https://doi.org/10.53964/jmer.2023015

https://www.innovationforever.com

Journal of
Modern Educational Research

ISSN 2790-3192 (Online)

Open Access

Research Article

Research on Intelligent Teaching Mode Based on Classroom Behavior Monitoring of Body Motion

Yimin Chen[1], Wanyi Li[1*], Hanrui Weng[1], Jiaying Zheng[1], Zhuoxian Qian[1], Jingmin Huang[1], Jiaqi Lun[2], Qiang Chen[1], Qian Zhang[1], Yilin Wu[1]

[1]School of Computer Science, Guangdong University of Education, Guangzhou, Guangdong Province, China

[2]Guangzhou Nansha Jinsha School, Guangzhou, Guangdong Province, China

[*]**Correspondence to: Wanyi Li, PhD, Lecturer,** School of Computer Science, Guangdong University of Education, Guangzhou, 510303, Guangdong Province, China; E-mail: luther1212@163.com

Received: September 3, 2023 **Revised:** September 18, 2023 **Accepted:** October 26, 2023 **Published:** December 31, 2023

Abstract

Objective: Classroom behavior detection is of great significance in the field of education, as it can assess students' participation and focus.

Methods: This paper introduces a novel classroom behavior detection method based on PP-YOLOv2. Leveraging computer vision and deep learning, the proposed approach involves the collection and annotation of student sample datasets, accompanied by thorough data preprocessing. The utilization of the Mish Activation function enhances the model's learning ability and improves the accuracy of behavior detection.

Results: This study is of great significance for real-time monitoring, enabling the evaluation of student behavior to improve teaching effectiveness and promote personalized learning. The experimental results show that this method exhibits good performance in classroom environments, providing educators with an efficient and accurate tool for behavior detection.

Conclusion: Further research endeavors can expand the application scope and optimize algorithms to enhance performance. In summary, the proposed method shows promise in revolutionizing classroom behavior detection, offering a more efficient and accurate means of assessing and improving students' educational experiences.

Keywords: classroom behavior detection, deep learning, PP-YOLOv2

Citation: Chen Y, Li W, Weng H, Zheng J, Qian Z, Huang J, Lun J, Chen Q, Zhang Q, Wu Y. Research on Intelligent Teaching Mode Based on Classroom Behavior Monitoring of Body Motion. *J Mod Educ Res*, 2023; 2: 15. DOI: 10.53964/jmer.2023015.

1 INTRODUCTION

"Body motion" is an art form or technique centered around the movement and posture of the human body, emphasizing the portrayal of the body, coordination of motions, and expression of aesthetics. This dynamic form of movement finds applications in various disciplines,

Copyright © 2023 The Author(s). This open-access article is licensed under a Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, sharing, adaptation, distribution, and reproduction in any medium, provided the original work is properly cited.

including dance, gymnastics, yoga, athletic competitions, theatrical performances, and other body-related arts. Moreover, this technology plays a pivotal role in several domains, including:

- Three-dimensional physical education in sports, music, and dance[1-2].
- Character animation and the creation of 3D character stereoscopic films[3-5].
- Implementation in driverless technology, augmented reality, and intelligent review assistance systems.
- Utilization in motion pattern recognition, detection, tracking, and monitoring[6-7].

Harnessing computer vision techniques, such as gesture estimation[8], action recognition[9], and target tracking[10], enables the accurate location and tracking of individuals within a classroom setting. This capability facilitates the identification and understanding of teaching behaviors, representing a significant direction in the realm of "artificial intelligence in education".

In a classroom environment, behaviors exhibited by both teachers and students serve as reflections of teaching styles and student concentration levels, crucial for assessing the classroom atmosphere and the attainment of educational objectives. Traditional analysis methods, predominantly reliant on external feedback sources such as post-class evaluations, suffer from issues such as subjectivity, one-sided evaluations, low efficiency, and limited room for improvement.

To address these challenges, this article introduces target detection technology based on deep learning, leveraging the PP-YOLOv2 model for classroom behavior detection. Through a substantial number of experiments conducted on two open-source large datasets. The trained model is then deployed in a genuine classroom environment, facilitating precise testing of teaching behavior. This research presents a practical solution for classroom behavior analysis in education, holding the potential to enhance teaching effects and personalized learning implementation.

In the study of monitoring students' classroom behavior, teachers can also benefit from this system. Here are some areas that teachers may benefit from:

Personalized teaching support: By monitoring students' behavior, teachers can better understand each student's learning habits and needs. This helps teachers provide more personalized teaching support tailored to different students. For example, if a teacher discovers that a student is often distracted, they can take measures to help the student improve their attention and focus.

