Scholar Publishing Group
International Journal of Big Data Intelligent Technology
https://doi.org/10.38007/IJBDIT.2025.060103
ISSN 2790-0932 Vol. 6, Issue 1: 33-46

Classroom Attention Detection Based on Computer Vision and Artificial Intelligence

Lian Xue[1,a*]

[1]*School of Computer and Computing Science, Hangzhou City University, Hangzhou 310015, Zhejiang, China*

[a]*xuel@hzcu.edu.cn*

[*]*Corresponding author*

Keywords: Attention Analysis, Graph Convolutional Neural Network, Gesture Recognition, Human-object Interaction

Abstract: The academic level of students is the core issue that schools and parents pay attention to. There is a close relationship between students' academic level and students' attention status in class. Students who can maintain their concentration for a long time tend to have higher academic level. Therefore, it is very necessary to improve students' attention level in class. To this end, the attention of the students needs to be detected first. Computer vision algorithms typically employ third-person image or video data. Due to the large number of students in the classroom, the algorithm is difficult to extract small objects and occluded objects, so it cannot accurately detect students' attention. The attention detection model proposed in this paper included three modules: gaze point estimation, gaze target recognition and attention level analysis. The gaze estimation module used a deep learning algorithm that combines saliency detection with attention shifting. In order to improve the accuracy of the algorithm, this paper tried three attention mechanisms, namely spatial attention, channel attention and mixed attention. The specific approach was to embed the attention module after the convolution block of the Convolutional Neural Network (CNN). In order to achieve the goal of obtaining students' classroom attention data in batches and accurately, this paper proposed a first-person video-based student classroom attention detection method. Compared with the third-person video, the first-person video had the advantages of one-to-one correspondence between the video and the students, and the video content was consistent with the gaze behavior, so it could make the students' attention detection more accurate. The experiments in this paper showed that the sum of the ratio of student 1's fixation to learning-related goals was 0.51, while that of student 2 was 0.56. Therefore, student 2's attention state was more concentrated, and the final algorithm model was better than the baseline.

1. Introduction

Attention refers to the behavioral and cognitive process of selectively focusing on some specific information. Relevant studies have shown that selective and sustained attention has a positive effect on learning, and students who pay attention in class tend to be better academically. In order to improve the level of students' attention in the classroom, the students' current attention situation should be known. Ultimately, the purpose of improving teaching quality and teaching students in accordance with their aptitude is achieved.

The traditional detection methods of students' attention include questionnaire survey method and instrument detection method. Although the questionnaire survey method can save time, money and human resources, and facilitate statistical analysis, it usually cannot guarantee the quality and reliability of the experimental results. This is because these methods are largely subject to subjective factors. With the continuous development of science and technology, computer vision algorithms are gradually applied to the detection of students' attention.

Attention plays a huge role in students' learning process, and many scholars have studied it. Through the extraction of yarn body and hairiness, Li S Y created a two-scale attention model in feature extraction, which could fully simulate human attention to the overall and detailed information of yarn at different viewing distances [1]. Oord S tested the effectiveness of short-term (six) individual cognitive-behavioral program interventions for Attention-Deficit/Hyperactivity Disorder (ADHD) students. Some patients with ADHD were treated, and random follow-up was conducted after treatment. For further treatment, adaptation to the therapy was required [2]. The teaching and learning process takes place formally in the classroom which is the main learning space of the school education system. Norazman N believed that it should be implemented effectively to achieve the satisfaction and comfort expected by students. So that students were no longer affected by the lack of a good educational environment, and the classroom capacity could be increased to achieve the purpose [3]. maharmeh used the Functional Behavior Assessment (FBA) technique, with the effect of: The technology was being used to provide special needs center education services to patients with autism spectrum disorder and attention deficit hyperactivity disorder. And sensitive data would not be shared on the network [4]. Abate A F found that in traditional classroom education, teachers perceived or gained students' participation through observation. Distance education suffered from a lack of feedback from the expressions and behavior of students in class. His proposed approach was considered a software architecture that ran locally and on student personal computers. Sensitive data was not shared over the network [5]. However, due to insufficient data sources and unprofessional research equipment, the above research is only in the theoretical stage and cannot be practiced.

Research on student attention using computer vision and artificial intelligence is an innovative project. Among them, Caroline conducted a six-week study of selected kindergarten students and teachers to understand their perceptions of teaching and learning in traditional indoor classrooms versus new outdoor classrooms. Data showed that their perception of well-being, joy and interest increased when teaching in outdoor classrooms [6]. Fu K proposed image captioning in the model. The shift of attention between visual regions in the system enabled the transition to impose an ordering cue in visual perception. This alignment characterized the flow of latent meaning, which encoded what the visual scene and textual descriptions semantically shared [7]. Attention is a mechanism that alters the processing of sensory stimuli by enhancing the perception of task-related information and inhibiting the perception of task-irrelevant stimuli. Research by Attigodu G showed that a single sensory modality minimized the impact of stimuli in unattended sensory modalities and reduced the integration of stimuli in multiple sensory modalities [8]. Effective methods proposed by Liu S were a bottom-up single-shot Convolutional Neural Network and a top-down operation

strategy. Through refinement, an intelligent agent with an attention mechanism was built to locate sub-regions that might contain relevant objects and zoom in to fire the detector again [9]. Sangeroki B A learned to detect chest diseases using Convolutional Neural Network of chest X-ray images, which was the best method. However, CNN usually had a slow running time, so he used the attention mechanism to integrate into a lightweight CNN model [10].

In this paper, a convolutional network algorithm was proposed to provide a structural basis for gaze estimation algorithms. Three templates of the attention detection model were proposed to detect students' classroom attention. The novelty is that the three attention mechanisms used fit the theme very well, and they can complement and contrast each other to a certain extent.

