. At the beginning of the class, the students sat upright and, as the class went on, some students began to pay attention and accompanied by various small movements, which led to the fluctuation of face detection. The total number of faces detected by the hybrid model proposed in this study is higher than that detected by the coarse and fine models in all time periods, which proves that our model is still reliable with the change of students' posture and the passage of time. From the results, the accuracy of the coarse detection model is the lowest, and the precision detection is in the middle. The model in this article absorbs the advantages of the two models and gets better results.

From Figure 4, the overall change of the accuracy detection result is inconsistent with the experimental result of total number of faces, among which coarse detection and fine detection have some similarity. The change of the algorithm in this study is small as time goes by, and it is proved from the side that the detection time length randomly selected will not affect the mixed model too much. From the perspective of the first 60 seconds, the detection accuracy of the model in this study decreases slightly with time, whereas that of the other two models decrease linearly. The only difference is that coarse detection drops to the lowest point and begins to rise linearly after 45 seconds, whereas fine detection moves backward by about 15 seconds compared with coarse detection. From the point of view of the minimum detection accuracy, the minimum detection accuracy of the model in this study is close to the maximum accuracy

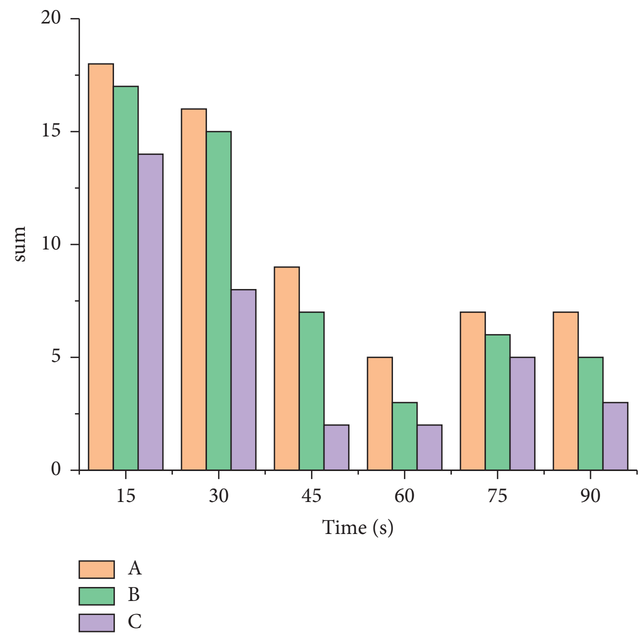

Figure 3: Face detection number result comparison.

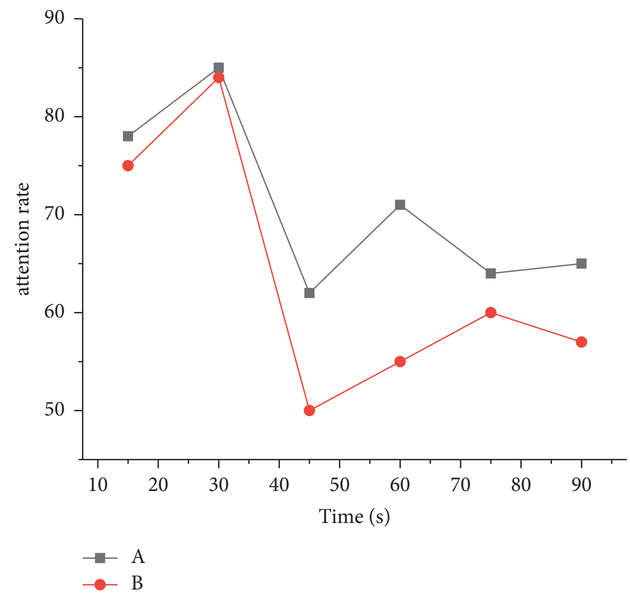

Figure 5: Experimental results of the students' attention rate in the class.

