Advances in Artificial Intelligence, Big Data and Algorithms
G. Grigoras and P. Lorenz (Eds.)

doi:10.3233/FAIA230890

Research on Student Classroom Attention Detection Based on Multi-Feature Fusion

Jihu SUN, Long MA, Song HE[1], Qiushi XI and Jun HU
Air Force Early Warning Academy, Wuhan, Hubei, China

Abstract: With the proposal of the concept of "AI+education", the research on learner attention has received great attention from scholars. Student classroom attention is an important aspect of educational research, which directly affects their learning status and efficiency in class, and also provides objective data support for teachers to adjust classroom rhythms and teaching methods. This paper proposes a multi-feature fusion method for attention detection based on artificial intelligence, which applies multiple features such as eye, mouth, and facial expression to determine whether students are in a state of focused learning. Compared with only detecting body posture or single facial feature detection, this method improves the accuracy of detection.

Keywords: classroom attention, Face features, multi-feature fusion

1. Introduction

Traditionally, student behavior status annotation is mainly carried out by professional personnel in the classroom to observe students' classroom performance, record students' behavioral actions, head posture, and participation level in classroom teaching activities, and then judge their learning status based on experience. However, this method has low efficiency. With the widespread application of information technology such as artificial intelligence and computer vision in the field of education, attention evaluation is also developing towards intelligentization and informatization. Fully mining and analyzing students' learning status data and conducting personalized teaching can stimulate students' enthusiasm for learning and thinking and improve learning outcomes.For the determination of learning focus, Gupta et al. [1] analyze students' learning investment through four different emotions displayed by students in the classroom: high positive, low positive, high negative, and low negative. Derkach et al. [2] identify student learning states by recognizing head postures. Psaltis et al. [3] determine students' learning focus in teaching activities based on eye movements. Liu et al. [4] use wearable brainwave detection devices for attention evaluation, but the devices are specialized and cost-intensive. This paper builds a classroom attention detection model based on computer vision processing technology and machine learning algorithms to detect students' classroom attention, mainly selecting eye, mouth, and facial expression detection to comprehensively evaluate students' learning focus.

[1] Corresponding author: Song HE, Air Force Early Warning Academy; e-mail: 339446624@qq.com

2. System Design.

This system integrates various technologies such as image acquisition, image processing, and attention detection algorithms. It proposes a method of attention detection based on multi-feature analysis. After the system starts, it calls external image capture devices to monitor students' facial images in real time. It captures facial graphics through video frames and extracts features from human face images through grayscale conversion and Histogram of Oriented Gradient (HOG) feature extraction algorithms. Key points of eyes and mouth are extracted, and the distance relationship between key points is calculated. The system uses Eye Aspect Ratio (EAR) and Mouth Aspect Ratio (MAR) to identify the state of eyes opening and closing and yawning of students, thus determining whether they are in a focused state or not. Convolutional neural networks are used to detect and classify facial expressions. Through comprehensive judgment of the above three situations, if the judgment is that the student is not in a focused state, an alarm will be triggered, and if they are in a focused state, the detection will continue.

3. Feature Point Extraction

This paper uses the Gradient Boosting Decision Tree (GBDT) algorithm [5] to select 68 facial feature points as the raw material for face concentration detection. Input image data with annotated key points of the face is first extracted and processed. Since the size of the face varies, affine transformation is used to affinely transform the facial key points to a unit space to unify the size and coordinate system. Then randomly sampling pixels within the initial key point range are used as corresponding feature pixels to match each local key point. Finally, the average face shape model is used to correct the matching results and iterate until convergence. The eye feature points is No.37 to No.48, as shown in Figure 1.

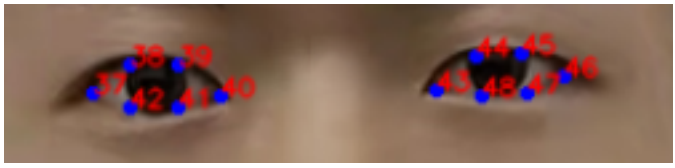

Figure 1. Eye Feature Points

The mouth feature points is No.49 to No.68, as shown in Figure 2.

