Research on student classroom attention recognition based on multimodality

Lu Lin*, Lili Shi
Department of Software Engineering, School of Informatics, Xiamen University, Xiamen, Fujian, China

ABSTRACT

With the continuous improvement of teaching contents and methods, the concept of students as the main body of teaching has been deepened. Efficient and accurate identification of students' classroom concentration has become an effective means to improve classroom efficiency and is also a focus of attention in the field of education. In this paper, we propose a concentration recognition model based on the multimodal fusion of students' head posture and facial expressions and validate and experiment on a self-built concentration database. The experimental results demonstrate that non-negative emotions are beneficial to concentration, and the student's head orientation can reflect the student's concentration more clearly. The multimodal fusion method performs better than the unimodal method. This will reduce the workload of manual data collection and effectively improve the accuracy of data collection. It also strengthens the home-school association and has implications for the development of educational theory and practice.

Keywords: Attention, head pose estimation, emotion recognition

1. INTRODUCTION

Attention is a generally accepted indicator of students' classroom efficiency and learning gains[1]. To improve the quality of classroom teaching and promote students' overall development, it is necessary to change the single assessment method in traditional classroom teaching and pay attention to students' classroom concentration. However, in traditional teaching, there are problems such as the inability to accommodate all students, single evaluation criteria, and lagging feedback. Therefore, how identifying and quantify students' learning status is an urgent problem to be solved.

Montagnini et al.[2] used students' grades as a basis to analyze the efficiency of students' learning and the effort they put in learning by analyzing their academic performance. Chen et al.[3] concluded that the persistence of female students' classroom attention was higher than that of male students by observation method. However, these methods have disadvantages such as high subjectivity, high labor cost, and inability to be replicated on a large scale.

In terms of studying the link between physiological signals and classroom concentration, Zhang et al.[4] used a wearable device to read a series of physiological signals including the visual focus of students and analyzed their classroom attention. Other researchers have analyzed students' classroom attention changes, their patterns, and related influences through eye-tracking[5]. However, such methods rely on high-cost data collection equipment and are based on a single judgment.

In this paper, we combine computer vision technology with the field of education and propose a multimodal approach to determine the concentration level of students' classroom attention. The method collects real classroom data through a camera, recognizes and analyzes students' head posture and facial expressions respectively, and then fuses the two behavioral patterns to determine students' concentration level. We established a database of students' classroom concentration to quantify, analyse, and visually present classroom concentration. On the one hand, the automatic collection of classroom data reduces human and material resources. On the other hand, we integrated multimodal information, conducted an exploration of the connection between classroom performance and classroom attention, and verified the validity of the findings on the self-built student classroom concentration database.

* 953155791@qq.com

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125063D
 · 0277-786X · doi: 10.1117/12.2662889

2. MATERIALS AND METHODS

2.1 Database

We built a student classroom concentration database by considering the meaning behind students' behavior from their real learning status. The source data were obtained from a real classroom in a domestic university and an online classroom during the epidemic.

For the offline classroom, the recording period was three months, with 48 sessions of 45 minutes each. Two cameras (Sony alpha 7 III) were used for source data collection, one at the instructor's podium and the other in the left front corner of the classroom. The camera resolution was 1440 × 1080 and the frame rate was 25 fps. For the online classroom, 40 college student volunteers, 16 female, and 24 male were recruited for this study. The volunteers used the webcam on their computers to make video recordings. Each recording should last no less than 10 minutes.

After data collection and screening, 11,782 valid images with clear images and little noise were selected from them for manual concentration labeling. According to the Likert scale, this paper divided students' classroom concentration into five levels, and each level was divided into 5 points. Through multiple labeling, the plural was taken as the concentration level and the mean as the concentration score. A total of thirty people performed the data marking. If the difference in labels was too large, it was handed over to a teacher experienced in teaching to make a decision.

2.2 Overview

Xu et al.[6] concluded that in the case of severe head deflection, the attention of the experimenter was significantly reduced. Meanwhile, a study proved that emotions play an important role in the cognition of individuals[7]. Therefore, in this paper, head posture and facial expressions are used as evaluation indicators, and two behavioral patterns are combined to implement a multimodal concentration recognition and analysis system.

