Neural Computing and Applications (2025) 37:13237–13249
https://doi.org/10.1007/s00521-024-10529-z

S.I.: NEURAL COMPUTING AND INTELLIGENT EDUCATION APPLICATIONS

Research on the impact of pointing gestures based on computer vision technology on classroom concentration

Jianyang Shi[1] · Zhangze Chen[1] · Jia Zhu[1] · Jian Zhou[1] · Qing Wang[2] · Xiaodong Ma[1]

Received: 8 October 2023 / Accepted: 3 October 2024 / Published online: 3 December 2024
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2024

Abstract
Classroom concentration is an essential manifestation of learners' engagement in the classroom, and it is a critical factor in adjusting learning states and optimizing teaching processes. A thorough exploration of the factors that affect classroom concentration and their effects is of great significance for enhancing it. This study proposes a classroom concentration recognition model by integrating the two dimensions of emotional and behavioral concentration using computer vision technology. The model's effectiveness in recognizing classroom concentration is verified through a comparative experiment with EEG equipment. Based on this model, this study further investigates the impact of teachers' pointing gestures on learners' classroom concentration. The results of ANOVA show that when teachers use pointing gestures, they can improve learners' concentration in class in a short time, but this positive effect has boundary effects. Our research results can help improve teachers' teaching behaviors and enhance learners' learning effects to some extent.

Keywords Pointing gestures · Classroom concentration · Computer vision · Teaching behavior

1 Introduction

In the teaching process, concentration is a prerequisite for effective learning and an essential guarantee for achieving good learning results. Classroom concentration refers to an individual's ability to selectively process pertinent information while disregarding external distractions amidst many stimuli [1]. When fully concentrated, learners tend to demonstrate a stationary state, directing their attention toward the learning stimuli and persisting in learning activities to accomplish the desired learning objectives [2]. There is strong evidence that learners have different learning performances in various states of concentration. In particular, when in a high-concentration state, students' motivation to learn is significantly enhanced, thereby facilitating efficient processing and retention of information [3, 4]. Therefore, accurate assessment and effectively controlling students' classroom concentration are critical.

1.1 Assessment of learning concentration

Accurately evaluating learning concentration has remained a pivotal subject in the field of education for a considerable duration. There are two traditional methods of concentration assessment: one is by observing learners' external behavior (such as body language, facial expression, etc.) to judge their concentration [5] and the other is by learners' self-report of concentration state [6], both of which have certain subjectivity and lag. With technological advancements in data collection, many scholars have explored

✉ Jianyang Shi
shijianyang@zjnu.edu.cn

Zhangze Chen
zjnuczz@zjnu.edu.cn

Jia Zhu
jiazhu@zjnu.edu.cn

Jian Zhou
zhoujian@zjnu.edu.cn

Qing Wang
wq2481@zjnu.edu.cn

Xiaodong Ma
felicity@zjnu.edu.cn

[1] College of Education, Zhejiang Normal University, No. 688 Yingbin Avenue, Jinhua 321004, Zhejiang, China

[2] School of Computer Science and Technology, Zhejiang Normal University, No. 688 Yingbin Avenue, Jinhua 321004, Zhejiang, China

automatic analysis of concentration based on learner data. A prevalent approach involves the utilization of cameras to gather nonintrusive explicit behavior data, followed by extracting pertinent features and identifying concentration levels through diverse machine learning methodologies [7, 8]. For instance, Hu et al. [9] proposed a bimodal model that integrates features from multiple modalities to evaluate concentration levels. Although this method allows for real-time and automatic assessment of learners' concentration, its effectiveness requires further evidence. In addition, eye movement data, such as fixation points, duration and count, saccade trajectory, and other factors, can also be used to evaluate learners' attention span, the difficulty of learning materials, and attention levels [10, 11]. D'Mello et al. [12], for example, focused on online reading scenarios and used global and local features, such as overall fixation rate, fixation duration and count, saccade length, reading time for words of varying lengths, skipped word count, and duration of initial fixation, to assess learners' concentration. Additionally, analyzing the frequency spectrum of electroencephalogram (EEG) signals can provide insights into learners' concentration levels [13, 14]. Although methods based on physiological data, such as eye-tracking and wearable EEG devices, can accurately assess learners' concentration levels, they often entail significant investment costs and may impede students' learning experiences. Hence, this study aims to combine computer vision technology and physiological data synergistically, facilitating the automated evaluation of learners' concentration levels without requiring invasive wearable EEG devices.

Accurate classroom concentration assessment can provide teachers with high-quality feedback, yet our primary objective is to optimize students' classroom concentration to enhance their learning outcomes. Therefore, it is crucial to identify the factors that influence classroom concentration. Many studies have shown that teachers' nonverbal behavior can promote learners' learning. Using gestures is one of the important features of teachers' nonverbal behavior in teaching. It can affect learners' cognitive processes, attention allocation, semantic integration, and social emotions, thus affecting their learning effect [15, 16]. Therefore, the impact of teachers' gestures on learners has received much attention.

1.2 The influence of gestures on learners

Gestures refer to specific actions that occur when a person uses their arms. In classroom teaching, teachers often use three kinds of gestures [17]: (a) pointing gestures, which indicate objects or locations, typically with an extended finger or hand; (b) descriptive gestures, which describe semantic content through the shape or movement trajectory of hands aspects of the shape to literally or metaphorically evoke a mental image of the shape in the listener's mind; and (c) beat gestures, which are simple, up-and-down rhythmic movements that do not depict semantic content, but instead align with the prosody or discourse structure of speech. Numerous studies have demonstrated the positive impact of teacher gestures on learners' learning outcomes [18]. More specifically, gestures serve as a valuable supplement to language information, enhancing learners' comprehension and memory abilities. In addition, acting as visual cues, gestures effectively facilitate learners' understanding of intricate concepts and information, concurrently fostering heightened concentration and engagement levels. For example, Pi et al. [19], employed eye movement tracking technology to capture subjects' eye movements while learning tasks. By comparing eye movements and learning performance under different gesture teaching conditions, they discovered that both pointing and descriptive gestures direct learners' attention toward the teacher and the learning materials displayed on the screen, but pointing gestures are better at encouraging learners to switch their attention between the teacher and learning materials, while descriptive gestures are better at promoting learners to allocate their attention within the learning materials, as mentioned above. Research shows that teachers can use different types of gestures to adjust students' attention allocation during teaching to promote their learning. EEG studies focusing on gestures have also revealed a significant correlation between learners' concentration levels and their observation of distinct types of gestures [20]. Notably, when students observed their teacher's beat or descriptive gestures, the amplitudes of α and β oscillations were higher than when they watched pointing gestures. α and β waves are linked to visuospatial attention allocation and activation of the sensorimotor cortex [21], suggesting that learners require the engagement of the sensorimotor cortex when processing gestures and performing visual spatial attention allocation cognitive tasks. Extensive research has consistently demonstrated the positive impact of teacher gestures on learners' learning outcomes, although the underlying mechanisms behind the favorable effects of various gesture types may exhibit variation. However, some scholars suggest that there may be a boundary effect on the positive impact of teacher gestures on learners, especially when gestures are highly redundant with speech or do not align with the teaching theme [22, 23]. In such cases, using gestures may not be beneficial for learning.