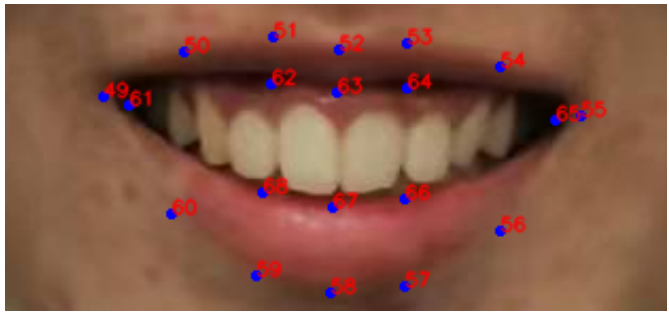

Figure 2. Mouth Feature Points

4. Detection of Eyes Feature

According to the medical research, the human eye blinks more than ten thousand times a day. Under normal circumstances, a person's eyes blink 15 times per minute, with an average of 4-5 seconds between each blink and a closure duration ranging from 100ms to 400ms. However, the degree of eye closure can vary with physiological state changes. When concentrating, the number of blinks decreases, the degree of eye closure is smaller, fatigue increases the frequency of blinking, the interval between blinks becomes shorter, and the duration of eye closure also increases. Therefore, it can be used to collect data on the degree of eye closure and closure time of the tested person as a feature indicator for concentration.

The key to collecting data on closed-eye behavior lies in how to observe and discriminate closed-eye behavior and use it as a basis for determining concentration. Using aspect ratio (aspect ratio) can solve this problem and improve the accuracy of discrimination while reducing errors caused by environmental, equipment, distance, and other elements differences. To further improve the accuracy of calculation results, this design introduces the Eye Aspect Ratio (EAR) as a concentration determination standard and uses the distance between feature points to calculate the eye's aspect ratio, improving both the efficiency and accuracy of the algorithm. The calculation method of eye aspect ratio is shown in Figure 3.

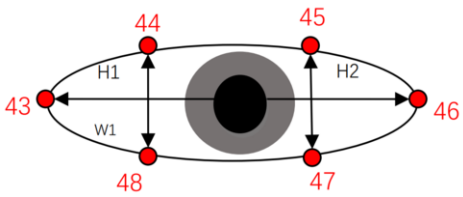

Figure 3. Mouth Feature Points and calculation

To improve the accuracy of calculation, we need to calculate the Euclidean distance between each feature point separately. Let eye0-eye5 be No.43 to No.48. Therefore, the EAR calculation formula is shown in Equation (1) as follows:

$$EAR=\frac{\|eye1-eye5\|+\|eye2-eye4\|}{2\|eye0-eye3\|} \tag{1}$$

In the formula, $\|$ eye1-eye5 $\|$ represents the upper and lower eyelid distance H1 calculated by feature points 44 and 48, $\|$ eye2-eye4 $\|$ represents the upper and lower eyelid distance H2 calculated by feature points 45 and 47, both of which are the upper and lower eyelid distances. The meaning of adding them together and dividing by two is to take the average, reducing errors. $\|$ eye0-eye3 $\|$ represents the distance between the left and right corners of the eyes calculated by feature points 43 and 46.

As the eyes close, the EAR value gradually decreases. When the EAR value is lower than the detection threshold, which is denoted as etv (eyes threshold value), the eye is considered to be closed. For a given frame in a video stream, the eye closure detection condition can be determined as shown in formula (2).

$$\text{The state of the eyes} = \begin{cases} \text{Closed} & EAR < etv \\ \text{Normal} & EAR > etv \end{cases} \tag{2}$$

According to the experimental results, when the eye threshold etv is 0.2, the effect is the best. Therefore, 0.2 is used as the eye threshold in this paper.