2. Student Attention Evaluation Methods

2.1 Convolutional Network Algorithm

"Dual stream" is divided into spatial stream and temporal stream. The main contributions of the two-stream convolutional network are as follows: First, a two-stream convolutional network combining spatiotemporal information is proposed; second, it is proved that multi-frame dense optical flow can effectively extract temporal features; finally, it is proved that multi-task learning can both expand training data and improve the training effect [11]. The two-stream convolutional network is the basic structure of the gaze estimation algorithm in the method. The structure diagram of the two-stream convolutional network is shown in Figure 1.

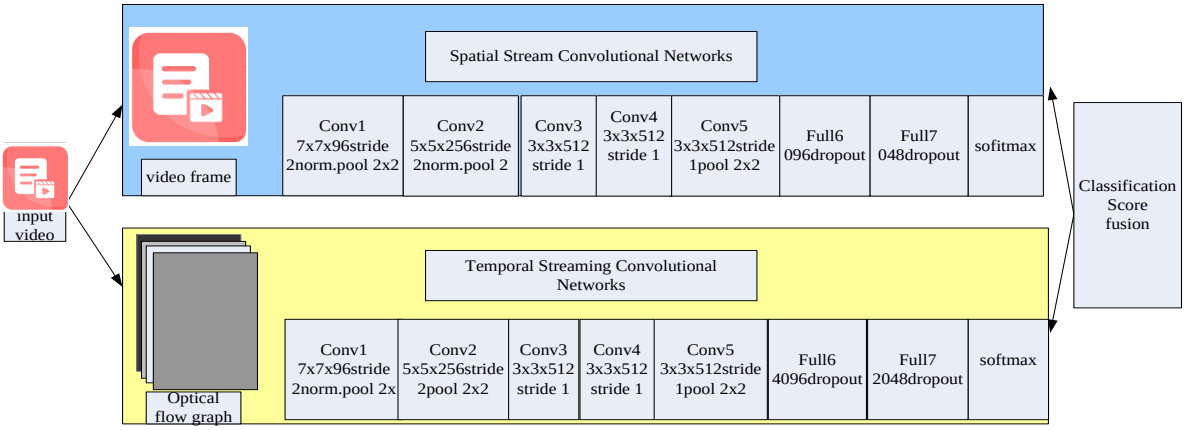

Figure 1. Two-stream convolutional network structure diagram

As shown in Figure 1, when observing a moving target, the image of the target on the retina is constantly moving and changing, just like a series of light and shadows passing quickly on the retina, so it is vividly called light flow. The optical flow field can be obtained by calculating the motion speed of all pixels in a frame of image. The speed is a vector. This includes both size and orientation [12]. For example, the pixel P coordinate in the t-th frame is (x, y). In frame +1 the pixel P moves to the (x,y) position. The key question is how to determine the position of pixel P in frame +1.

Head pose estimation is essentially the collection and recognition of head features using techniques such as computer vision. An estimate of the head pose is analyzed. These methods have various limitations and advantages due to differences in detection metrics.

This method is more suitable for pose estimation when the images to be tested are all frontal faces. However, the complexity of time and space required for its detection is relatively large. And the detection would be affected by some other factors such as identity, which leads to poor detection

performance. The method based on the detector array ignores some changes that are not related to the attitude, which improves the accuracy to a certain extent. But when training for each pose, the workload of its detector is heavy. It is necessary to provide positive and negative training samples, which would make the detection process cumbersome [13]. However, this is more easily affected by the resolution and depends on the localization accuracy of the feature points. In this paper, a head pose estimation algorithm based on Convolutional Neural Network is used to realize it. The structure of the Convolutional Neural Network is shown in Figure 2.

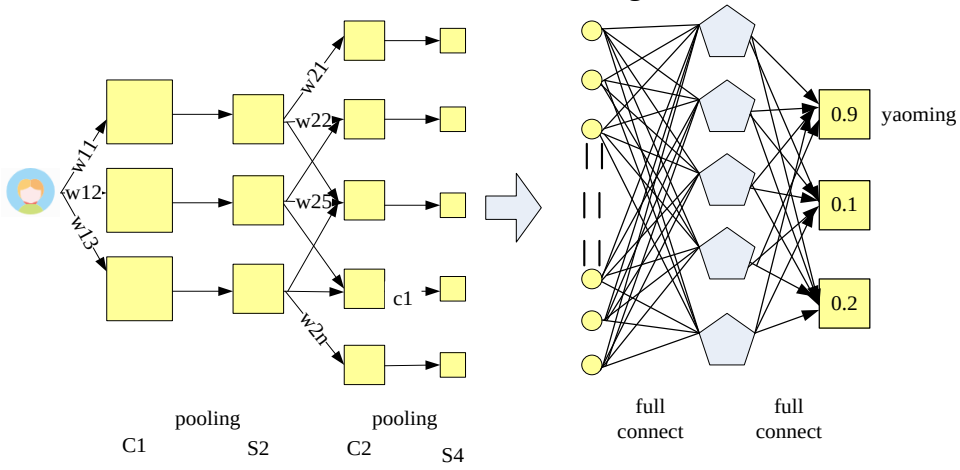

Figure 2. Convolutional Neural Network structure

As shown in Figure 2, the classifier is used to divide it, so as to obtain the result of the Convolutional Neural Network detection. A schematic diagram of a CNN is shown in Figure 3.