Based on the above research, we can infer a certain correlation between teachers' gestures and learners' levels of concentration. However, the specific relationship still needs further validation. In order to break through the teaching black box and reveal the intrinsic connection between teaching gestures and classroom concentration,

thereby assisting teachers in regulating learners' classroom concentration to promote learning outcomes, we conducted in-depth research on the correlation between teachers' pointing gestures and students' classroom concentration using computer vision technology. During the research process, the challenges faced include: first, how to evaluate the effect of the classroom concentration identification model. The current classroom concentration identification method based on computer vision technology obtains the learner's explicit behavior data based on the visual model. Then, it uses the corresponding evaluation indicators and calculation methods to arrive at the final classroom concentration level. In the above methods, the accuracy of the visual model does not represent the quality of the classroom concentration recognition model. Second, in actual classroom teaching, the nonverbal behaviors of teachers are complex and diverse, which makes it very difficult to accurately identify between teachers' pointing gestures and non-pointing gestures. Finally, based on the above two works, we explored the impact of teachers' pointing gestures on learners' classroom concentration in actual classroom teaching. The following is a summary of the contributions of our work:

1. This study proposes a classroom concentration recognition model. It adopts a new model evaluation method to evaluate the performance of the classroom concentration recognition model from a physiological level.
2. This study reveals the impact of teachers' pointing gestures on learners' classroom concentration, providing a reference for teachers to improve their teaching behaviors.

2 Methodology

This section presents the classroom concentration recognition method proposed in this study and the pointing gesture recognition method based on the ResNet50 neural network [24].

2.1 Classroom concentration recognition

According to the three-dimensional representation model [25], classroom concentration consists of three dimensions: emotional, cognitive, and behavioral. These three dimensions are organically combined but relatively independent. Emotional concentration refers to students' emotional experiences during the learning process; cognitive concentration relates to students' cognitive strategies and self-monitoring; behavioral concentration mainly refers to students' active learning and interactive learning behaviors. In these three dimensions, cognitive concentration refers to internal processes, while only emotional and behavioral concentration is reflected in visible clues. Therefore, this research mainly focuses on emotional concentration and behavioral concentration. In students' external behaviors, facial expressions are the most direct way to reflect their emotional concentration status, while head posture can reflect their behavioral concentration status. Therefore, as shown in Fig. 1, the classroom concentration recognition model mainly includes four steps: face detection, facial expression recognition, head posture detection, and classroom concentration calculation. The student's video data first identifies the learner's face information through the face detection module and obtains the student's face image by cropping. Subsequently, all the obtained face images are input to the head pose detection and facial expression recognition modules to get each learner's facial expression category and head rotation angle. Then, based on the corresponding evaluation indicators, the classroom concentration value is calculated. We will cover these steps in detail in the following sections.

2.1.1 Face detection

Face detection is crucial for subsequent tasks, such as expression recognition, head pose detection, etc. However, face detection has many challenges in practical use. First, a face can appear in various poses, including rotation, tilt, and pitch. Accurately locating the critical feature points under these postural changes is challenging, especially in the non-frontal perspective. Second, the appearance of a face varies significantly between different ages and genders. Face detection systems must adapt to these differences to perform accurate localization in various situations. Finally, a face's expression can vary, from smiling to frowning. Expression changes may lead to changes in the location of the key feature points, which poses a challenge for accurate face localization. In response to the above problems, this study uses RetinaFace [26] to perform face detection.

RetinaFace has made improvements to address the above issues. As shown in Fig. 2, the model consists of three modules: feature pyramid network, context module, and cascade multitask loss. The feature pyramid network uses ResNet as the backbone to extract features from C2 to C5 layers. It generates P2–P6 layer feature pyramids using top-down and horizontal connections to handle faces of different scales. Context modules have been added at each layer of the feature pyramid to increase receptive fields and enhance context modeling capabilities. A 1×1 convolutional layer is used to calculate the multitask loss at each level of the feature pyramid, including face classification, face box regression, five 2D facial landmark regression, and 1k 3D vertex regression. To further improve the

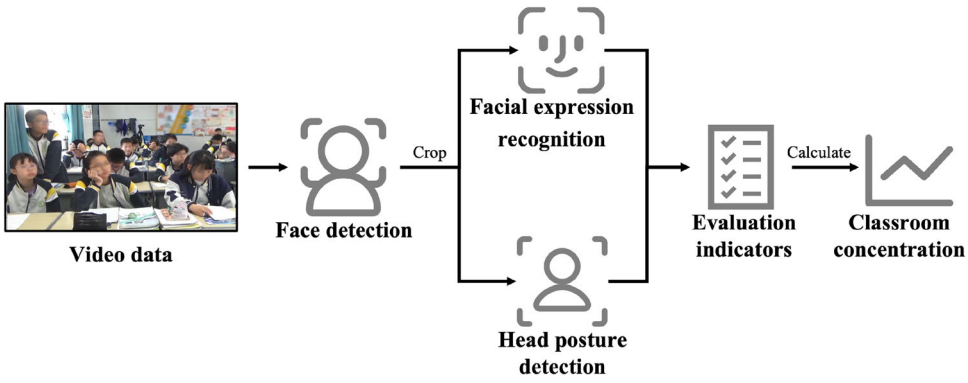

Fig. 1 The overall process of classroom concentration recognition model

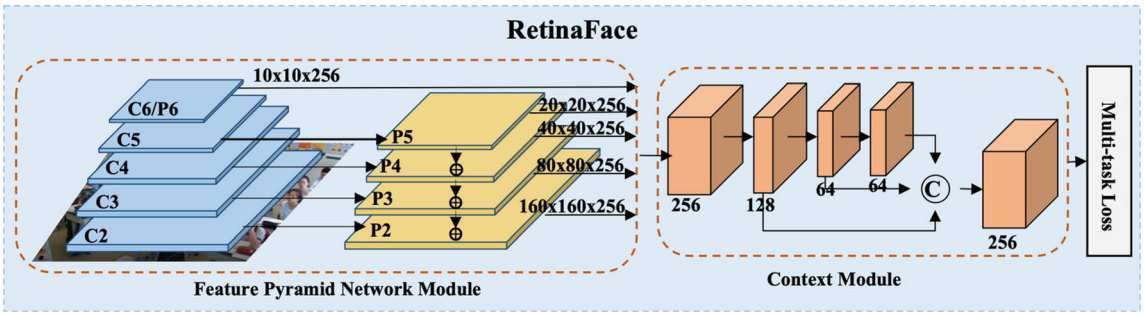

Fig. 2 RetinaFace model structure diagram

accuracy of face localization, a cascade regression method is used, where the face box is predicted based on the normal anchor points first, and then more accurate face boxes are predicted based on the regression of anchor points. The calculation method of the loss function is shown in Eq. (1).

$$\mathcal{L} = \mathcal{L}_{\text{cls}}\left(p_i, p_i^*\right) + \lambda_1 p_i^* \mathcal{L}_{\text{box}}\left(t_i, t_i^*\right) + \lambda_2 p_i^* \mathcal{L}_{\text{pts}}\left(l_i, l_i^*\right) + \lambda_3 p_i^* \mathcal{L}_{\text{pixel}} \tag{1}$$