When the eye closure state of the test subject is determined, the video frame counting begins. The total number of frames counted is recorded as a counter. When the counter value exceeds the system trigger value (cf1), an alarm mechanism is triggered. The system will emit an audible alarm and display an alarm message on the page. This dual alarm provides a warning to the test subject to prevent them from falling asleep during the test. The trigger value can be set according to the needs. In this system, the trigger value is set at 48, which means that the alarm will be triggered when the counter value is greater than 48. If you want to enhance the sensitivity of the system, you can reduce the trigger value and shorten the time for frame counting. Otherwise, vice versa.

5. Detection of Mouth Feature

Yawning is one of the typical characteristics of lack of concentration. Therefore, the yawning behavior of the subject can be used as an important indicator of concentration judgment. Yawning is a case where the mouth is open to a certain extent. Similar to the closed-eye judgment, this paper calculates the mouth aspect ratio MAR (Mouse Aspect Ratio) as a measure of the degree of mouth opening. The aspect ratio of the mouth is calculated as shown in Figure 4.

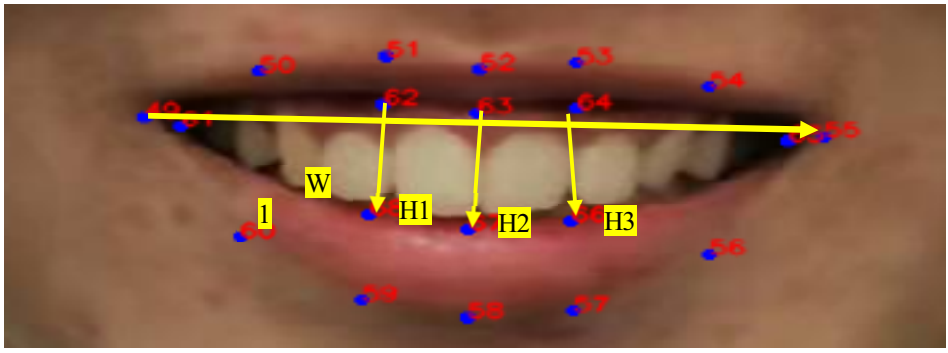

Figure 4. Mouth feature points and calculation

Suppose 49-68 points are m0-m19, then the MAR calculation is shown in formula (3)

$$MAR=\frac{\|m13\text{-}m19\|+\|m14\text{-}m18\|+\|m15\text{-}m17\|}{3\|m0\text{-}m6\|} \tag{3}$$

In the formula, $\|$ m13-m19 $\|$ represents the upper and lower lip distance H1 calculated by feature points 62 and 68, $\|$ m14-m18 $\|$ represents the distance between the upper and lower lips calculated at feature points 63 and 67, $\|$ m15-m17 $\|$ represents the distance between the upper and lower lips calculated at feature points 64 and 66, The three distances are all taken to be the distance between the upper and lower eyelids, and the sum of these distances is divided by three to calculate an average value. This helps to reduce errors in the measurement. $\|$ m0-m6 $\|$ represents the distance between the left and right corners of the mouth calculated by the feature points 49 and 55

The detection threshold for yawning is called the "mouth threshold value" (mtv). When the MAR exceeds the MTV, a preliminary determination of yawning is made and frame counting begins. Research has shown that a yawn lasts approximately 6 seconds. With this characteristic, by counting frames or time accumulation, significant normal changes in the mouth such as speaking loudly can be distinguished from real yawning, further improving the accuracy of the determination. When the cumulative frame count COUNTER2 reaches the trigger value (cf2), it is determined that yawning has occurred, triggering an alarm.

The judgment of yawn is shown in formula (4).

$$\text{The state of the Mouth} = \begin{cases} \text{Yawn} & \text{MAR}>\text{mtv} \cap \text{COUNTER2}>\text{cf2} \\ \text{Normal} & \text{MAR}<\text{mtv} \end{cases} \quad (4)$$

According to the experimental results, the accuracy is the highest when the mouth threshold mtv is 0.45. Therefore, the mouth threshold in this paper adopts the optimal threshold of 0.45.

