DOI: 10.53469/jerp.2025.07(08).12

Comparison and Analysis of Student Behavior Recognition Methods in University Classrooms

Liqiong Lu[1,5], Lingyue Hu[2], Qin Lei[4], Dong Wu[1,5], Yongheng Chen[1], Tonglai Liu[3,*]

[1]School of Computer Science and Intelligence Education, Lingnan Normal University, Zhanjiang 524048, China
[2]College of Big Data and Computer Science, Guangdong Baiyun University, Guangzhou 510450, China
[3]College of Information Science and Technology, Zhongkai University of Agriculture and Engineering, Guangzhou, 510225, China
[4]School of Mathematics and Statistics, Lingnan Normal University, Zhanjiang 524048, China
[5]Guangdong Provincial Key Laboratory of Development and Education for Special Needs Children, Lingnan Normal University, Zhanjiang 524048, China
Correspondence Author, tonglailiu@zhku.edu.cn

Abstract: ***Student behavior intelligence recognition is a very important component of intelligent evaluation in university classrooms. Firstly, 1000 images containing various types of student behaviors in university classrooms were collected to construct a student behavior recognition dataset. These student behaviors include listen, read, write, bow, lie, yawn, drink, play, glance and trick. Then, classic CNN object recognition methods including SSD, Faster RCNN, YOLOV8, YOLOV11 and YOLOV12 were used to recognize student behaviors in university classrooms on the above dataset and the recognition performance of different methods was compared and analyzed. The experimental results show that YOLOV12 has the best student behavior recognition performance, with an mAP value of 0.521 and YOLO V11 has the fastest inference speed and the second highest recognition performance with an mAP value of 0.519.***

Keywords: Student behavior recognition, University Classroom, CNN.

1. Introduction

Classroom student behavior is the most fundamental and important component of teaching activities. Analyzing it can not only help teachers understand students' listening and learning status, thus forming effective student evaluations, but also help teachers analyze teaching effectiveness from the perspective of students' listening, thereby improving subsequent teaching methods. In traditional classrooms, teachers understand students' classroom state by observing their classroom behavior and interacting with the teacher. This type of method is time-consuming and laborious, and it is difficult for teachers to observe the learning status of all students, making it difficult to accurately evaluate the classroom effect. With the sweeping of artificial intelligence technology throughout society, the education industry is inevitably facing various reforms under the impact of new technologies. Among them, the intelligence of teaching and management has become a key focus of education reform. How to use computer vision technology to intelligently analyze classroom behavior, help teachers automatically evaluate teaching effectiveness and improve teaching quality is one of the urgent technical fields that need to be studied at present.

Student classroom behavior recognition has become a very important part of intelligent education. In recent years, more and more scholars have started researching in this field. From a technical perspective, student classroom behavior recognition methods can be divided into two categories: traditional machine learning based classroom behavior recognition methods and deep learning based classroom behavior recognition methods. traditional machine learning based methods usually manually extract features from images and then combine them with machine learning methods for classroom behavior recognition. Wu et al. [1] combined multi-scale HOG features, human skeleton information of images with SVM methods for recognizing classroom behavior. Altuwair et al. [2] proposed a automatic multi-modal approach to extract composite engagement value feature including key frame feature, emotion feature, mouse and keyboard feature, then used this composite engagement value feature as an input to the Naive Bayes (NB) classifier to analyze three modalities representing students' behaviors: emotions from facial expressions, keyboard keystrokes, and mouse movements. Wei et al. [3] used the Scale-Invariant Feature Transform (SIFT) descriptor combined with the Local Log Euclidean Multivariate Gaussian (L2 EMG) descriptor to extract richer local features and then the extracted features were fed into fuzzy BLS for recognizing student behaviors. Tradition machine learning based methods have some limitations such as manual feature design and selection, sensitive to variations and noise in the data and demanding a large amount of labeled data for training.