Among them, $\mathcal{L}_{\text{cls}}\left(p_i, p_i^*\right)$ is the face classification loss. p_i is the predicted probability that the i-th anchor is a face; p_i^* takes the value 1 when the anchor is positive and 0 when the anchor is negative. $\mathcal{L}_{\text{box}}\left(t_i, t_i^*\right)$ is the face box regression loss. $t_i = \left\{t_x, t_y, t_w, t_h\right\}_i, t_i^* = \left\{t_x^*, t_y^*, t_w^*, t_h^*\right\}_i$, t_i, and t_i^* are the coordinates of the predicted face frame and the actual face box, respectively. $\mathcal{L}_{\text{pts}}\left(l_i, l_i^*\right)$ is the loss of facial key points, $l_i = \left\{l_{x_1}, l_{y_1}, \ldots, l_{x_5}, l_{y_5}\right\}_i$, $l_i^* = \left\{l_{x_1}^*, l_{y_1}^*, \ldots, l_{x_5}^*, l_{y_5}^*\right\}_i$, l_i and l_i^* represent the predicted coordinates of the five facial key points and the five actual coordinate points of the positive anchor, respectively. $\mathcal{L}_{\text{pixel}}$ is the loss of dense regression. The loss adjustment parameters λ_1, λ_2, and λ_3 are set to 0.25, 0.1, and 0.01, respectively, which means that in the supervision signal, the importance of bounding boxes and key point positioning is increased, which also reflects the relatively low importance of loss in the dense task.

2.1.2 Expression recognition

For facial expression recognition, this study uses the ResMaskingNet facial expression recognition model [27] and trains it on the FER2013 public dataset [28]. The model proposes a new masking method to improve the accuracy of facial expression recognition tasks. It also uses a segmentation network to refine feature maps, enabling the model to focus on relevant details and make correct decisions. In addition, the model combines deep residual networks and U-Net [29] structures to generate residual masking networks.

As shown in Fig. 3, the ResMaskingNet first generates coarse-grained features from the feature map through the residual layer. Then, it calculates refined features from the coarse-grained feature map through the mask block. Finally, it combines the coarse-grained and fine-grained features to obtain the final feature map. Specifically, given an input feature map $F \in R^{C \times W \times H}$, firstly, F will pass through the residual layer R to generate a coarse feature map $F_{\text{R}} = R(F)$, $F_{\text{R}} \in R^{C' \times W' \times H'}$. Then, the mask module calculates an activation map F_M with values ranging from 0 to 1. Finally, Eq. (2) outputs the refined residual mask block feature map.

$$F_N = F_{\text{R}} + F_{\text{R}} \otimes M(F_{\text{R}}) \tag{2}$$

F_{R} is the coarse-grained feature after transformation through the residual layer, M refers to using mask blocks

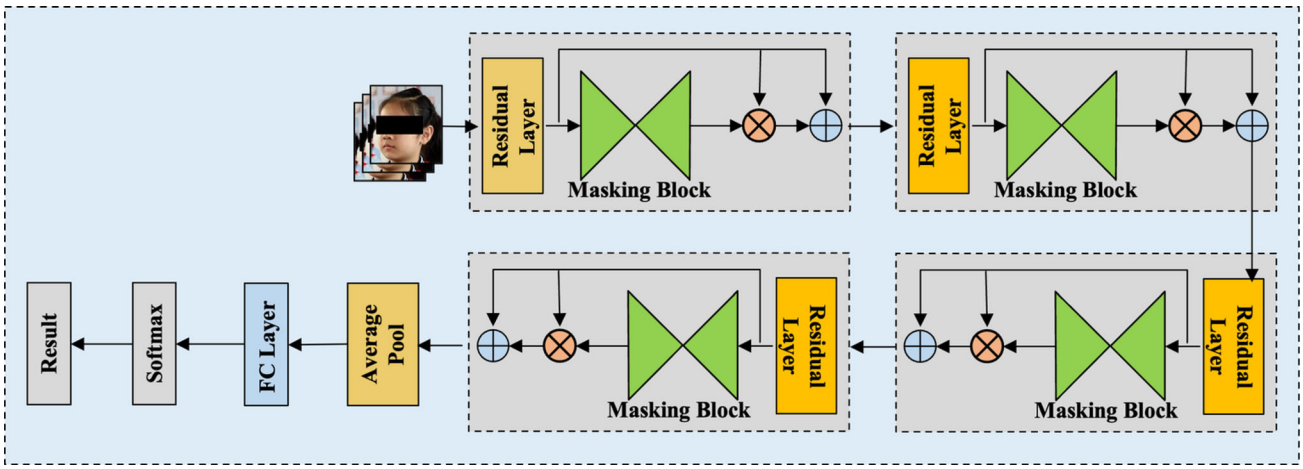

Fig. 3 ResMaskingNet model structure diagram

for calculation, and $\otimes$ means calculating by element-wise multiplication.

2.1.3 Head posture detection

This study used the HopeNet head pose detection model [30] for the head pose detection task. There are many problems in the actual use of head posture detection. Firstly, the data required for the training set is relatively large, so it is necessary to fit the input data distribution during training as much as possible. Secondly, the data distribution is not uniform. Usually, the most data set is the face facing the camera, while the amount of data for head up, side face, and head down is relatively small. This situation makes the model training easily lead to overfitting, leading to low recognition accuracy in natural scenes. In addition, using one loss function to target three different labels leads to poor model convergence. HopeNet has made improvements to the above problems. First, HopeNet proposes discretizing the regression space and introducing a classification layer to learn about data distribution to solve the problem of complex model convergence. In addition, HopeNet adjusts the effect of different distribution data on network models by setting three various losses to improve head pose detection accuracy.

As shown in Fig. 4, the HopeNet model first extracts image features using the backbone network and then puts these features into a fully connected layer to obtain a feature of 66 dimensions.

In this study, we adopt ResNet50 as the backbone network. Each backbone network can be extended to predict angles with three fully connected layers that share the network's previous convolutional layers. Then, the model divides the 3D angles according to the preset angle intervals to obtain the label of each angle and calculates the cross-entropy error during the division process. For example, if the interval is set to 3 degrees, the pitch, yaw, and roll angles will be divided into 66 intervals, resulting in labels for each angle. For three angles (yaw, pitch, roll), HopeNet employs three separate loss functions, each loss being a combination of binary pose classification and regression. The loss function calculates the loss for each angle, consisting of cross-entropy loss and mean squared error. The weight coefficient is added before the mean squared error, and different weights can be chosen for different network models to adjust the loss function. The specific loss function is shown in the Eq. (3).

$$\mathcal{L} = H(y, \hat{y}) + \alpha \cdot \mathrm{MSE}(y, \hat{y}) \tag{3}$$

$\mathcal{L}$ is the loss function, H is the cross-entropy, and MSE is the mean square error. Cross entropy H is used to control the angle base point of the classification regression angle, and mean square error MSE is used to control the expected adjustment based on the angle base point. y represents the training results, and $\hat{y}$ represents the annotated results. Among them, α is the loss weight factor in the training model, which is set to a value of 1 in this study.