The system is mainly divided into a head posture module, facial expression recognition module, and concentration calculation module. The system first captures classroom video through a camera and crops the target object. The target video is recognized frame by frame for head posture and the concentration of the target student is predicted by regression analysis. If the score is lower than the threshold, it is judged as not focused. If the score is higher than the threshold, it enters the expression module and performs expression recognition and concentration judgment on the target face. Combining the two, the final concentration level of the student is output.

The workflow is shown in Figure 1.

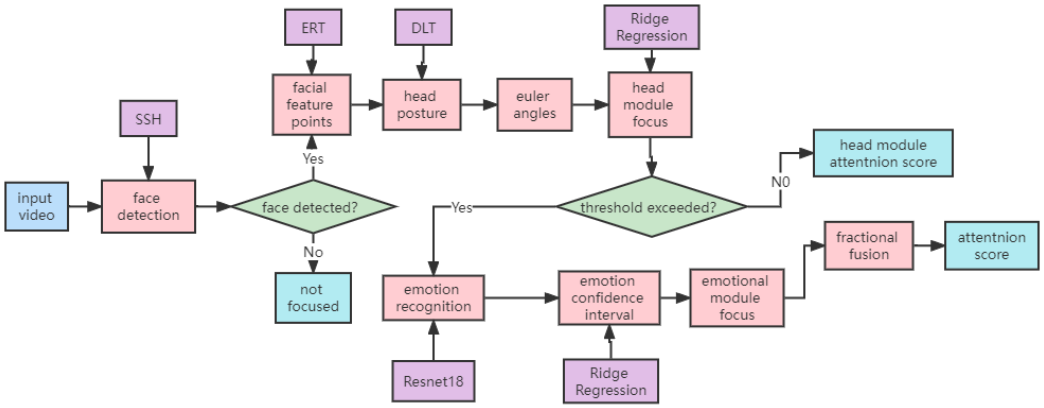

Figure 1. Concentration model workflow.

2.3 Euler angles for head posture

In this study, a model-based approach is used to estimate the user's head pose based on the geometric relationships or feature points of the object. For head rotation, the Euler angles are used to represent the motion of the object, which are pitch, roll, and yaw angles. The basic idea is to rotate the standard head model to a certain angle until the 3D feature points on the 2D projected image synthesized by the standard model coincide with the feature points on the real image.

To accommodate multi-scale faces in the classroom, SSH[8] is used as the algorithm for the face detection module in this paper. The detected face feature maps are then passed to the face feature point capture module as parameters. We use the ERT[9] algorithm to capture 68 feature points of the face for determining the target motion. The method regresses the face shape from the current shape to the true shape step by step by building GBDT.

In this module, there are three coordinate systems. They are the world coordinate system, the camera coordinate system, and the image coordinate system. The face feature points of the real face are located in the image coordinate system, and the feature points of the standard model of the face are located in the world coordinate system. When the rotation translation vector of the object is obtained, the points in the world coordinates can be converted to the points in the camera coordinates. The points in the camera coordinates are then converted to the image coordinate system using DLT by the inherent parameters of the camera, such as focal length and optical center.

The basic principle is shown in Figure 2.

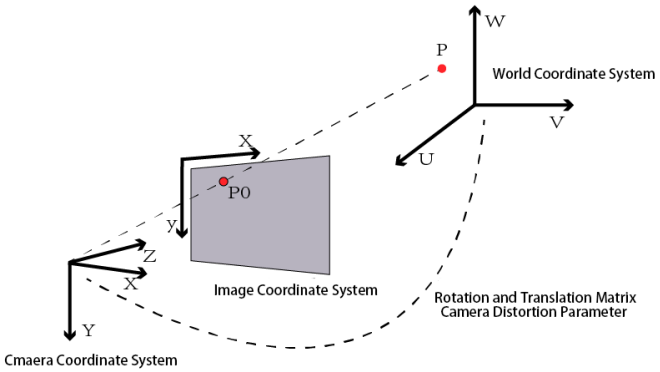

Figure 2. Fundamentals.

The resulting example of a student's head posture is shown in Figure 3.

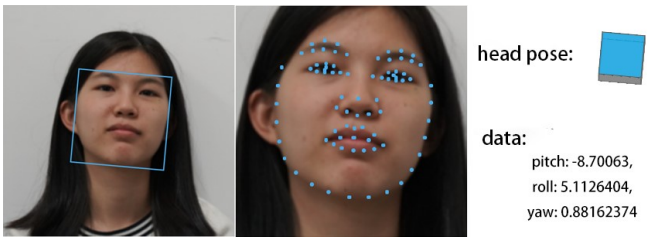

Figure 3. An example of head posture.