Improving teaching strategies: students' behavioral data enables teachers to evaluate the effectiveness of their teaching strategies. Adjustments can be made to better meet students' needs and enhance teaching quality.

Early intervention and support: Monitoring students' behavior helps teachers identify problems or difficulties that students may encounter as early as possible, allowing teachers to provide timely intervention and support to help students overcome obstacles and prevent further issues.

Classroom management: Understanding students' behavior patterns aids teachers in better managing the classroom. Appropriate measures based on students' behavior can maintain order and classroom discipline, ensuring a conducive learning environment.

Enhancing educational research: Teacher participation and data collection contribute to educational research progress. Feedback and observations assist researchers in understanding student behavior and learning processes, improving research methods and tools.

Overall, monitoring students' classroom behavior not only contributes to their learning and development but also provides valuable information about teaching processes and methods. This system benefits both students and educators, ultimately improving teaching quality and meeting students' needs more effectively.

2 RESEARCH ON THE TARGET DETECTION ALGORITHM

2.1 The Defects of Traditional Target Detection Algorithm

(1) In the regional recommendation phase, traditional target detection algorithms generate numerous regions of interest by employing different-scale sliding windows across multiple iterations of input images. However, this approach introduces redundant computational overhead, adversely impacting the algorithm's operational speed. The use of fixed-size sliding windows further complicates achieving a perfect match with the target.

(2) The feature extraction phase is limited to capturing low-level characteristics of the image[11]. These characteristics have restricted expressive capabilities and are highly task-dependent. Consequently, significant changes may require a redesign of the algorithm to adapt to new requirements.

(3) Traditional target detection algorithms[12] often divide the process into three independent stages, making it difficult to attain a globally optimal solution. Additionally, the algorithm's design relies on the designer's prior knowledge of detection objectives and specific tasks, limiting its adaptability.

To address these deficiencies[13], this article aims to discover more refined computational methods to accelerate the target detection algorithm, meeting real-time requirements. Simultaneously, the goal is to design a broader range of detection algorithms to address the limited expressiveness of artificial characteristics. By adopting novel computational methods and diverse algorithms, this article seeks to enhance the performance and robustness of

the target detection algorithm. The goal is to create a system better suited for various detection tasks and capable of adapting to changes in the target.

2.2 Development of Target Detection Algorithm

Before the emergence of deep learning, the traditional target detection method was mainly composed of three parts: regional selection (sliding window), feature extraction (such as SIFT[14], HOG, etc.), and classifiers (such as SVM, Adaboost, etc.). However, there are two major problems with these traditional methods. First of all, the sliding window selection strategy lacks targetedness, the time complexity is high, and there is a window redundant. Secondly, the characteristics of hand-designed characteristics are poor. For the situation where the target changes are large, the algorithm needs to be redesigned. After the appearance of deep learning, target detection has made a huge breakthrough. The two most noticeable directions are:

(1) Detective detection algorithms based on regional recommendations are represented by R-CNN (such as R-CNN[15], SPP-Net[16], Fast R-CNN[17], Faster R-CNN[18], etc.), and use two stages of methods. First of all, a recioner area (such as Selective Search) or convolutional neural network (CNN) (such as RPN) is generated and then classified and returned to these candidate areas.

(2) Based on the deep learning target detection algorithm based on the regression method, with YOLO as the representative (such as YOLO, SSD, etc.), a single CNN directly predicts the category and location of different targets. Therefore, the target detection network based on deep learning can be divided into two categories: one is the TWO-Stage detection network, such as the R-CNN series algorithm. This type of algorithm is proposed by generating regional proposals and classifying and returned in these candidate areas to classify and return to these candidate areas. Realize target detection; the second is the ONE-Stage detection network, such as the YOLO series algorithm, SSD RetinaNet, etc., which directly predicts the category and location of the target in a single CNN network.

2.3 Application of the Target Detection Algorithm in this Article

The primary objective of this study is to analyze the learning progress of video students, providing real-time feedback to teachers on students' learning behavior and facilitating subsequent teaching adjustments. Given the scale of video data and the need for real-time analysis, a deep learning-based object detection method was applied. Specifically, the YOLO-V3 algorithm was selected as the foundational algorithm, with targeted optimizations tailored to our data characteristics. YOLO-V3 exhibits rapid training speeds while preserving accuracy, effectively addressing the challenges associated with processing large-scale video data.

Through optimization measures, the aim is to enhance the precision and real-time analysis of students' learning situations, delivering timely and precise feedback to teachers, and supporting their teaching adjustments and decision-making for further improvement. The study's primary objective is to enhance detection speed while maintaining adequate accuracy to fulfill the practical requirements of teaching applications.