2.1.4 Classroom concentration calculation

Classroom concentration includes the concentration of students' thinking and emotions in the classroom, as well as their positive and negative behaviors. Therefore, when calculating classroom concentration, not only students' emotional concentration but also their behavioral concentration should be considered. Emotional concentration refers to students' emotional experiences during learning, including positive and negative emotions. To obtain the students' classroom emotional concentration state through expression, this article defines a specific description of classroom expressions and their corresponding evaluation scores based on relevant research [31], as shown in

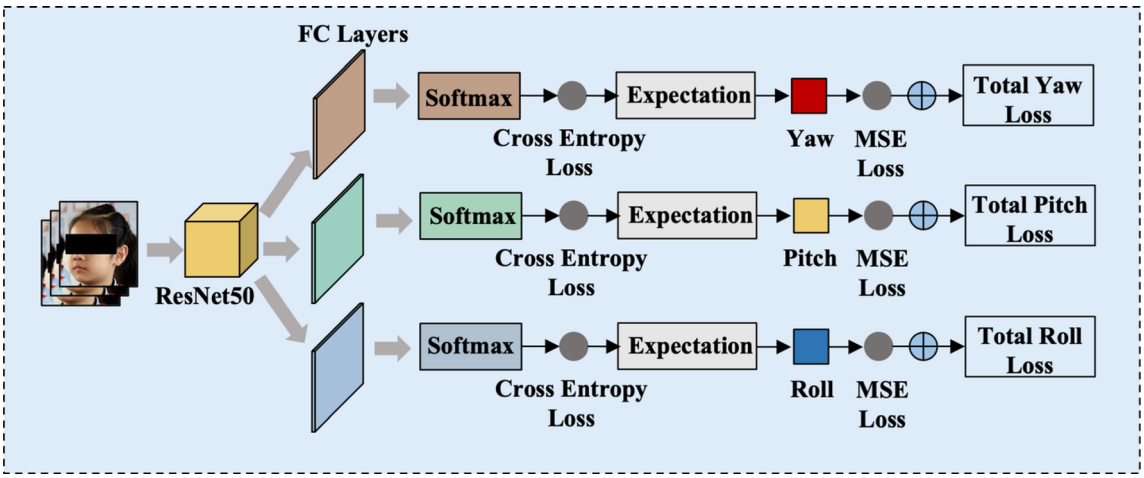

Fig. 4 HopeNet model structure diagram

Table 1 Classroom expression and their corresponding weights

Expression category	Student expression state description	Score
Disgust	Eyebrows lowered, corners of mouth down, pouting	−3
Angry	Eyes wide open, pupils smaller, eyebrows pressed down, nostrils dilated	−2
Sad	Facial muscles are tight, eyebrows and eyelids are raised, and pupils first become larger and then smaller	−1
Fear	Facial muscles are tight, eyebrows and eyelids are raised, and pupils first become larger and then smaller	0
Neutral	Facial muscles are relaxed, evenly distributed, without significant changes	1
Happy	Extend the lips outward and upward, tilt the corners of the mouth upwards, and lift the cheek muscles upwards	2
Surprise	Eyes wide open, pupils dilated, upper eyelids and eyebrows raised, mouth wide open	3

Table 1. For the calculation of emotional concentration, the specific calculation formula is shown below:

$$f_s = \begin{cases} -3^*p_1 - 2^*p_2 - 1^*p_3 + 0^*p_4 + 1^*p_5 + 2^*p_6 + 3^*p_7 \\ 0 \end{cases} \tag{4}$$

where p_i $(1 \le i \le 7,\ i \in Z)$ represents the probability that a face is recognized as a different expression. Suppose a face is detected in the face detection stage. In that case, the expression probability p_i of the face is multiplied by the corresponding expression evaluation score and then added to obtain its emotional concentration score. If no face is detected during the face detection stage, the emotional concentration score is 0. To further facilitate subsequent calculations and highlight the intuitive meaning, the concentration scores based on facial expressions are normalized within the [0–1] range and then expanded to [0–100], as illustrated in Eq. (5).

$$f_s^* = \frac{f_s' - \min(f_s)}{\max(f_s) - \min(f_s)} * 100, f_s' \in f_s \tag{5}$$

f_s^* is the emotional concentration score after the original emotional concentration score f_s is normalized and extended to [0, 100].

Behavioral concentration mainly refers to students' positive and negative learning behaviors. By observing students' head posture, we can promptly and effectively discover whether students focus on teaching content. To obtain students' behavioral concentration state through head posture, this paper refers to previous research [32] and defines the correspondence between head posture and

Table 2 Concentration level classification of head deflection angle

Pitch	Yaw			
	0°–20°	20°–40°	40°–60°	> 60°
0°–20°	v_1	v_2	v_3	v_4
20°–40°	v_2	v_2	v_3	v_4
40°–60°	v_3	v_3	v_3	v_4
> 60°	v_4	v_4	v_4	v_4

behavioral concentration state. As shown in Table 2, head rotation's pitch and yaw angles are divided into four levels: very serious v_1, serious v_2, not serious v_3, and very not serious v_4.

Based on each student's emotional concentration score and their level of head posture deviation, we can calculate the average classroom concentration at a specific moment. The specific calculation formula is as follows:

$$f_r = \frac{\sum_{n=1}^{N} f_s^* * \alpha}{N} \tag{6}$$

here N represents the number of faces detected, f_s^* represents the emotional concentration of a student, α represents the weight of the corresponding head posture, and v_1, v_2, v_3, and v_4 represent weight values corresponding to 1, 0.75, 0.5, and 0.25, respectively.

2.2 Teacher pointing gesture recognition

This section introduces a pointing gesture recognition method based on ResNet50. The method mainly includes two steps: teacher recognition and pointing gesture recognition. The specific process is shown in Fig. 5.

2.2.1 Teacher recognition

The premise of recognizing teachers' pointing gestures in classroom teaching is to identify the teacher objects in each frame of the classroom video. Through tracking and photographing two classes in a middle school in Zhejiang Province for a semester of math classes, we found that teachers spent most of their time near the podium when teaching-learning materials. This activity area is reflected in the video as a rectangular area surrounded by a podium and blackboard. Therefore, we assume that teachers are always located near the podium when using pointing gestures to teach. As shown in Fig. 5, we identify the teacher in the classroom video by leveraging the significant spatial boundary between the teacher and the students. Specifically, consider the color attribute of the electronic whiteboard. First, convert an RGB image into a grayscale image, and then set the threshold value to 150 to convert the grayscale image into a binary image, where contour detection is performed. Since the electronic whiteboard has the largest outline area, we can obtain the outline of the electronic whiteboard through contour detection. After determining the outline of the electronic whiteboard, we can obtain the upper left coordinate (X_lt, Y_lt) of the bounding rectangle of the outline and the width W_{rect} and height H_{rect} of the bounding rectangle. To obtain the