In recent years, the most prominent technology in deep learning, Convolution Neural Networks (CNNs) have been widely used in various fields of computer vision and have achieved remarkable results. Deep learning based classroom behavior recognition methods always used CNN as the basic technology. Cheng et al. [4] proposed a Deep Convolutional Generative Adversarial Network for Student Action Recognition (DCGANSAR) containing two stages: constructing the Deep Convolutional Generative Adversarial Network (DCGAN) to obtain pre-trained weights in the discriminator, and using the discriminator of DCGAN to classify actions. Zhao et al. [5] used Efficient Transformer Block (ETB) improved the capability to recognize occluded students and utilized Efficient Convolution Aggregation Block (ECAB) to improve the accuracy of student behavior recognition. Zhang et al. [6] proposed a method to detect hand-raising of students. In this method, they designed Spatial Context Augmentation (SCA) to mitigate feature map information loss at the highest level and Multi-Branch Dilated Convolution (MBDC) to enlarge the receptive field and reduce false detection. Student behavior recognition is closely

related to facial expressions, body posture, and other factors, leading to the emergence of methods that combine body posture and facial expressions with student behavior recognition. Abdallah et al. [8] pre-trained the model on a facial expression dataset. Then, the trained model was transfered to classify students' behavior. Lin et al. [9] used the open pose framework to collect skeleton data and then feature extraction was performed to generate feature vectors that represent human postures to recognize student behaviors. The classic object recognition methods, especially the YOLO series algorithms, have injected new vitality into the field of student behavior analysis. In recent years, many methods have emerged that combine student behavior characteristics with classic object recognition methods. Based on YOLO V3[10], YOLO V5[11], YOLO V7[12] and YOLO V8 [13], many classroom behavior recognition methods [14-17] were proposed too.

In this paper, firstly, a dataset of classroom student behaviour is constructed, which contains 1000 images and identifies 10 behaviors of students. These images are all obtained from real university classrooms. Using these images as the data set for recognizing classroom student behaviour has great practical and guiding significance. Considering that student behavior involves student portrait privacy, this dataset cannot be made public. Subsequently, based on the deep learning platform, the general methods in the field of object recognition are realized, including SSD [19], Faster RCNN [20], YOLOV8 [13], YOLO V11[21] and YOLO V12[22]. These methods are applied to the constructed student behavior recognition dataset, and the recognition performance is compared and analyzed. The experimental results show that YOLO V11 and YOLO V12 have better recognition performance, with a score of 0.519 and 0.521 in *mAP*.

2. Classroom Student Behavior Recognition Dataset

A dataset of classroom student behavior recognition was constructed obtained from the real university classrooms. This dataset consists of a total of 1000 images, with 800 in the training set and 200 in the testing set. There are 10 types of student behavior including listen (listening to lectures), read, write, bow (bowing their heads), sleep, yawn, drink (drinking water), play (playing with their phones), glance (looking left and right), and trick (engaging in behaviors unrelated to learning). Considering that student behavior involves student portrait privacy, this classroom student behavior dataset cannot be made public. In the images of this dataset, student seats are very densely packed, with larger target positions for students in the front row and smaller target positions for students in the back row. Additionally, there are situations where students in the front and back rows obstruct each other, which poses significant challenges for recognizing classroom student behavior.

Two images of real university classroom were selected randomly as shown in Figure 1. There are all kinds of student behavior including listen, write, bow, sleep, glance and other behaviors in Figure 1. The behavior of each student constantly changes throughout the entire class, and the students in the front and back rows occupy different spaces in the image, with occasional occlusions.

Figure 1: Sample images of university classroom student behavior

The behavior of students in university classrooms is not evenly classified. Generally speaking, there are more situations such as listening, reading, and bowing, while other types are relatively few. The total number of each type of student behavior have been listed in Figure 2, which shows that among 1000 images, the number of listen was the highest, reaching 4580, followed by the number of bow. The sample types of other student behaviors were relatively fewer, with the lowest being the number of drink and play, both less than 120. The uneven sample size also poses certain difficulties for student behavior recognition, especially for those with fewer samples. Therefore, when designing student behavior recognition algorithms, it is necessary to accurately mine the characteristics of different student behaviors in order to cope with the imbalance in sample size.