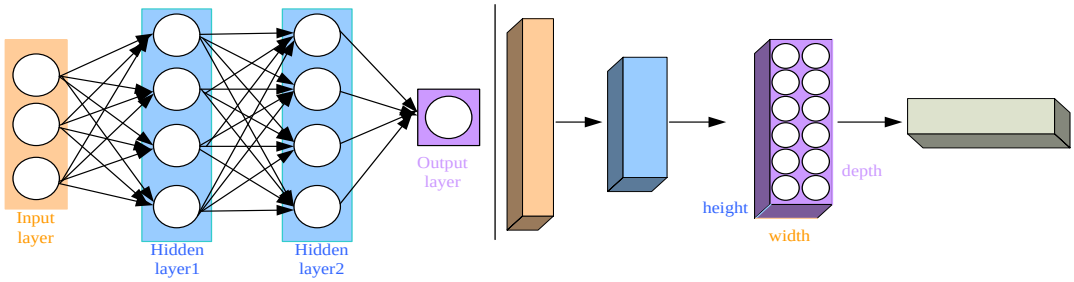

Figure 3. Schematic diagram of a CNN

As shown in Figure 3, in order to solve the application problem of neural networks on images, CNN came into being. Convolutional Neural Network use a three-dimensional version of neurons to solve this problem. Specifically, a neuron has three dimensions: width, height, and depth. Still taking the picture in Figure 3 as an example, its input size is 32×32×3. Each neuron is only connected to a small part of the picture, not fully connected. In the last layer (output layer), the two dimensions of width and height are compressed, leaving only the dimension of depth. The output scores for each class are obtained [14]. The optimization method of Convolutional Neural Network still adopts the method based on gradient descent.

2.2 Detecting Human Attention with Human Skeleton

ST-GCN's definition of action is video-oriented. Although the definition of attention discussed in this paper is picture-oriented, the utilization of human skeletons in ST-GCN greatly inspired the

approach. Therefore, ST-GCN is introduced in detail below. The ST-GCN flow chart is shown in Figure 4.

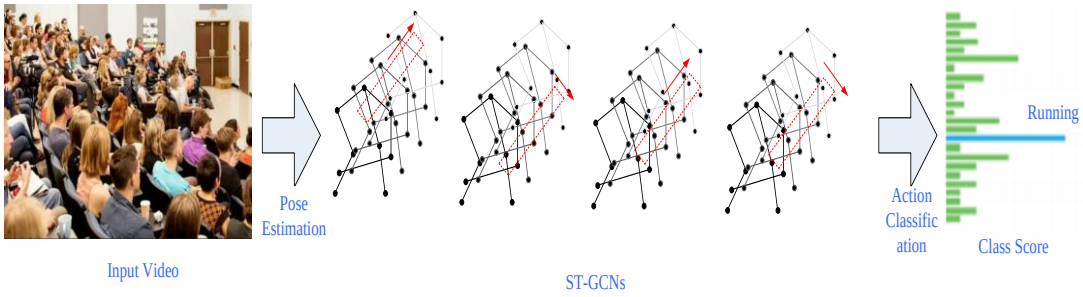

Figure 4. ST-GCN flow chart

As shown in Figure 4, the model is based on a series of skeleton graphs with two classes of edges. The first category is the edges connecting adjacent joints within the same frame, and the other category is the edges connecting the same joint within different frames. On this basis, ST-GCN establishes a multi-layer spatiotemporal graph convolution, which enables information to be embedded in the spatiotemporal dimension.

2.3 Using Faces to Detect Student Attention

The Haar-Like operator acts as a window in the actual calculation process. The window is enlarged at a certain ratio, and the sample is slid in the image with a certain step size [15]. During the calculation of eigenvalues, the pixels in the white block area are summed. The sum of the pixels in the black block area is subtracted as the one-dimensional feature collected by this operator. In this way, until the operator changes from the initial size to the maximum and traverses the entire image at the same time, the operator's feature acquisition process for this image is counted as complete. Similarly, other types of operators have to go through the same calculation and sampling process.

The definition formula for the problem of calculating the speed of Haar-Like features is as follows:

$$\mathrm{I}(a,b)=\sum_{a'\le a,b'\le b} i(a',b') \quad (1)$$

In the formula, I represents the integral graph, and I(a, b) represents the value at the point (a, b) of the integral graph coordinate. i represents the original image, and i(a', b') represents the pixel value at the point (a', b') of the original image. The fast and equivalent iterative calculation formula of Formula 1 is as follows:

$$\mathrm{W}(a,b)=S(a,b-1)+i(a,b) \quad (2)$$

$$I(a,b)=I(a-1,b)+W(a,b) \quad (3)$$

Due to the data characteristics of the integral graph, it is extremely convenient and fast to use it to calculate the value sum of the pixels of the rectangle at any position in the image. If the sum of the values of the pixels in the D area is calculated, the formula is as follows:

$$\mathrm{F}_D=I_1+I_4-(I_2+I_3) \quad (4)$$

In the initialization stage of Adaboost algorithm, the initialization parameters of each weak classifier are the same. In the process of using Haar-Like feature to train the classifier, Adaboost algorithm would train a weak classifier according to each Haar-Like feature operator. The expression of the weak classifier is as follows:

$$h(a, f, p, \theta) = pf(a) < p\theta \quad or(h(a, f, p, \theta) = 0, other) \tag{5}$$

The selection process of the optimal threshold of the weak classifier above needs to go through multiple iterations. Finally, all the weak classifiers with the best threshold selected are linearly weighted to form a strong classifier, and the expression of the combination is as follows:

$$S(a) = 1, \sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t \quad or(S(0) = 0, other) \tag{6}$$

Among them, T is the number of weak classifiers, and αt is the weighting parameter of the t-th weak classifier. The average probability is:

$$\frac{1}{2}(\sum_{t=1}^{T} \alpha_t \cdot 1 + \sum_{t=1}^{T} \alpha_t \cdot 0) = \frac{1}{2} \sum_{t=1}^{T} \alpha_t \tag{7}$$