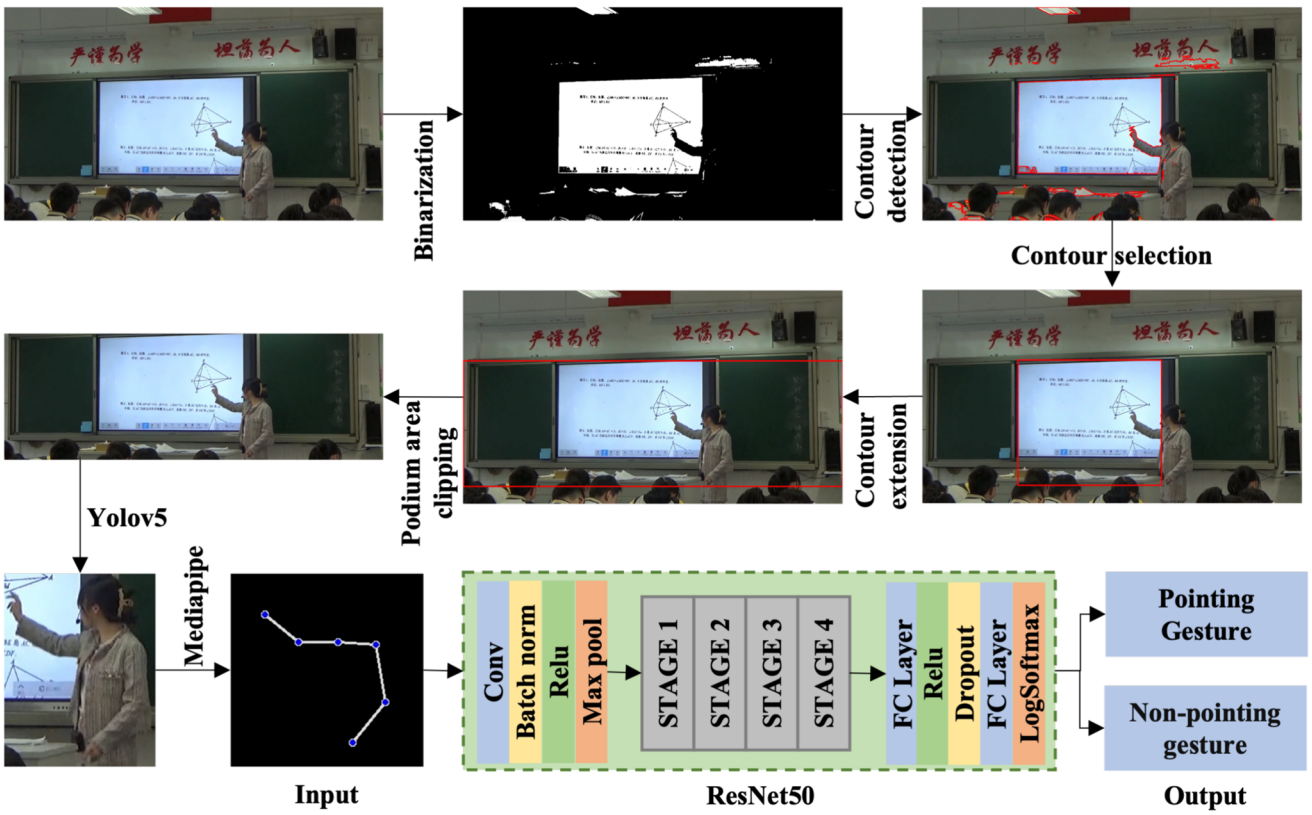

Fig. 5 The process of pointing gesture recognition

coordinates of the four vertices of the bounding rectangular area enclosed by the podium and the blackboard, we appropriately amplify the height of the external rectangle. The formula for calculating the coordinates of the four vertices of the rectangle in the teacher activity area is as follows:

$$\begin{aligned} X_1 &= 0 \\ X_2 &= W_{\text{pic}} \\ Y_1 &= Y_{lt} \\ Y_2 &= Y_1 + H_{\text{rect}} * 1.3 \end{aligned} \tag{7}$$

where W_{pic} is the width of the entire image. Within the rectangular area defined by the four coordinate points(X_1, X_2, Y_1, Y_2), the teacher object can be extracted through Yolov5 [33].

2.2.2 Pointing gesture recognition

Since hand posture depends on the position of the teacher's joints, we use the Mediapipe algorithm to detect human joints. Specifically, the sequence of teacher images captured in the teacher activity area is used as input to Mediapipe to obtain the connection points S_i of the characters in the picture. Mediapipe can detect 33 human body joints, but we only need the hand joints.

Therefore, $11 \leq i \leq 16,\ i \in Z$. For the recognition of pointing gestures, we used ResNet50. During the training process, we utilized the pretrained weights of the ResNet50 model on the ImageNet dataset. This approach not only improved the training speed and stability of the model but also enabled it to achieve good classification results even when the dataset is relatively small. And for the gesture classification task, we fine-tuned the ResNet50 model. As shown in Fig. 5, we replaced the last layer (fully connected layer) of the ResNet50 model with a new neural network layer that contains two fully connected layers, a ReLU activation function, a dropout layer, and a LogSoftmax layer to make ResNet50 suitable for binary classification tasks. The datasets used in the training process all come from real classroom videos and contain two labels: pointing gestures and non-pointing gestures. A total of 2183 RGB teacher images were included, of which 1310 were in the training set, and 873 were in the test set. After 100 rounds of training, the accuracy of the model on the training and testing sets was 0.8576 and 0.8394, respectively.

3 Experiment

In this study, we conducted two experiments. One experiment evaluated the effect of the classroom concentration recognition model, and the other explored the impact of pointing gestures on students' classroom concentration. In this section, we will detail the procedures and results of these two experiments.

3.1 Evaluation of the effectiveness of the classroom concentration recognition model

To evaluate the effectiveness of our proposed classroom concentration recognition model and ensure the accuracy of subsequent research, we conducted a comparative experiment to analyze the correlation between EEG devices and classroom concentration recognition modes in monitoring learners' classroom concentration fluctuations. The specific experimental process will be detailed in the following chapters.

3.1.1 Participants

Graduate students ($N = 5$) from a university in China participated in this study. They are between 23 and 24 years old ($M_age = 23.8$, $SD_age = 0.4$). All participants have normal vision or are corrected to normal. In addition, no one has a history of neurological diseases. As the subjects will study advanced mathematics, the recruited subjects are from nonmathematical majors. The participants participated in the experiment voluntarily. Before the experiment, all the participants signed an informed consent form and received experimental remuneration after the experiment.

3.1.2 Measurements

Experimental equipment, Sony HDRCX680 (recording video of the participant's class at a frame rate of 25 fps) and portable head-mounted EEG device (with built-in Think Gear AM chip [34]), used to collect EEG data of participants and measure their concentration. Participants were asked to wear the EEG device to watch a 20-minute video of an advanced mathematics lesson. To eliminate the influence of the invasiveness of the EEG equipment on the participants as much as possible, we conducted five experiments on the same group of participants. The teaching video in each experiment explained the "limit of the function," but the content was different.

3.1.3 Data collection and analysis

During the experiment, the classroom concentration of each participant per second can be obtained through EEG devices to calculate the average classroom concentration per second. After each video recording, we sampled the recorded classroom video second by second and obtained 1200 classroom pictures. These 1200 classroom pictures are then input into the classroom concentration recognition model, and the average classroom concentration per second is obtained based on model analysis.