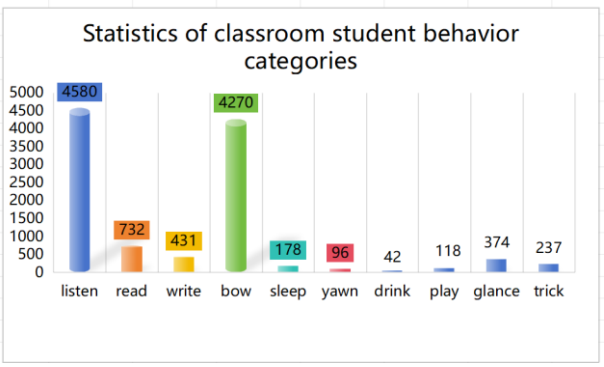

Figure 2: Statistics of classroom student behavior categories in all images

3. Comparison and Analysis of Classroom Student Behaviors Based on CNN

3.1 Experimental Set

In this paper, SSD and Faster RCNN methods are implemented based on Tensorflow 2.4, YOLO V8 and YOLO V11 methods are implemented based on Pytorch 1.9 and YOLO V12 are implemented based on Pytorch2.2. The above methods are all trained and tested using Geforce 3060 graphics card. Because student behavior recognition includes small and large target detection in the image, this paper resets the size of the default anchors when implementing SSD and Faster RCNN methods. For the size of input image being (300,300), the anchor_size is set to [21, 45, 99, 153, 207, 261, 315] in SSD and [32, 256, 512] in Faster RCNN. Other parameters in the experiment process, such as learning rate, are consistent with the initial setting of the method, and the value of bath size is adjusted according to the size of GPU memory. YOLO V8, YOLO V11 and YOLO V12 use yolo8m.pt, yolo11m.pt and yolo12m.pt to initialize the parameters respectively.

3.2 Performance Evaluation Indicators for Classroom Student Behavior Recognition

Accuracy (Precision, *P*), regression rate (Recall, *R*) and comprehensive indicators mean average precision (*mAP*) are used to evaluate the recognition performance of classroom student behaviors.

$$P = \frac{TP}{FP+TP};$$

$$R = \frac{TP}{TP+FN};$$

Among them, TP (true positive) represents the number of correctly recognized samples; FP (false positive) represents the number of error recognized samples; FN (false negative) represents the number of undetected samples. Here, when the IOU between the recognition box and the ground truth box exceeds 50%, it is considered correct recognition. *mAP* is the average of the average precision values for all categories, calculated based on the precision recall curve.

3.3 Comparative Analysis of Recognition Methods of Classroom Student Behaviors

Table 1 lists the recognition results of different methods on the dataset of classroom student behavior constructed in this paper. From the experimental results listed in Table 1, YOLO V12 has the highest value, reaching 0.521, followed by YOLO V11 and YOLO V8 with a *mAP* value of 0.519 and 0.484, followed by SSD and Faster RCNN. It is obviously, YOLO series methods have significantly better recognition performance than SSD and Faster RCNN. Moreover, the values of Faster RCNN are almost less than half of those of YOLO V11 and YOLO V12, which shows that YOLO V11 and YOLO V12 are more suitable for the recognition of classroom student behaviors. In addition, from the accuracy and regression rate points of view, YOLO V12 has the highest accuracy and regression rate, having absolute recognition performance advantage. From the perspective of running speed, YOLO V11 can complete 1.23 iterations per second, while YOLO V12 only completes 0.44 iterations per second. Obviously, YOLO V11 has better training and inference speed than YOLO V12.

Table 1: Comparison of recognition performance

Method	*P*	*R*	*mAP*	*Speed(it/s)*
SSD	0.448	0.306	0.352	-
Faster CNN	0.336	0.398	0.215	-
YOLO V8	0.459	0.535	0.484	1.21
YOLO V11	0.477	0.556	0.519	**1.23**
YOLO V12	**0.496**	**0.577**	**0.521**	0.44

Table 2 lists the detailed AP values for each student behavior category. It is obviously, YOLO series methods obtained better performance in each type of student behavior. YOLO V12 has the best performance in *AP-read*, *AP-sleep* and *AP-glance*. YOLO V11 has the best performance in *AP-listen*, *AP-write*, *AP-bow* and *AP-yawn*. YOLO V8 has the best performance in *AP-drink*, *AP-play* and *AP-trick*. Especially for the recognition of students' listen behavior, which accounts for the largest proportion in the dataset, the *AP* values of the YOLO series methods have all reached 0.9 or above.