The mapping process from the feature points of the initialized face to the feature points of the manually annotated face. In the Newton descent method, when the Hessian matrix is positive definite, the minimum value can be solved iteratively by solving a system of linear formulas. An initial estimate $x \in RP*1$ is used to represent a p-dimensional column vector. The updated iterative formula of the Newton descent method is as follows:

$$a_{k+1} = a_k - H^{-1}(a_k) J_f(a_k) \tag{8}$$

In SDM, the update iteration formula of its descent method can be written as the approximate Formula 10 of the update iteration formula of Newton's method according to the chain derivation method of the matrix:

$$J_f(a_k) = J_h^T(a_k)(\phi_k - \phi_*) \tag{9}$$

$$a_{k+1} = a_k - 2H^{-1}(a_k) J_h^T(a_k)(\phi_k - \phi_*) \tag{10}$$

Among them, a_k represents the facial features of point a. The formula can be simplified to:

$$\Delta a_1 = R_0 \phi_0 + b_0 \tag{11}$$

So far, the offset Δx of the feature point movement of the phased solution has been transformed into the form of a linear function. The process of solving coefficients R_0, b_0 does not have to devolve into solving the inverse and Jacobian of the Hessian matrix. It can be solved only by reducing the previous formula to the least squares form. A lot of complex computing processes are bypassed. The least squares form is as follows:

$$loss = \left\| h(d(a_0 + \Delta x)) - \phi_* \right\|_2^2 \tag{12}$$

The coefficients R, b_0 in Formula 12 can be derived from the least squares formula. d is the characteristic value of point x and its value is known, so the first step iteration result Δx can be obtained. The process of solving the target feature points by the gradient descent method supervised by SDM is a regression solving process. After several iterations using the initial face feature point information, the position of the target feature point is obtained. The objective function is as follows:

$$f(a_0 + \Delta x) = \left\| h(d(a_0 + \Delta x)) - \phi_* \right\|_2^2 \tag{13}$$

A schematic diagram of the average angle error is shown in Figure 5.

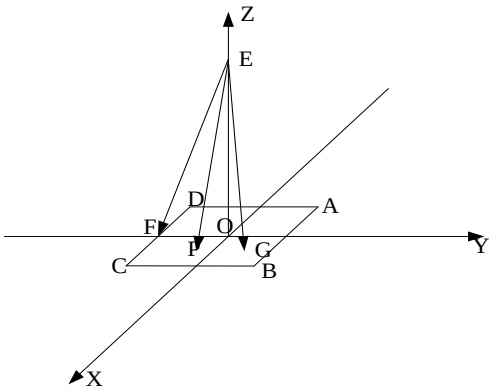

Figure 5. Schematic diagram of average angle error

As shown in Figure 5, in the space Cartesian coordinate system, the quadrilateral ABCD in the XOY plane represents an image with a size of 224 × 224. The center of the image coincides with the coordinate origin O, F is the midpoint of CD, P is the estimated position of the gaze point, G is the real position, and E is a point on the Z-axis. The angle α between EF and EO is 30°, and the angle 0 between EP and EG is the angle error of gaze point estimation. Set EP=, EG=r, from the inner product and outer product formulas of vectors, the following can be obtained:

$$\vec{r}_1 \cdot \vec{r}_2 = |\vec{r}_1||\vec{r}_2| \cos\theta \quad (14)$$

$$|\vec{r}_1 \cdot \vec{r}_2| = |\vec{r}_1||\vec{r}_2| \sin\theta \quad (15)$$

Divide Formula 15 by Formula 14 to get:

$$\frac{|\vec{r}_1 \times \vec{r}_2|}{\vec{r}_1 \cdot \vec{r}_2} = \frac{\sin\theta}{\cos\theta} = \tan\theta \quad (16)$$

From Formula 16, it can be obtained as follows:

$$\theta = \arctan \frac{|\vec{r}_1 \times \vec{r}_2|}{\vec{r}_1 \cdot \vec{r}_2} \quad (17)$$

The formula for the average angle error AAE is as follows:

$$AAE = \frac{1}{N} \sum_{x=1}^{N} \theta_1 \quad (18)$$

Precision refers to the correct proportion of all objects detected. The recall rate refers to the proportion of correctly detected objects to the total number of objects. Therefore, the higher the precision and recall, the closer to 1, the better the performance of the target detection algorithm. The formulas for calculating precision and recall are as follows:

$$\mathrm{Pr}ecision = \frac{TP}{TP + FP} \quad (19)$$

$$\mathrm{Re}call = \frac{TP}{TP + FN} \quad (20)$$

Compared with the PR curve, the specific value can more intuitively show the performance of the target detection algorithm. The AP metric is usually used. The higher the value of AP, the closer it is to 1, the better the performance of the target detection algorithm. The formula for calculating AP is as follows:

$$AP = \int_0^1 P(r)dr \quad (21)$$

Among them, p represents precision, r represents recall, and p is a function with r as an independent variable. This is equivalent to finding the area under the PR curve. However, the above integral is an ideal calculation method. In practice, the approximate AP algorithm is usually used to approximate the integral:

$$AP = \sum_{i=1}^{N} P(k)\Delta r(k) \tag{22}$$

Among them, N represents the number of prediction boxes belonging to a certain category in the test set, and p(k) represents the value of precision when there are i prediction boxes with confidence greater than or equal to the threshold. Δr(k) represents the change in the recall value when the number of prediction frames with a confidence greater than or equal to the threshold changes from k-1 to i (by reducing the threshold).

3. Student Attention Experiments

3.1 Student Attention Detection

In units of 20 seconds, the entire first-person video is divided into three stages: early, middle and late stages. The duration of students in each state of attention (concentrated or not) in each stage is counted, and the line chart of changes in attention state is shown in Figure 6.