After the five experiments, we collected five sets of data, each set of data containing 1200 classroom concentration values based on model recognition and 1200 classroom concentration values based on an EEG device. Then, SPSS 26 was used as the data analysis tool to conduct correlation analysis on these five sets of data. As shown in Table 3, the Pearson correlation coefficients of the five experiments were 0.413, 0.401, 0.574, 0.553, and 0.568, indicating that the classroom concentration values identified by the model were positively correlated with the classroom concentration values measured by EEG, and all of them had significant statistical effects ($p < 0.01$).

In addition, there are differences between the correlation coefficients of classroom concentration obtained by the two measurement methods in the first two experiments and the correlation coefficients of the last three experiments. The possible reason is that the invasive data collection of EEG devices makes learners' external behaviors and internal psychological resource input inconsistent. This leads to the significant differences between the concentration fluctuations monitored by EEG and those analyzed by the model in the first two experiments. However, with the increase in the number of experiments, the perception of the subjects to the portable EEG devices is gradually weakened; that is, the impact of the invasive data collection on the subjects is becoming less and less, and their external behaviors and internal psychological resource input are gradually consistent. The final correlation coefficient is stable at about 0.5.

In conclusion, the experiment results demonstrate that the classroom concentration recognition method proposed in this study can effectively monitor learners' concentration and provide teachers with high-quality feedback. However, the ultimate goal of this research is to explore the underlying factors affecting classroom concentration and provide guidance for improving teaching practices. We were inspired by current research on teaching behaviors and conducted an in-depth exploration of teachers' pointing gestures based on the classroom concentration recognition model.

3.2 The impact of pointing gestures on classroom concentration

In order to reveal the impact of teachers' pointing gestures on students' classroom concentration, we used the classroom concentration recognition model and pointing gesture recognition model proposed in this study to conduct in-depth research on the effect of pointing gestures on classroom concentration in real classrooms. Before the experiment, drawing upon previous research on classroom concentration and teachers' pointing gestures, we put forth the following hypotheses:

H1: Using pointing gestures in classroom teaching can improve students' concentration.

H2: There is a boundary effect in the use of pointing gestures in the classroom, i.e., the teacher's excessive use of pointing gestures in a short time is not conducive to improving learners' concentration.

3.2.1 Data collection and analysis

We recorded five eighth-grade mathematics classes in a middle school in Zhejiang, each lasting 40 min. Then, using our proposed classroom concentration and gesture recognition models, we calculated the average classroom

Table 3 Correlation analysis of two types of classroom concentration measurement results

Experiment number	Method	*M*	*SD*	*p*	Pearson correlation
1	Model	47.14	24.343	< 0.01	0.413
	EEG device	64.81	24.829		
2	Model	47.26	24.649	< 0.01	0.401
	EEG device	65.23	25.071		
3	Model	47.01	24.794	< 0.01	0.574
	EEG device	66.91	24.213		
4	Model	48.23	24.208	< 0.01	0.553
	EEG device	66.89	24.215		
5	Model	46.36	24.227	< 0.01	0.568
	EEG device	66.21	23.850		

concentration per minute and the frequency of the teacher's pointing gestures. Finally, we used ANOVA in SPSS 26 to perform variance analyses with the pointing gesture condition as a factor.

3.2.2 Teachers' use of pointing gestures in the classroom

Regarding teachers' use of pointing gestures in the classroom, the statistical results in Fig. 6 show that teachers used pointing gestures during most of the teaching process, with their time proportion exceeding 50%, and there is a significant difference in the use time between pointing and non-pointing gestures(MD = 9.38, $p < 0.001$). The above results demonstrate the importance of pointing gestures in teaching.

3.2.3 The effect of pointing gestures on classroom concentration

As shown in Fig. 7, the results of ANOVA show that learners' classroom concentration is significantly different ($M = 64.02$, SD = 3.92, $p < 0.001$) when teachers use pointing gestures and when they do not. Specifically, teachers' use of pointing gestures can significantly enhance learners' classroom concentration compared to the absence of pointing gestures. Therefore, hypothesis H1 is true. These findings underscore the importance of integrating pointing gestures into teaching strategies to improve learners' classroom concentration.

3.2.4 The effect of different levels of pointing gestures on classroom concentration

To further explore the influence of teachers' pointing gestures on learners' classroom concentration, we classified the frequency of teachers' pointing gestures within one minute into three levels: low (PG = 0), normal

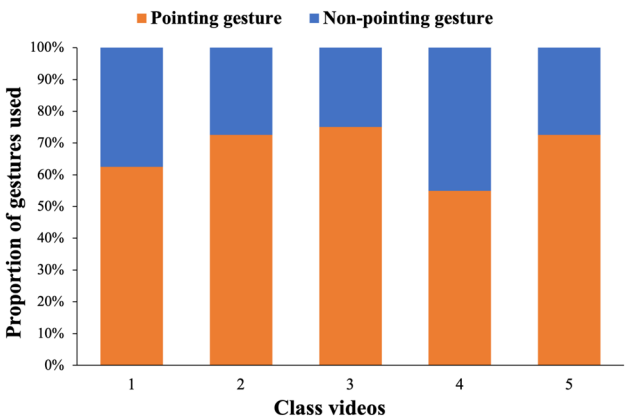

Fig. 6 The use of teacher pointing gestures in five classroom videos

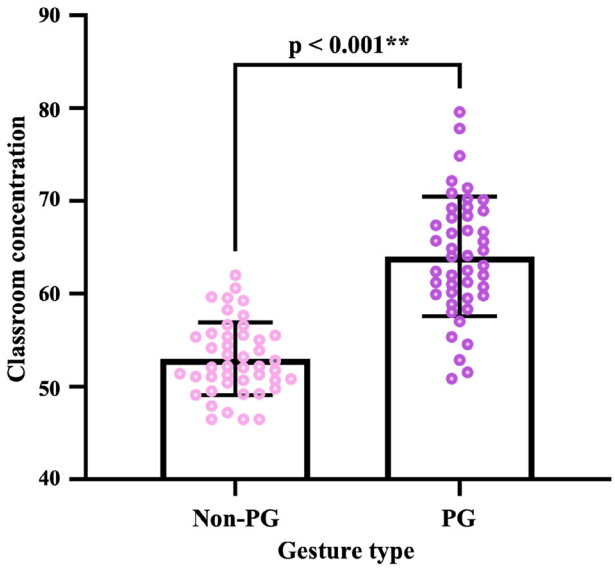

Fig. 7 The effect of pointing gestures on classroom concentration

(0 < PG ≤ 3), and high (PG > 3). The ANOVA analysis in Fig. 8 and the post hoc test (LSD) results indicated significant differences in classroom concentration among the three levels ($F = 62.836$, $p < 0.001$).

Specifically, compared to low-level pointing gestures, both normal-level (MD = 9.38, $p < 0.001$) and high-level (MD = 16.29, $p < 0.001$) pointing gestures significantly enhanced students' classroom concentration. Nevertheless, the positive effect of high-level pointing gestures on learners' classroom concentration was less pronounced than normal-level pointing gestures (MD = −6.91, $p < 0.001$). Therefore, hypothesis H2 is true.