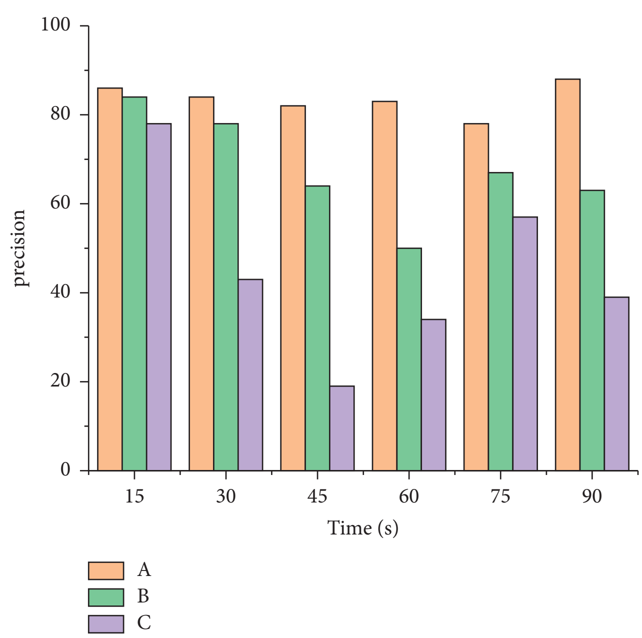

Figure 4: Comparison of face detection accuracy results.

of rough detection, which is about 78%, and 25% higher than the minimum accuracy of fine detection. By comparing the above experimental results from multiple angles, the model in this study not only achieves a good total number of faces but also has relatively stable detection accuracy.

As can be seen from Figure 5, in the first 1/3 time points randomly selected, the students' attention rate is the highest corresponding to assassin students' concentration. After that, the attention rate of students began to decline and began to fluctuate slightly after a certain time. It is worth noting that both algorithms get the lowest attention rate at 45 seconds; that is, when the class time is halfway through, students' attention is the lowest. In actual classroom teaching, it is not recommended that the teacher explain the most important content during this period, and the main content should be concentrated in the first half of the class. As the class size generally does not exceed 50, the number of individual objects is small for the budget algorithm, so the highest detection rate of the two algorithms is similar. However, from the average result, the model in this study is about 10% higher than the fine detection model. More importantly, the method of judging the students' attention rate by the students' position will reduce the accuracy due to the different directions of students' eyes. However, students' vision can be adapted to more scenes, such as large conference rooms and studio rooms, based on the difference between target students' and surrounding students' vision, without extra manual marking of students' locations. The hybrid model integrates the advantages of various algorithms and can be used in various situations.

Facial expression is the most direct reflection of students' psychological emotions. In order to better test the effectiveness of the model, we trained the model on a large number of public facial expression datasets and put 3000 test pictures from FER20L3 into the trained facial expression recognition model for testing. The model got seven expressions, but anger and fear were rarely seen in the classroom. According to the psychological dichotomies (positive and negative aspects), this paper divides seven kinds of expressions into two categories. The latitude of negative emotion represents the emotional experience of the individual showing negative or angry emotion, while the dimension of positive emotion reflects the individual showing positive emotion. So we classify happiness, surprise, and neutrality as positive emotions, anger, disgust, fear, and sadness as negative emotions.