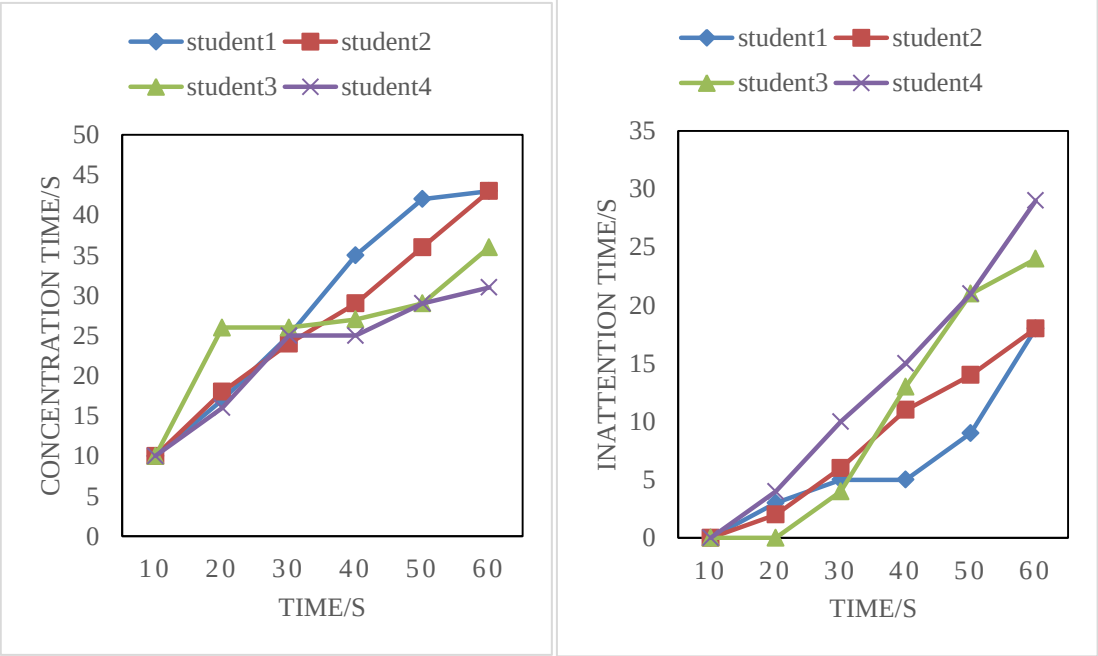

Figure 6. Line chart of attention state change

As shown in Figure 6, the following conclusions can be drawn: The attention of the 4 students in the early stage is in a state of concentration. Student 2 is completely focused in the early stage, but relatively scattered in the middle and later stages. Student 1's attention state changes quite obviously, and as time goes by, his attention becomes less and less concentrated. In the three stages, student 3's attention time is longer than his inattention time, and his attention time is the highest, so this student is the best performer among the 4 students. Student 4 is in a state of high concentration in the early and mid-term, but he becomes distracted in the later stage.

The statistical histogram of the attention status of all students is shown in Figure 7.

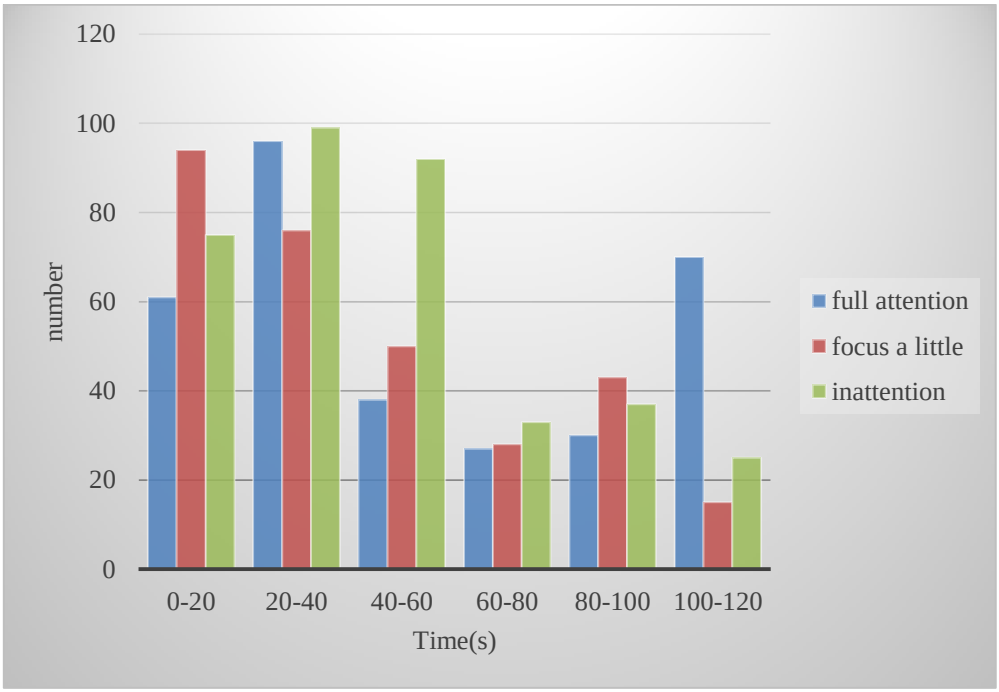

Figure 7. Histogram of segmentation statistics of the attention status of all students

Figure 8 and Figure 9 show the experimental results of the SSD algorithm corresponding to different evaluation indicators on the FPVSC dataset, as well as the precision and recall rates when the confidence threshold is set to 0.4.