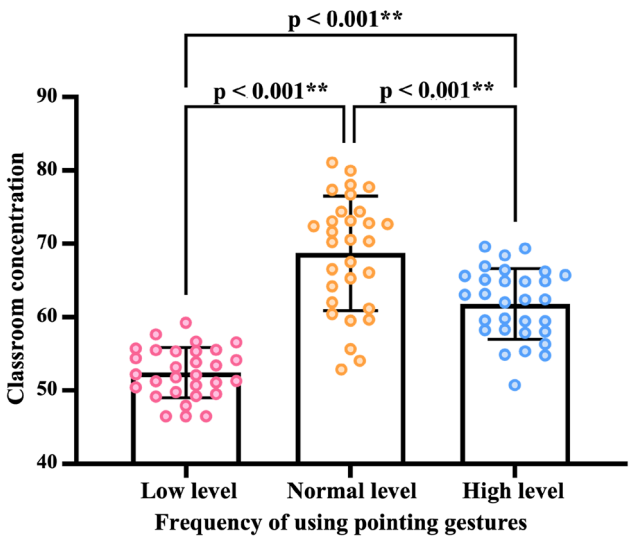

Fig. 8 The effect of different levels of pointing gestures on classroom concentration

4 Discussion

4.1 Automatic assessment of classroom concentration based on computer vision technology

This study introduces a novel model for recognizing classroom concentration by leveraging head pose and facial expression data, which is further validated for its efficacy through EEG equipment. The experimental results demonstrate the successful monitoring of classroom concentration fluctuations using our method. The human brain plays a vital role in attention control, influencing individuals' responses to the external environment and regulating their behavior [35, 36]. Distinct patterns of neural activity in specific brain cortical areas manifest when an individual enters a state of focused attention [37]. These patterns of neural activity reflect the brain's processing of particular stimuli in the environment, including perception and sensitivity to different types of information. During states of concentration, there is a significant increase in neural activity in the prefrontal region of the brain's cortex, which is crucial for attentional control and executing complex tasks. Similarly, variations in patterns of neural activity can indicate levels of human attention and can be used to evaluate classroom concentration accordingly. These changes are also reflected in learners' behavior, including their head posture and facial expressions. Attentive learners, for example, demonstrate active listening, maintaining eye contact with the teacher or the blackboard, and sitting upright. Their facial expressions may also exhibit signs of concentration, such as smiling, frowning, or pursed lips [38]. These manifestations indicate learners' level of concentration and reflect the quality of the new knowledge they acquire during the learning process [39].

Our classroom concentration assessment model uses computer vision technology to identify learners' head postures and facial expressions, combined with a large number of brain science research and machine learning algorithms, to achieve automatic assessment of learners' classroom concentration. However, it is essential to note that head posture and facial expressions are only one of the manifestations of a learner's concentration state [40, 41]. The learner's internal concentration state may vary from person to person, and head posture and facial expressions cannot fully reflect the learner's internal concentration state. Therefore, adopting a comprehensive approach that integrates multiple assessment methods, such as EEG and behavioral tasks, can provide a more comprehensive assessment of learners' concentration, enabling a more accurate evaluation of their performance in the learning process [42]. In future research, we can enhance the classroom concentration assessment model by incorporating additional biological signals and behavioral features, such as heart rate variability and sound analysis, to improve the accuracy and reliability of concentration assessment.

4.2 The influence of pointing gestures on classroom concentration

The present study meticulously investigated the influence of pointing gestures on students' classroom concentration. The outcomes derived from the one-way ANOVA analysis revealed a substantial enhancement in learners' classroom concentration when teachers employed pointing gestures. Consequently, H1 is corroborated. However, the results of one-way ANOVA for different levels of pointing gestures show that pointing gestures have a boundary effect on improving classroom attention, which is consistent with H2. Selective attention theory posits that individuals cannot simultaneously attend to all presented stimuli. Instead, they selectively focus on specific stimuli of interest while consciously disregarding others [43] proficiently aids learners in swiftly identifying and processing pertinent materials exhibited on the blackboard or electronic whiteboard [44]. Pointing gestures effectively mitigate distractions by directing students' visual attention toward the instructional content. Additionally, the teacher's deliberate incorporation of pointing gestures cultivates an interactive and immersive learning environment, thereby encouraging learners to actively engage in knowledge construction by allocating their limited cognitive resources. Consequently, when students observe pointing gestures during classroom instruction, they display behavioral attentiveness, augmenting their overall concentration [45]. However, the positive impact of teacher gestures on classroom concentration is limited, especially over a short period [46]. Previous research has indicated that the simultaneous presentation of different information through speech and gestures can encourage learners to integrate information across modalities actively [47]. However, when the information in both modalities becomes highly redundant, the surplus information may impose a cognitive burden on learners' working memory. Therefore, it is important to strike a balance between the use of teaching gestures and the cognitive load.

For example, in our study, we investigated two teaching scenarios. In the first scenario, the teacher presents a question to the students concerning the sum of the interior angles of a regular hexagon. Simultaneously, the teacher points to the regular hexagon displayed on the electronic whiteboard, which shows only one regular hexagon. In the second scenario, when the teacher explains the proof principle of unity of triangles to students, he points to the

angles and sides involved while explaining. In the first scenario, the teacher's pointing gestures are completely superfluous to his words, because the teacher's words have already been expressed very clearly. Even without gestures, the regular hexagon is familiar to middle school students. In the second scenario, the teacher's pointing gesture provides crucial information for the students' visual exploration. The excessive use of pointing gestures within a brief timeframe is predominantly similar to scenario one and involves repetitive language. Such pointing gestures that fail to provide meaningful information and excessively overlap with speech lack the ability to deliver sufficient external information to learners and overburden their cognitive resources, ultimately resulting in a decline in student concentration during class.

5 Conclusion

In conclusion, regulating students' classroom concentration is critical in optimizing the teaching process and improving learning outcomes. This study validated the effectiveness of the classroom concentration recognition model in recognizing classroom concentration and investigated the impact of pointing gestures on students' classroom concentration. Our findings demonstrate that the classroom concentration recognition model can effectively monitor the fluctuation of classroom concentration, and pointing gestures can enhance students' concentration but have a boundary effect. While our study provides important insights into optimizing classroom concentration and guiding teachers' behavior, we acknowledge that teachers' other nonverbal behaviors, such as language, gaze, and spatial position, may also affect students' concentration. Future research should explore these factors better to understand their impact on students' classroom concentration.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grant (62077015), the National Key R&D Program of China under Grant (2022YFC3303600), and the Natural Science Foundation of Zhejiang Province under Grant (LY23F020010).