6. Detection of Facial Expression

Analyzing students' facial expressions can help teachers understand their listening status and make timely adjustments to the classroom mode. In 1971, American scholars Ekman and others conducted extensive facial expression experiments to divide human faces into six basic emotions, including happiness, surprise, fear, sadness, disgust, and anger. Students' emotional states during class can be categorized as positive, negative, or neutral emotions. Positive emotions include happiness and surprise. When students show positive emotions, it indicates that they are in a state of willingness to accept what is being taught and are currently thinking actively. Negative emotions include sadness, anger, fear, and disgust, indicating that students are not interested in what the teacher is teaching or are not paying attention during class, which suggests poor concentration. Neutral emotions suggest that students' concentration levels are average.

By using cameras to collect real-time facial images, facial feature extraction modules extract facial features and classify them according to emotion categories, ultimately outputting the classification results. The training dataset includes two parts: general face detection data set and actual collection data set. The general face detection data set selects FER2013, JAFFE, and CK+, while the actual collection data set is collected by cameras. Each image contains facial information and its corresponding emotional information, with seven emotions: anger, disgust, fear, happiness, sadness, surprise, and normal.

7. Experimental Results

In order to detect the overall function of the system and verify the advantages of multi-feature fusion detection method, the accuracy of multi-feature and single-feature detection is compared. A total of 4000 pictures were tested, including 2000 focused samples and 2000 unfocused samples. Among them, the accurate number of samples detected by single feature of mouth was 2886, and the corresponding accuracy rate was 72.15 %. The accurate number of samples detected by single feature of eye was 3225,

and the corresponding accuracy rate was 80.63 %. The accurate number of samples detected by single feature of expression was 3340, and the corresponding accuracy rate was 83.5 %. The accurate number of samples detected by fusion of two-dimensional features of eye and mouth was 3465, and the corresponding accuracy rate was 86.63 %. The accurate number of samples detected by fusion of three-dimensional features of eye, mouth and expression was 3641. The corresponding accuracy was 91.03 %. The accuracy results of each model are shown in Table 1.

Table 1 Comparison of attention detection performance

Feature Type	accuracy
multi-feature	91.03%
eye features and mouth features	86.63%
expression feature	83.50%
eye features	80.63%
mouth features	72.15%

According to the results of table 1, the multi-feature fusion model has the highest accuracy rate, reaching 91.03 %. The accuracy rate is 4.4 % higher than that of eye and mouth features, 7.53 % higher than that of eye single features, and 18.88 % higher than that of mouth single features. Therefore, the multi-feature fusion model has the best detection performance.

References

[1] Gupta S K, Ashwin T S, Guddeti R M R. Students' affective content analysis in smart classroom environment using deep learning techniques[J]. Multimedia Tools and Applications, 2019, 78(18): 25321-25348.

[2] Derkach D, Ruiz A, Sukno F M. Tensor Decomposition and Non-linear Manifold Modeling for 3D Head Pose Estimation[J]. International Journal of Computer Vision, 2019, 127(10): 1565-1585.

[3] Psaltis A, Apostolakis K C, Dimitropoulos K, et al. Multimodal Student Engagement Recognition in Prosocial Games[J]. Ieee Transactions on Games, 2018, 10(3): 292-303.

[4] Liu Xi, Zhan Zengrong. Wearable brain wave concentration evaluation system [J].Computer programming skills and maintenance, 2021(12): 139-140+170.

[5] De Kegel Alexandra,Baetens Tina,Peersman Wim,Maes Leen,Dhooge Ingeborg,Van Waelvelde Hilde.Ghent developmental balance test: a new tool to evaluate balance performance in toddlers and preschool children[J]. Physical therapy.2012, 92(6) : 841-852.