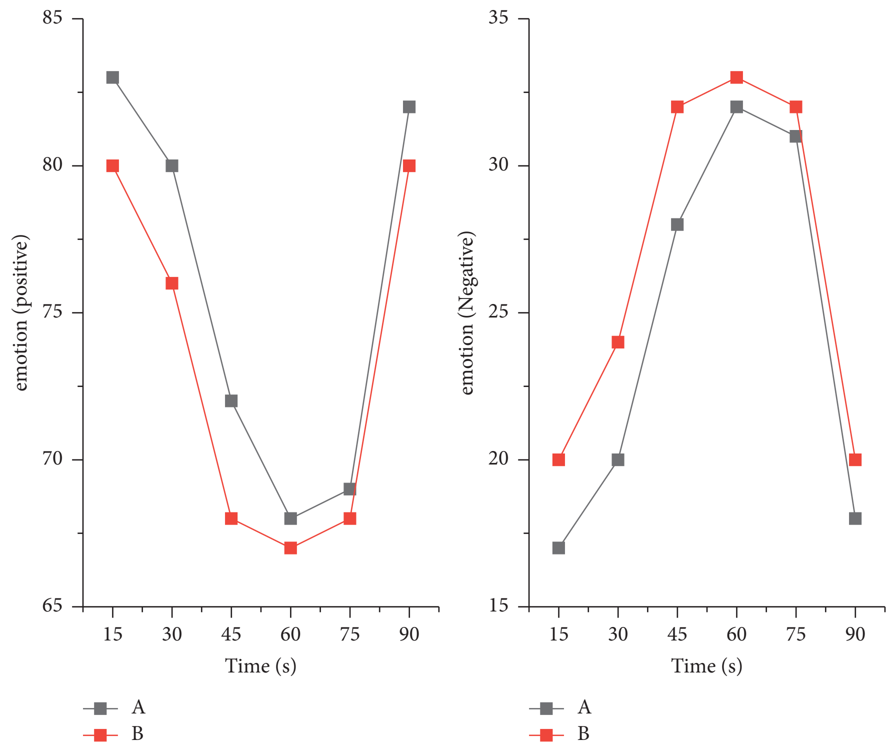

Figure 6: Student facial expression recognition results.

The experimental results are shown in Figure 6. By comparing the experimental results of algorithm and manual marking, it can be concluded that the accuracy rate of facial expression recognition model for students in natural class is 90%. More importantly, students' positive emotions continued to decline in the first 2/3 of the period from the beginning of class, and their negative emotions reached the highest level in the period from 1/2 to 5/6. The experimental results and analysis are in line with the actual situation, so school administrators should timely adjust the distribution of the classroom content according to these results, so as to improve the learning efficiency of students in a limited time.

5. Conclusions

In education, the classroom has always been the most important occasion for students and teachers to study and communicate, and the behavior state of students in the classroom has also attracted academic attention and research. This study collects students' classroom behavior state based on video images and analyzes the evaluation methods and shortcomings of classroom teaching at present. The basic methods of classroom behavior state are reviewed, face detection algorithm and expression recognition algorithm are introduced, and the implementation details are described. Based on the traditional algorithm, the hybrid face detection algorithm is improved, and a face expression recognition model based on visual t is established for students, and the feasibility and accuracy of the system are verified in the public dataset. It proves that the intelligent analysis of students' classroom videos will help teachers and other school administrators to make teaching scientific and improve teaching quality. Along with the continuous development of intelligent devices such as sensors, the physiological signal and other modal information will be added to the more video intelligent analysis of college students, through a variety of modal signal complement each other, will all aspects of the analysis of students' comprehensive emotional state, to help teachers and school administrators scientific teaching, further improve the quality of teaching.

Data Availability

The dataset used in this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declared that they have no conflicts of interest regarding this work.

References

[1] C. Zhou, "Teacher: organizer and manager of student learning -- discussion on the role orientation of teachers in the new era," *Education*, vol. 000, no. 6, p. 1, 2018.

[2] W. Qiao, "Evaluation and analysis of teachers' online guidance behavior based on ahp and neural network," *Journal of Tianjin TV University*, vol. 3, no. 4, pp. 23–27, 2007.

[3] W. Gao, *Construction of Evaluation Standard of Classroom Teaching Behavior Oriented to Promoting Students' Active Learning -- Research Based on Delphi Survey Method*, Educational Research and Experiment, 2013.