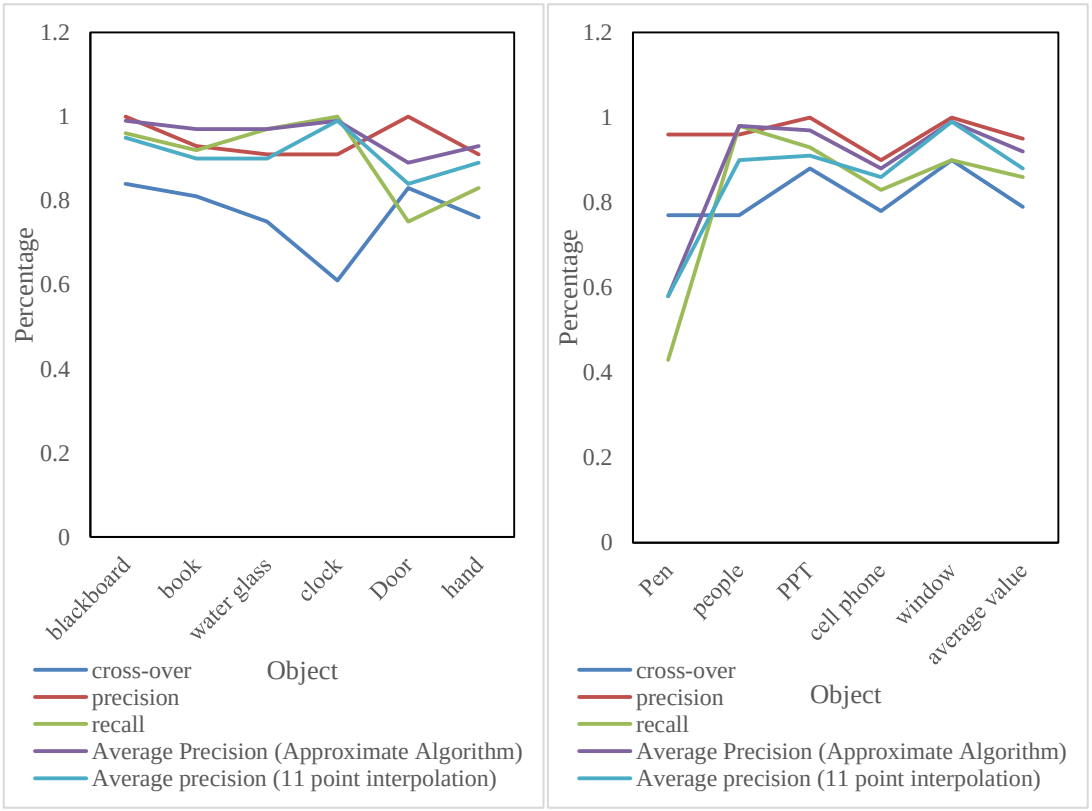

Figure 8. Experimental results of SSD algorithm on FPVSC dataset

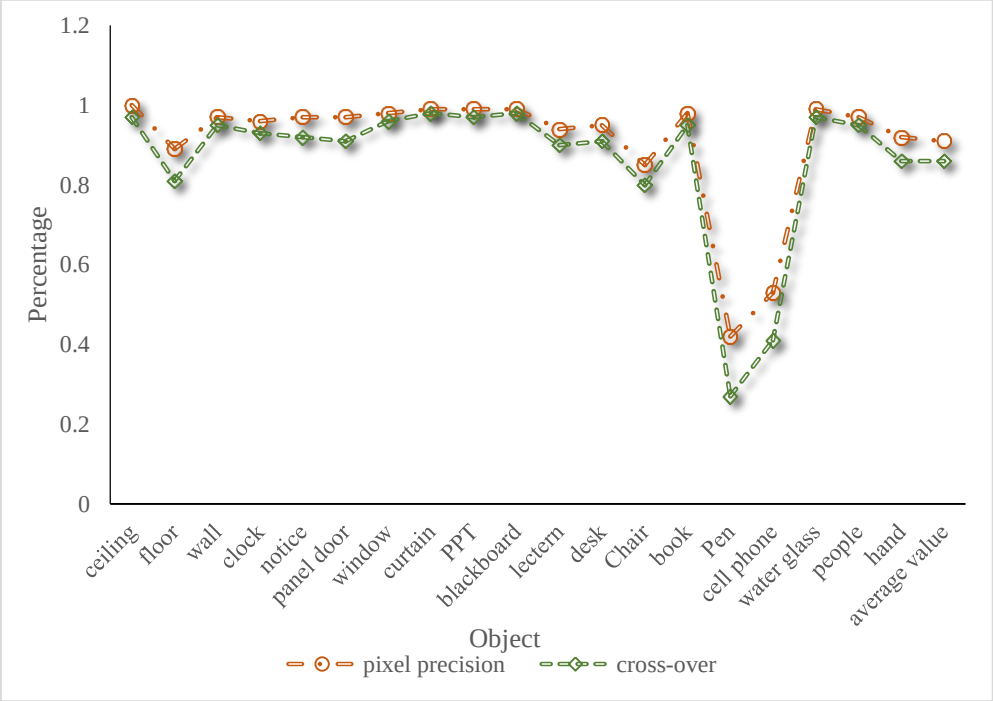

Figure 9. Experimental results of the Fast-SCNN algorithm on the FPVSC data set

As shown in Figures 8 and 9, the reason why the recall and average precision of the pen are much lower than other categories is that the shape of the pen is long and thin. It would be blocked by the hand when holding the pen to write. The detection results of the remaining 10 categories are all good, and the expected experimental results are achieved.

3.2 Detection Experiment of Head Pose

For the head pose, based on the estimation of pitch angle and yaw angle obtained by the inference module, this paper uses a different range based on the angle to divide. Subsets of different states are obtained and the evaluation grade division is determined [16]. The preliminary classification level is shown in Table 1.