YOLO V11 and YOLO V12 have good performance in recognizing classroom student behaviors. The reason is that the YOLO series methods are superior in design, mainly reflected in the following aspects. (1) YOLOv11 introduces several fundamental improvements aimed at optimizing detection speed and accuracy. One of the most significant improvements is the shift from a pure CNN architecture to a transformer based architecture; (2) YOLOv12 completely abandons traditional CNN architecture for the first time and adopts pure attention mechanism (Vision Transformer) as the backbone network, breaking the design paradigm of YOLO series that has long relied on CNN; (3) YOLO V11 adopts an improved backbone and neck architecture to enhance feature extraction capability, and introduces C2PSA module (cross stage local self attention mechanism) to improve the efficiency of context information capture, especially enhancing the detection accuracy of small targets and complex scenes; (4) YOLOv12 introduces the region attention module, which transforms global attention into local attention, thereby reducing computational costs. At the same time, it introduces the R-ELAN (Residual Efficient Layer Aggregation Network) module, which introduces residual shortcut paths and scaling factors on the basis of the efficient layer aggregation network (ELAN), retaining feature integration capabilities while reducing computational costs and parameter/memory usage.

Table 2: Comparison of AP for each category

Method	*AP-listen*	*AP-read*	*AP-write*	*AP-bow*	*AP-sleep*	*AP-yawn*	*AP-drink*	*AP-play*	*AP-glance*	*AP-trick*	*mAP*
SSD	0.865	0.438	0.307	0.764	0.641	0.002	0.080	0.078	0.199	0.149	0.352
Faster CNN	0.752	0.234	0.044	0.612	0.282	0.001	0.001	0.003	0.101	0.090	0.215
YOLO V8	0.917	0.522	0.524	0.832	0.599	0.224	**0.232**	**0.284**	0.404	**0.305**	0.484
YOLO V11	**0.957**	0.562	**0.657**	**0.863**	0.665	**0.333**	0.169	0.253	0.487	0.282	0.519
YOLO V12	0.948	**0.694**	0.599	0.853	**0.715**	0.311	0.192	0.278	**0.520**	0.277	**0.521**

4. Conclusion

This paper focuses on the recognition of classroom student behavior. Firstly, a dataset containing 1000 images was constructed for classroom student behavior. Then, five methods including SSD, Faster RCNN, YOLO V8, YOLO V11 and YOLO V12 were used to compare and analyze the recognition performance of classroom student behavior on this dataset. The experimental results show that YOLO V12 has the best performance, with a *mAP* value of 0.521, followed by YOLO V11 with a *mAP* value of 0.519. The recognition performance of SSD and Faster RCNN is not ideal. It is worth mentioning that YOLO V11 has better training and inference speed than YOLO V12. Specifically, for classroom student behavior recognition tasks, the former's speed is about three times that of the latter. Overall, if there are high requirements for recognition performance and inference speed, YOLO V11 can be chosen as the basic method and improved; If more emphasis is placed on recognition performance, improving based on YOLO V12 is a better choice.

Based on the YOLO series methods, we plan to further analyze the characteristics of classroom student behavior, such as obstruction between students, large changes in the size of the positions occupied by students in the front and back rows, and design appropriate feature extraction mechanisms to improve the performance of classroom student behavior recognition.

Acknowledgments

The authors are grateful to the anonymous reviewers and the helpful suggestion given by the partners. The research was supported by the Guangdong province philosophy and social science planning project (no. GD24CJY21), Guangdong Basic and Applied Basic Research Foundation (no. 2023A1515011230), Heyuan Social Science and Agriculture Project (no. 2023015]), Social Sciences Research Project of the Ministry of Education (no. 24YJA860002) and Education Model Innovation Team (no. LT2404).

References

[1] Wu D, Chen J, Deng W et al. The Recognition of Teacher Behavior Based on Multimodal Information Fusion[J]. Mathematical Problems in Engineering, 2020, 2020(1):1-8.