Data availability Data analyzed in the project are available upon request. Please contact the corresponding author.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Moore DW, Anderson A, Glassenbury M et al (2013) Increasing on-task behavior in students in a regular classroom: effectiveness of a self-management procedure using a tactile prompt. J Behav Edu 22:302–311
2. Smithson EF, Phillips R, Harvey DW et al (2013) The use of stimulant medication to improve neurocognitive and learning outcomes in children diagnosed with brain tumours: a systematic review. Eur J Cancer 49(14):3029–3040
3. Napoli M (2004) Mindfulness training for teachers: a pilot program. Complement Health Pract Rev 9(1):31–42
4. Risko EF, Anderson N, Sarwal A et al (2012) Everyday attention: variation in mind wandering and memory in a lecture. Appl Cogn Psychol 26(2):234–242
5. Bhanji F, Gottesman R, de Grave W et al (2012) The retrospective pre-post: a practical method to evaluate learning from an educational program. Acad Emerg Med 19(2):189–194
6. Bunce DM, Flens EA, Neiles KY (2010) How long can students pay attention in class? A study of student attention decline using clickers. J Chem Edu 87(12):1438–1443
7. Deng Q, Wu Z (2018) Students' attention assessment in elearning based on machine learning. In: IOP Conference series: earth and environmental science, IOP Publishing, p 032042
8. Renawi A, Alnajjar F, Parambil M, et al (2021) A simplified real-time camera-based attention assessment system for classrooms: pilot study. Edu Inf Technol: 1–18
9. Hu M, Wei Y, Li M et al (2022) Bimodal learning engagement recognition from videos in the classroom. Sensors 22(16):5932
10. Rayner K (1998) Eye movements in reading and information processing: 20 years of research. Psychol Bull 124(3):372
11. Mahon A, Clarke AD, Hunt AR (2018) The role of attention in eye-movement awareness. Atten Percept Psychophys 80:1691–1704
12. D'Mello S, Cobian J, Hunter M (2013) Automatic gaze-based detection of mind wandering during reading. In: Educational data mining
13. Behzadnia A, Ghoshuni M, Chermahini S (2017) EEG activities and the sustained attention performance. Neurophysiology 49(3):226–233
14. Chiang HS, Hsiao KL, Liu LC (2018) EEG-based detection model for evaluating and improving learning attention. J Med Biol Eng 38:847–856
15. Hostetter AB, Alibali MW (2008) Visible embodiment: gestures as simulated action. Psychon Bull Rev 15:495–514
16. Ping R, Goldin-Meadow S (2010) Gesturing saves cognitive resources when talking about nonpresent objects. Cogn Sci 34(4):602–619
17. Alibali MW, Nathan MJ, Wolfgram MS et al (2014) How teachers link ideas in mathematics instruction using speech and gesture: a corpus analysis. Cogn Instr 32(1):65–100
18. Koumoutsakis T, Church RB, Alibali MW et al (2016) Gesture in instruction: Evidence from live and video lessons. J Nonverbal Behav 40:301–315
19. Pi Z, Zhang Y, Yang J et al (2019) All roads lead to Rome: instructors' pointing and depictive gestures in video lectures promote learning through different patterns of attention allocation. J Nonverbal Behav 43:549–559
20. Pi Z, Zhang Y, Yu Q et al (2022) Neural oscillations and learning performance vary with an instructor's gestures and visual materials in video lectures. Br J Edu Technol 53(1):93–113
21. Brooks JR, Passaro AD, Kerick SE et al (2018) Overlapping brain network and alpha power changes suggest visuospatial attention effects on driving performance. Behav Neurosci 132(1):23

22. Yeo A, Ledesma I, Nathan MJ et al (2017) Teachers' gestures and students' learning: sometimes "hands off" is better. Cogn Res Princ Implic 2:1–11
23. Dargue N, Sweller N (2018) Not all gestures are created equal: the effects of typical and atypical iconic gestures on narrative comprehension. J Nonverbal Behav 42:327–345
24. He K, Zhang X, Ren S, et al (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
25. Fredricks JA, Blumenfeld PC, Paris AH (2004) School engagement: potential of the concept, state of the evidence. Rev Edu Res 74(1):59–109
26. Deng J, Guo J, Ververas E, et al (2020) Retinaface: Single-shot multi-level face localisation in the wild. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition (CVPR)
27. Pham L, Vu TH, Tran TA (2021) Facial expression recognition using residual masking network. In: 2020 25th International Conference on pattern recognition (ICPR), IEEE, pp 4513–4519
28. Goodfellow IJ, Erhan D, Carrier PL, et al (2013) Challenges in representation learning: A report on three machine learning contests. In: International conference on neural information processing, Springer, pp 117–124
29. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, pp 234–241
30. Ruiz N, Chong E, Rehg JM (2018) Fine-grained head pose estimation without keypoints. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 2074–2083
31. D'Mello S, Graesser A (2012) Dynamics of affective states during complex learning. Learn Instr 22(2):145–157
32. Xu X, Teng X (2020) Classroom attention analysis based on multiple euler angles constraint and head pose estimation. In: MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part I 26, Springer, pp 329–340
33. Wu W, Liu H, Li L et al (2021) Application of local fully convolutional neural network combined with yolo v5 algorithm in small target detection of remote sensing image. PloS ONE 16(10):e0259283
34. Rebolledo-Mendez G, Dunwell I, Martínez-Mirón EA, et al (2009) Assessing neurosky's usability to detect attention levels in an assessment exercise. In: International Conference on human–computer interaction, Springer, pp 149–158
35. Posner MI, Petersen SE (1990) The attention system of the human brain. Ann Rev Neurosci 13(1):25–42
36. Corbetta M, Shulman GL (2002) Control of goal-directed and stimulus-driven attention in the brain. Nat Rev Neurosci 3(3):201–215
37. Foxe JJ, Snyder AC (2011) The role of alpha-band brain oscillations as a sensory suppression mechanism during selective attention. Front Psychol 2:154
38. Ekman P, Friesen WV (1971) Constants across cultures in the face and emotion. J Personal Soc Psychol 17(2):124
39. D'Mello S, Jackson T, Craig S, et al (2008) Autotutor detects and responds to learners affective and cognitive states. In: Workshop on emotional and cognitive issues at the international conference on intelligent tutoring systems, pp 306–308
40. Wang L (2018) Attention decrease detection based on video analysis in e-learning. Transactions on Edutainment XIV pp 166–179
41. Zhai X, Xu J, Chen NS et al (2023) The syncretic effect of dual-source data on affective computing in online learning contexts: a perspective from convolutional neural network with attention mechanism. J Edu Comput Res 61(2):466–493
42. Mühl C, Jeunet C, Lotte F (2014) EEG-based workload estimation across affective contexts. Front Neurosci 8:114
43. Broadbent DE (2013) Perception and communication. Elsevier, Amsterdam
44. Greiffenhagen C, Sharrock W (2005) Gestures in the blackboard work of mathematics instruction. International society for gesture studies
45. Alibali MW, Nathan MJ, Church RB et al (2013) Teachers' gestures and speech in mathematics lessons: forging common ground by resolving trouble spots. ZDM 45:425–440
46. Krahmer E, Swerts M (2007) The effects of visual beats on prosodic prominence: acoustic analyses, auditory perception and visual perception. J Mem Lang 57(3):396–414
47. So WC, Yi-Feng AL, Yap DF et al (2013) Iconic gestures prime words: comparison of priming effects when gestures are presented alone and when they are accompanying speech. Front Psychol 4:779

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.