[4] X. Xu, "Research on illegal problems and countermeasures of teacher's," *Classroom Teaching*, vol. 4, pp. 7–9, 2021.
[5] J. Jiao, "Camera and face recognition technology in classroom," *China Information Technology Education*, vol. 19, 2019.
[6] H. Xu and W. Liu, "Design and practice exploration of student classroom supervision system based on "internet +" thinking," *Communication and Copyright*, no. 11, pp. 145–147, 2017.
[7] S. Tang, "Research on application of improved neural algorithm in classroom face recognition situation analysis," *Information & Systems Engineering*, vol. 311, no. 11, pp. 124-125, 2019.
[8] N. A. N. Zhou, *Research on Indoor/outdoor Scene Recognition Based on Wearable Device*, Shanghai Jiaotong University, Shanghai, China, 2017.
[9] G. Qian and W. Yu, "Application status of intelligent wearable devices in senile patients with dementia," *Chinese Journal of Nursing*, vol. 25, no. 14, p. 4, 2018.
[10] Z. Wang, *Research and Development of Student Class Status System Based on Video Analysis*, Xinjiang University, Ürümqi, China, 2019.
[11] L. Jia, Z. Zhang, and X. Zhao, "Classroom student status analysis based on ARTIFICIAL Intelligence video processing," *Modern Educational Technology*, vol. 29, no. 12, p. 7, 2019.
[12] A. Sinha, R. Gavas, D. Chatterjee, R. Das, and A. Sinharay, "Dynamic assessment of learner's mental state for an improved learning experience," in *Proceedings of the Frontiers in Education Conference*, El Paso, TX, USA, October 2015.
[13] K. Fujisawa and K. Aihara, *Estimation of Interest from Physical Actions Captured by Familiar User Device*, Springer, London, UK, 2011.
[14] N. Nourbakhsh, Y. Wang, F. Chen, and R. A. Calvo, "Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks," in *Proceedings of the Australian Computer-human Interaction Conference*, pp. 420–423, Melbourne, Australia, November 2012.
[15] Z. Zhan, "Emotional and cognitive Recognition model of distance Learners based on intelligent Agent–Coupling between Eye movement tracking and facial expression recognition technology," *Modern Distance Education Research*, vol. 6, no. 5, 2013.
[16] Z. Zhu, S. Ober, and R. Jafari, "Modeling and detecting student attention and interest level using wearable computers," in *Proceedings of the IEEE International Conference on Wearable & Implantable Body Sensor Networks*, IEEE, Eindhoven, Netherlands, May 2017.
[17] L. Shen, "Characteristics of wearable devices and NXP solutions," *Electronics World*, vol. 28, no. 5, p. 1, 2021.
[18] Y. Cheng, *Research on Classroom Teaching Behavior Analysis Method Based on Video*, Central China Normal University, Wuhan, China, 2015.
[19] G. R. Bradski and A. Kaehler, *Learning OpenCV - Computer Vision with the OpenCV Library: Software that sees*, DBLP, Germany, 2008.
[20] L. Hou, Y. Wang, and S. Zhang, "The application of face detection technology in teaching evaluation," *Electronic World*, no. 24, pp. 37-38, 2016.
[21] M. Rahman, M. Debnath, S. Sharmin, L. Alam, S. Arefin, and M. Hoque, "Designing an empirical framework to measure the level of interest of Human," in *Proceedings of the International Conference on Electrical Information & Communication Technology*, IEEE, Khulna, Bangladesh, December 2016.
[22] A. Mehrabian, "Communication without words," *Journal of University of East London*, vol. 2, no. 4, pp. 53–55, 1968.
[23] F. Mantang, Q. Ma, and R. Wang, "Research on intelligent network teaching system based on facial expression recognition," *Computer Technology and Development*, vol. 17, no. 6, pp. 193–196, 2011.
[24] M. Cheng, M. Lin, and Z. Wang, "Research on intelligent teaching system based on expression recognition and sight tracking," *Distance Education in China*, vol. 7, no. 5, p. 6, 2013.
[25] B. Sun, Y. Liu, and J. Chen, "Emotional analysis based on facial expression in intelligent learning environment," *Modern Distance Education Research*, vol. 5, no. 2, pp. 96–103, 2015.
[26] B. Jiang, W. Li, and Z. Li, "Automatic recognition of learning confusion based on facial expression," *Open Education Research*, vol. 11, no. 4, pp. 101–108, 2018.
[27] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: automatic recognition of student engagementfrom facial expressions," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 86–98, 2014.
[28] L. Han, L. I. Yang, and Z. Zhou, "Research on modern distance education," vol. 2, no. 4, pp. 97–103, 2017, in Chinese.
[29] L. Chen, Z. Luo, and R. Xu, "Intelligent analysis tn of students' learning interest in classroom teaching environment," *Electronic Education Research*, vol. 39, no. 8, pp. 93–98, 2018.
[30] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, *Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments*, USA, 2008.
[31] V. Jain and E. Learned-Miller, *A Benchmark for Face Detection in Unconstrained Settings*, USA, 2010.
[32] M. Kostinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, IEEE, Barcelona, Spain, November 2012.
[33] T. Y. Lin, M. Maire, and S. Belongie, "Microsoft COCO: common objects in context," in *Proceedings of the European Conference on Computer Vision*, Springer International Publishing, Zurich, Switzerland, September 2014.
[34] C. P. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in *Proceedings of the Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, Bombay, India, January 1998.
[35] Q. Zhu, M. C. Yeh, K. T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *Proceedings of the IEEE Computer Society Conference on Computer Vision & Pattern Recognition*, IEEE, New York, NY, USA, June 2006.