2.4 Facial expression recognition

Psychologists have classified human expressions into seven categories through cross-cultural studies. They are anger, fear, disgust, happiness, sadness, surprise, and neural. The module will first preprocess the detected face images for size, grayscale normalization, and image segmentation. Next, Resnet18[10] is used to complete the recognition and classification of the seven expressions. A dropout[11] strategy is added before the fully connected layer to increase the model robustness. We remove multiple fully connected layers in traditional Resnet18 and classify directly after one fully connected layer and obtain the confidence level of the seven emotions of the target object.

A final example of the emotions obtained in Figure 4.

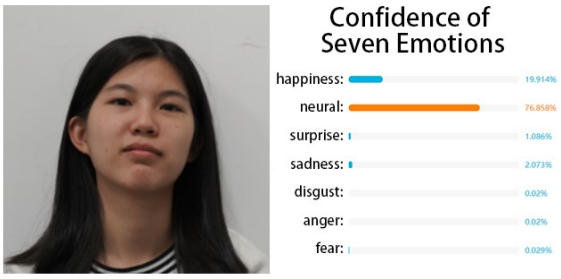

Figure 4. An emotional example.

2.5 Attention prediction

To avoid multicollinearity, this study uses Ridge regression[12] to model and predict students' classroom attention with three angles of head pose and seven emotional confidences as input features. In this paper, the fusion is based on the multimodal information score layer. Identify the characteristics of each modal of the student's classroom and perform regression, and then perform the final fusion of the scores obtained from the characteristics of each modal to obtain the final result.

3. RESULTS AND DISCUSSION

3.1 Univariate correlation analysis

We define 10 factors as independent variables. They are pitch yaw angle, fear, disgust, happiness, sadness, surprise and neural. We use the following formula to measure the linear correlation *r* between a single independent variable and attention.

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{S_X}\right)\left(\frac{Y_i-\bar{Y}}{S_Y}\right) \quad (1)$$

where *n* is the sample size, and $\bar{X}$ and *Y* are the sample mean values, respectively. *SX* and *SY* are the sample standard deviations, respectively. The larger the absolute value of *r*, the stronger the correlation.

After removing the invalid data that cannot be recognized by the face, the correlation coefficient between each independent variable and the degree of attention is shown in Table 1.

Table 1. The correlation coefficient.

Emotion	Correlation coefficient
Anger	-0.0565
Disgust	-0.1068
Fear	-0.0319
Happiness	0.0576
Neural	0.1424
Sadness	-0.0404
Surprise	**-0.0148**
Pitch	**0.3102**
Yaw	0.1139
Roll	0.2129

The data shows that the correlation between head posture and concentration is significantly higher than that of emotion. It can be inferred that the head orientation can better reflect the concentration of students in the classroom. The relationship between emotion and concentration is very weak. The yaw angle has a very weak correlation with concentration, the roll angle has a weak correlation, and the pitch angle has a moderate correlation. The pitch angle is most strongly correlated with concentration. It can be concluded that the situation of students bowing their heads and raising their heads is an important criterion for judging their attention.

Among the seven emotions, happy emotions and calm emotions in non-negative emotions were significantly positively correlated with students' concentration. Anger, depression, fear, sadness, and surprise in negative emotions were significantly negatively correlated with students' concentration. Therefore, improving students' emotional experience of learning is an effective way to improve classroom attention.

3.2 Model performance comparison

To verify the accuracy and applicability of the regression equation, this study conducted experiments on the data of several volunteers selected from a self-built database. The results are shown in Table 2. The three model structures are a separate head pose prediction model, a separate facial expression prediction model, and a multimodal prediction model. The root means square error is used as the evaluation index of the model.

As can be seen from the table, the basic performance of HPR is better than that of FER, and the performance of CR is the best. From this, it can also be inferred that head posture accounts for a larger proportion of the factors affecting students' concentration. Especially for volunteers 12 and 15, the FER performance dropped sharply. Tracing the original data, the confidence levels of the volunteer's seven emotions were similar, which led to the abnormal situation. The root means square error of different experimenters varies greatly, which proves that students have independent habitual movements and postures, and individual differences are obvious[13]. In conclusion, the multimodal attention recognition model has better performance and robustness than a single modality.