[2] Altuwairqi, K., Jarraya, S.K., Allinjawi, A., Hammami, M. Student behavior analysis to measure engagement levels in online learning environments[J]. Signal Image Video Process. 2021, 15:1387-1395.

[3] Wei Y, Lei F, Gao J, Li X. Student action recognition based on fuzzy broad learning system[C]. International Conference on Intelligent Education and Intelligent Research (IEIR), 2022, 128–135.

[4] Cheng, Y., Dai, Z., Ji, Y., Li, S., Jia, Z., Hirota, K., and Dai, Y. Student action recognition based on deep convolutional generative adversarial network[C]. Chinese Control and Decision Conference (CCDC), 2020, 128-133.

[5] Zhao, J., Zhu, H., and Niu, L. BiTNet: A lightweight object detection network for real-time classroom behavior recognition with transformer and bi-directional pyramid network[J]. Journal of King Saud University - Computer and Information Sciences, 2023, 35(8): 101670.

[6] Zhang, G., Wang, L., Wang, L., and Chen, Z. Hand-raising gesture detection in classroom with spatial context augmentation and dilated convolution[J]. Computers & Graphics, 2023, 110:151-161.

[7] Lin F, Ngo H, Dow C et al. Student Behavior Recognition System for the Classroom Environment Based on Skeleton Pose Estimation and Person Detection. [J]. Sensors (Basel, Switzerland), 2021, 21(16).

[8] Abdallah, T. B., Elleuch, I., and Guermazi, R. Student behavior recognition in classroom using deep transfer learning with VGG-16[J]. Procedia Computer Science, 2021, 192:951-960.

[9] Buono, P., De Carolis, B., D'Errico, F., Macchiarulo, N., and Palestra, G. Assessing student engagement from facial behavior in on-line learning[J]. Multimedia Tools and Applications, 2023, 82(9):12859-12877.

[10] Joseph Redmon, Ali Farhadi. YOLOv3: An Incremental Improvement [EB/OL]. [2018], https://alumni.soe.ucsc.edu/~czczycz/src/YOLOv3.pdf.

[11] Joseph Redmon. YOLO V5[EB/OL]. [2020-05-18]. https://github.com/ultralytics/yolov5.

[12] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors [EB/OL]. [2022-07-06]. https://arxiv.org/pdf/2207.02696.

[13] Ultralytics. YOLO V8 [EB/OL]. [2023-02-03]. https://github.com/ultralytics/ultralytics.

[14] Ali, M. Y., Zhang, X. D., and Harun-Ar-Rashid, M. Student activities detection of SUST using YOLOv3 on deep learning[J]. Indonesian Journal of Electrical Engineering and Informatics (IJEEI), 2020, 8(4): 757-769.

[15] Rashmi, M., Ashwin, T. S., and Guddeti, R. M. R. Surveillance video analysis for student action recognition and localization inside computer laboratories of a smart campus[J]. Multimedia Tools and Applications. 2021, 80, 2907-2929.

[16] Wang Z, Yao J, Zeng C, et al. Learning Behavior Recognition in Smart Classroom with Multiple Students Based on YOLOv5[EB/OL]. [2023-03-20]. https://arxiv.org/pdf/2303.10916.

[17] Ma L, Zhou T,Yu B, et al. Improving YOLOv7 for Large Target Classroom Behavior Recognition of Teachers in Smart Classroom Scenarios [J]. Electronics, 2024,13(18):3726-3726.

[18] Sun J, Li S, Zhang J. Classroom Behavior Recognition and Research Based on DLKAS-YOLO8n[J]. Academic Journal of Computing & Information Science. 2024, 7(11).

[19] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[J]. 14th European Conference on Computer Vision (ECCV), 2015, (9905):21-37.

[20] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 39(6):1137-1149.

[21] Khanam R, Hussain M.YOLOv11: An Overview of the Key Architectural Enhancements[EB/OK]. [2024-10-23]. https://www.arxiv.org/pdf/2410.17725.
[22] Tian Y, Ye Q, Doermann D. YOLOv12: Attention-Centric Real-Time Object Detectors[EB/OK]. [2025-02-18]. https://arxiv.org/pdf/2502.12524.