Table 1. Preliminary rating division table for yaw and pitch angles

	-79 °-45 °	-45 °-0 °	0 °-45 °	45 °-75 °
-60 °--30 °	D1	D2	D3	D4
-30 °-0 °	D5	D6	D7	D8
0 °-30 °	D9	D10	D11	D12
>30 °	D13	D14	D15	D16

As shown in Table 1, when the vertical rotation angle of the student's head is -60 ° - 30 °, that is, when the angle is lower than the horizontal line, the student is in a state of bowing his head. In the classroom, it can be considered that the student's head is focused on the book, and the student is concentrating on reading or writing. It is initially determined that the degree of concentration of students at this angle is better, and then the problem of the horizontal rotation angle is discussed [17]. When the horizontal rotation angle is -79 °-45 ° or 45 °-75 °, the student's head angle is relatively deviated from the frontal direction. It can be assumed that the student is bowing his head and talking to other people, or doing something else. When conducting concentration analysis, it can be

considered that the student has a higher probability of doing something unrelated to learning, and a lower evaluation is given. When the vertical rotation angle of the student's head is within the range of >30 °, that is, the student's head sight is higher than the general listening level. Therefore, it is believed that there is a certain probability that the current students are distracted, and a low concentration evaluation would be given when the concentration is rated [18]. Then, according to the method of synthesizing the vertical rotation angle and the horizontal rotation angle of the student's head, the overall angle corresponding to the degree of concentration is obtained. The table is a preliminary detailed rating table. Teachers can divide the weights according to the specific situation, so as to change the rating corresponding to the specific behavior, as shown in Table 2.

Table 2. Concentration ratings for head angle

	-79 °-45 °	-45 °-0 °	0 °-45 °	45 °-75 °
-60 °--30 °	Ⅳ	Ⅰ	Ⅰ	Ⅳ
-30 °-0 °	Ⅱ	Ⅰ	Ⅰ	Ⅱ
0 °-30 °	Ⅱ	Ⅰ	Ⅰ	Ⅱ
>30 °	Ⅳ	Ⅲ	Ⅲ	Ⅳ

As shown in Table 2, real-time module inference is performed at a single time point of each test, and the results of micro-expression recognition and pose estimation of the current student are obtained. After that, the test results are sent to the evaluation and analysis rating table, and then corresponding can be done [19]. If there is an obvious deviation in the predicted result, the teacher can divide the weights according to the situation in the classification of the rating, so as to obtain a rating table suitable for the class [20]. In this paper, the default micro-expression evaluation and head pose evaluation each account for 50% of the focus evaluation. Teachers can divide the 4 types of angle ranges of the two angles by setting the importance priority, and the priority setting number is "1, 2, 3, 4, 5" five priority types. The number 1 indicates that the rank has the lowest weight. The higher the number, the higher the weight of the class. The number 5 indicates that the rank has the highest weight.

In this experiment, one video frame is extracted per second, so the number of fixations per second is fixed at 1. Therefore, the fixation time is equal to the fixation frequency, and the fixation time ratio is equal to the fixation frequency ratio. The eye movement measurement indicators of the two students participating in the attention detection experiment are shown in Table 3.

Table 3. Eye movement metrics for student 1 and student 2

gaze at the target	number of gazes	Longest Gaze Duration	average fixation duration	fixation ratio	look back time
teacher	7/5	7/3	7/1.67	0.11/0.05	0/2
blackboard	7/14	4/5	1.75/3.5	0.11/0.14	3/9
PPT	5/7	2/5	1.67/3.5	0.08/0.07	3/5
book	13/25	6/12	2.17/4.17	0.20/0.26	10/23
Pen	1/4	1/2	1/2	0.01/0.04	0/2
hand	7/10	2/5	1.17/2	0.11/0.10	6/9
water glass	2/8	2/5	2/4	0.03/0.08	0/3
cell phone	6/3	2/1	1.5/1	0.09/0.03	4/2
clock	4/6	3/4	2/3	0.06/0.06	3/2
Door	3/3	2/1	1.5/1	0.05/0.03	2/2

window	3/7	2/4	1.5/3.5	0.05/0.07	2/4
background	6/5	2/1	1.17/1	0.10/0.05	6/4

As shown in Table 3, the data of student 1 is on the left side of the slash, and the data of student 2 is on the right side of the slash. The look-back time can indicate the students' interest in the gaze target, and the experimental results show that student 2 has a strong interest in books. The sum of student 1's fixation ratios to learning-related goals is 0.51, while that of student 2 is 0.56. Therefore, the attention state of student 2 is more concentrated.

Student pupil detection is also one of the important basis for the detection of students' gaze direction. The student pupil detection mainly adopts the Hough circle transform method which is adaptive to the illumination. The student pupil detection is performed when the pupils' eyes are open, so the detection needs to exclude the closed eyes. The specific detection results of the experiment are shown in Table 4:

Table 4. Accuracy of pupil detection

sample	total frames	The number of frames marked with eyes open	The number of frames in which the pupil was detected	Detection accuracy
Video1	4000	3699	3532	95.49%
Video2	4000	3571	3350	93.8 1%
Video3	4000	2954	2818	95.40%
Video4	4000	2697	2436	90.32%
Video5	4000	3380	3222	95.33%
Video6	4000	2611	2550	97.66%
Video7	4000	2357	2278	96.65%
Video8	4000	2890	2568	88.86%

As shown in Table 4, the detection results are shown in the following table, which also includes a group of control experiments. The experimental method uses annotated samples to classify by linear Support Vector Machine (SVM). The accuracy of attention detection and the accuracy of comparative experiments are shown in Table 5.