Table 2. A slightly more complex table with a narrow caption.

Volunteer number	Head posture model (HPR)	Facial expression model (FER)	Combined model (CR)
1	3.176	2.696	2.843
2	1.213	1.216	1.180
3	1.312	1.299	1.288
4	1.483	1.806	1.515
5	2.768	2.600	2.581
6	1.020	2.640	1.078
7	1.018	2.125	2.437
8	2.610	2.607	2.538
9	6.190	8.651	5.828
10	0.393	0.385	0.333
11	1.931	2.135	1.844
12	**2.240**	**5.629**	**2.051**
13	3.431	4.199	3.318
14	1.714	2.773	1.250
15	**1.667**	**12.136**	**1.500**
16	1.865	1.767	1.538

4. CONCLUSION

This study constructs a database of students' classroom concentration, proposes a simple and convenient attention identification method, and explores the relationship between students' classroom performance and attention. The method obtains features based on the head pose and facial expression and combines the two modalities for ridge regression modeling. The verification and experiments are carried out on the self-built database, and the experimental results are analyzed and demonstrated. It solves the problem that teachers cannot take care of students in traditional classrooms and helps students understand their learning status. It also reduces labor costs for classroom data collection. To sum up, it has theoretical and practical significance for promoting the development of education.

ACKNOWLEDGMENTS

This work was supported by the National College Student Innovation and Entrepreneurship Training Program of China (202110384258) and The Social Science Program of Fujian Province (FJ2020B062).

REFERENCES

[1] Li, X. and Yang, X., "Effects of learning styles and interest on concentration and achievement of students in mobile learning," Journal of Educational Computing Research, 54(7), 922-945(2016).
[2] Montagnini, A. and Castet, E., "Spatiotemporal dynamics of visual attention during saccade preparation: Independence and coupling between attention and movement planning," Journal of Vision, 7(14), 8(2007).
[3] Chen, C. M. and Wang, J. Y., "Effects of online synchronous instruction with an attention monitoring and alarm mechanism on sustained attention and learning performance," Interactive Learning Environments, 26(4), 427-443(2018).
[4] Zhang, X., Wu, C. W. and Fournier-Viger, P., "Analyzing students' attention in class using wearable devices," IEEE 18th Inter. Symp. on a World of Wireless, Mobile and Multimedia Networks, 1-9(2017).
[5] Rosengrant, D., Hearrington, D. and O'Brien, J., "Investigating student sustained attention in a guided inquiry lecture course using an eye tracker," Educational Psychology Review, 33(1), 11-26(2021).
[6] Xu, X. and Teng, X., "Classroom attention analysis based on multiple euler angles constraint and head pose estimation," Inter. Conf. on Multimedia Modeling, 329-340(2020).
[7] Vanderlind, W. M., Millgram, Y. and Baskin-Sommers, A. R., "Understanding positive emotion deficits in depression: From emotion preferences to emotion regulation," Clinical Psychology Review, 76, 101826(2020).
[8] Najibi, M., Samangouei, P. and Chellappa, R., "SSH: Single stage headless face detector," Proceedings of the IEEE Inter. Conf. on Computer Vision, 4875-4884(2017).
[9] Kazemi, V. and Sullivan, J., "One millisecond face alignment with an ensemble of regression trees," Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, 1867-1874(2014).
[10] Hao, Z., Ren, F. and Kang, X., "Classification of steel strip surface defects based on optimized ResNet18," 2021 IEEE Inter. Conf. on Agents (ICA), 61-62(2021).
[11] Poernomo, A. and Kang, D. K., "Biased dropout and crossmap dropout: Learning towards effective dropout regularization in convolutional neural network," Neural Networks, 104, 60-67(2018).
[12] Ndabashinze, B. and Üstündağ Şiray, G., "Comparing ordinary ridge and generalized ridge regression results obtained using genetic algorithms for ridge parameter selection," Communications in Statistics-Simulation and Computation, 1-11(2020).
[13] Raca, M. and Dillenbourg, P., "System for assessing classroom attention," Proc. of the Third Inter. Conf. on Learning Analytics and Knowledge, 265-269(2013).