Table 5. Accuracy of attention detection and comparative experimental accuracy

sample	total frames	The number of frames detected by the SVM method	The number of frames detected by this method	SVM method detection accuracy	The detection accuracy of this method
Video1	4000	919	3726	22.98%	93.15%
Video2	4000	1167	3606	29.18%	90.15%
Video3	4000	1921	3567	48.03%	89.18%
Video4	4000	840	3491	21.00%	87.28%
Video5	4000	2100	3695	52.50%	92.38%
Video6	4000	1822	3617	45.55%	90.43%
Video7	4000	1484	3624	37.10%	90.60%
Video8	4000	1609	3674	40.23%	91.85%

As shown in Table 5, it can be seen from the above control experiments that the samples labeled based on image data and directly classified by the sample labels are not distinguishable. Student

attention detection is a physiological state detection based on the detection of other indicators.

4. Conclusions

This paper analyzed the advantages and disadvantages of using various solutions for student attention analysis. A two-stage classification algorithm was proposed. The training of the whole algorithm only needed the annotation information based on the category, and no other additional annotation information was needed. According to this requirement, a dataset was made. The data set was collected from the surveillance camera data of the school, which was very close to the practical application. This algorithm model used a position correction module to model the relative position of the input target and the blackboard. All targets were aligned according to the relative positions to the blackboard by this position correction module. This type of algorithm was mostly used in business scenarios, and it was difficult to define and standardize system solutions related to student attention research. Therefore, there was no unified data set and algorithm solutions that could be referenced and compared. Therefore, this paper could not compare the performance of this system with other student attention detection systems. Each research team developed its own evaluation criteria based on the collected data. The criteria were applicable to some scenarios but not all scenarios. Therefore, it was difficult to make an accurate comparison of results based on different criteria. Aiming at the research goal of students' classroom attention detection, this paper proposed a detection method based on first-person video and established the corresponding first-person video data set of students' classroom.

References

[1] Li S Y, Xu B G, Fu H, Tao X M, Chi Z R. A two-scale attention model for intelligent evaluation of yarn surface qualities with computer vision [J]. Journal of the Textile Institute, 2018, 109 (6): 798-812.

[2] Oord S, Boyer B E, Dyck L V, Mackay K J, Meyer H D, Baeyens D. A Randomized Controlled Study of a Cognitive Behavioral Planning Intervention for College Students With ADHD: An Effectiveness Study in Student Counseling Services in Flanders: [J]. Journal of Attention Disorders, 2020, 24 (6): 849-862.

[3] Norazman N, Che-Ani A I, Ja'Afar N H, Khoiry M A. Standard compliance and suitability of classroom capacity in secondary school buildings [J]. Journal of Facilities Management, 2019, 17 (3): 238-248.

[4] maharmeh, lina, mahmoud. Reducing The Rate of Behavioral Problems for Students with ASD & ADHD using the Techniques of FBA [J]. International Journal for Research in Education, 2019, 43 (2): 9-9.

[5] Abate A F, Cascone L, Nappi M, Narducci F, Passero I. Attention monitoring for synchronous distance learning [J]. Future Generation Computer Systems, 2021, 125 (4): 774-784.

[6]Caroline, Guardino, Katrina, W., Hall, Erin, et al. Teacher and student perceptions of an outdoor classroom [J]. Journal of Outdoor and Environmental Education, 2019, 22 (2): 113–126.

[7] Fu K, Jin J, Cui R, Fei S, Zhang C. Aligning Where to See and What to Tell: Image Captioning with Region-Based Attention and Scene-Specific Contexts [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39 (12): 2321-2334.

[8] Attigodu G, Berthommier F, Nahorna O, Schwartz J L, Attigoduchandrashekara G, Olhanahorna J L, et al. effect of context, rebinding and noise, on audiovisual speech fusion. 14th annual conference of the international speech communication association [J]. Exp Brain Res, 2017, 184 (1): 39-52.

[9] Liu S, Huang D, Wang Y. Pay Attention to Them: Deep Reinforcement Learning-Based Cascade Object Detection [J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31 (7): 2544-2556.

[10] Sangeroki B A, Cenggoro T W. A Fast and Accurate Model of Thoracic Disease Detection by Integrating Attention Mechanism to a Lightweight Convolutional Neural Network [J]. Procedia Computer Science, 2021, 179 (11): 112-118.

[11] Luo X, Hu H. Selected and refined local attention module for object detection [J]. Electronics Letters, 2020, 56 (14): 712-714.

[12] Wang Y, Gu X. Using of Attention for Scene Text Detection [J]. Journal of Computer-Aided Design & Computer Graphics, 2022, 33 (12): 1908-1915.

[13] Zhou M, Zou Z, Shi Z, Zeng W J, Gui J. Local Attention Networks for Occluded Airplane Detection in Remote Sensing Images [J]. IEEE Geoscience and Remote Sensing Letters, 2020, 17 (3): 381-385.

[14] Liang Y, Qin G, Sun M, Yan J, Jiang H. MAFNet: Multi-style attention fusion network for salient object detection [J]. Neurocomputing, 2021, 422 (2): 22-33.

[15] Avoliot B J, Gardner W L. Authentic leadership development: Getting to the root of positive forms of leadership [J]. IEEE Engineering Management Review, 2017, 16 (3): 315-338.

[16] Baglio K J. Student Motivation in the Latin Classroom [J]. Journal of Classics Teaching, 2022, 23 (45): 75-78.

[17] Chen J, Wan L, Zhu J, Xu G, Deng M. Multi-Scale Spatial and Channel-wise Attention for Improving Object Detection in Remote Sensing Imagery [J]. IEEE Geoscience and Remote Sensing Letters, 2020, 17 (4): 681-685.

[18] Kempf E, Manconi A, Spalt O. Distracted Shareholders and Corporate Actions [J]. Review of Financial Studies, 2017, 30 (5): 1660-1695.

[19] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39 (6): 1137-1149.

[20] Zhao Z, Bao Z, Zhang Z, Cummins N, Sun S, Wang H, et al. Self-attention transfer networks for speech emotion recognition [J]. Virtual Reality & Intelligent Hardware, 2021, 3 (1): 43-54.