3 MODEL ADVANTAGE

3.1 Improvement of the Backbone Network (ResNet50-VD)

This study adopts an improved architecture in the field of target testing, using ResNet50 VD as the backbone of the entire network. To enhance the model's expressive power, a deformable convolution is introduced. This involves replacing some convolutional layers with deformable convolutions, and specifically, a deformable convolutional layer (DCN) convolution was used in the final layer of 3×3 convolution.

The entire network consists of three parts: backbone, neck, and head, as shown in Figure 1. By using ResNet50VD and deformable convolution, PP-YOLOv2 introduces increased complexity to the network, thereby improving the overall performance and accuracy of the model. The introduction of this improved architecture has significantly improved the target detection task, yielding significant improvements and demonstrating promising results in experiments.

In order to improve the target detection algorithm, this study adopts ResNet50-VD as an alternative to the entire architecture, replacing the larger DarkNet53 used in YOLOv3. This choice is based on the advantages of ResNet in widespread application advantages, its diversified applicability across various domains, and its superior performance in terms of optimization and parameters than factors such as DarkNet53. The experimental results show that by replacing Darknet53 with a ResNet50-VD, the map of the model has been significantly improved.

Furthermore, to prevent any potential decline in performance during this substitution, the direct replacement of ResNet50-VD was avoided. Instead, DCN was used to replace specific convolutional layers. Notably, the 3×3 convolutional layer in the final stage was replaced with a DCN, rather than replacing the entire network. The introduction of DCN does not significantly increase the number of parameters or FLOPS. However, in practical applications, an excessive number of DCN layers can considerably extend inference time. Striking a balance between accuracy and speed, this study strategically replaced only the final stage's 3×3 convolutional layer with DCN[19].

The implementation of this improvement strategy

https://doi.org/10.53964/jmer.2023015

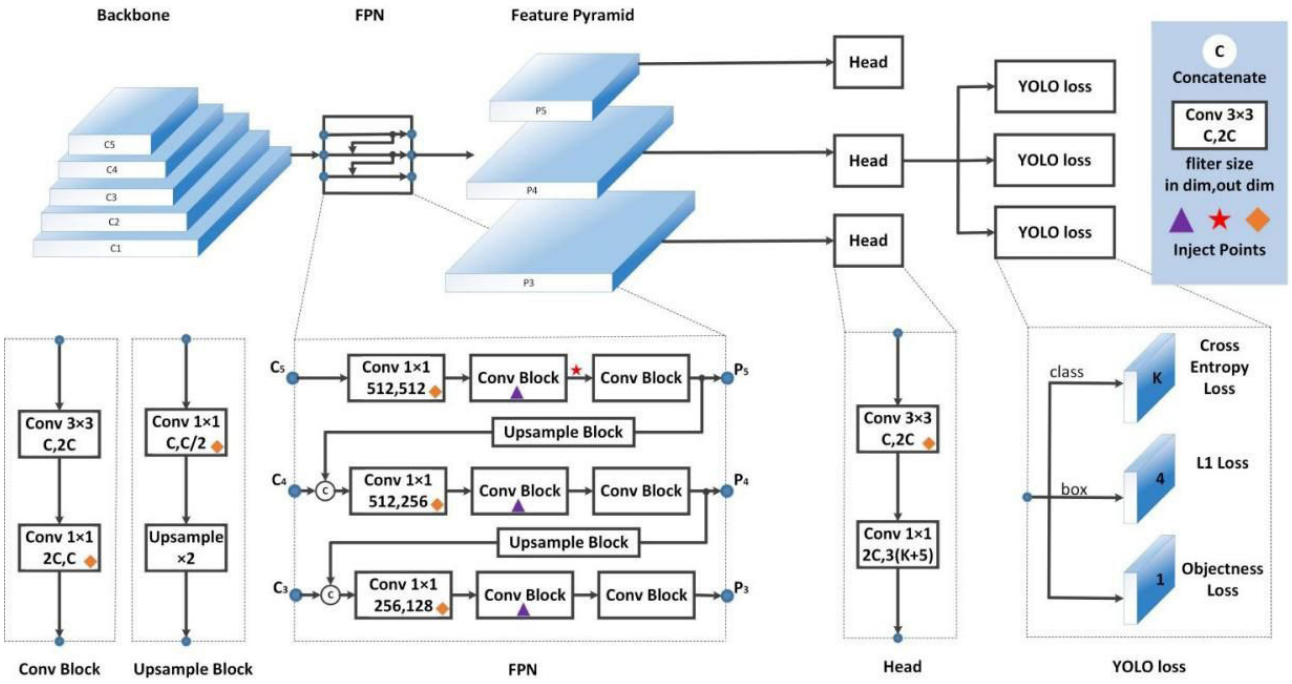

Figure 1. Backbone network.

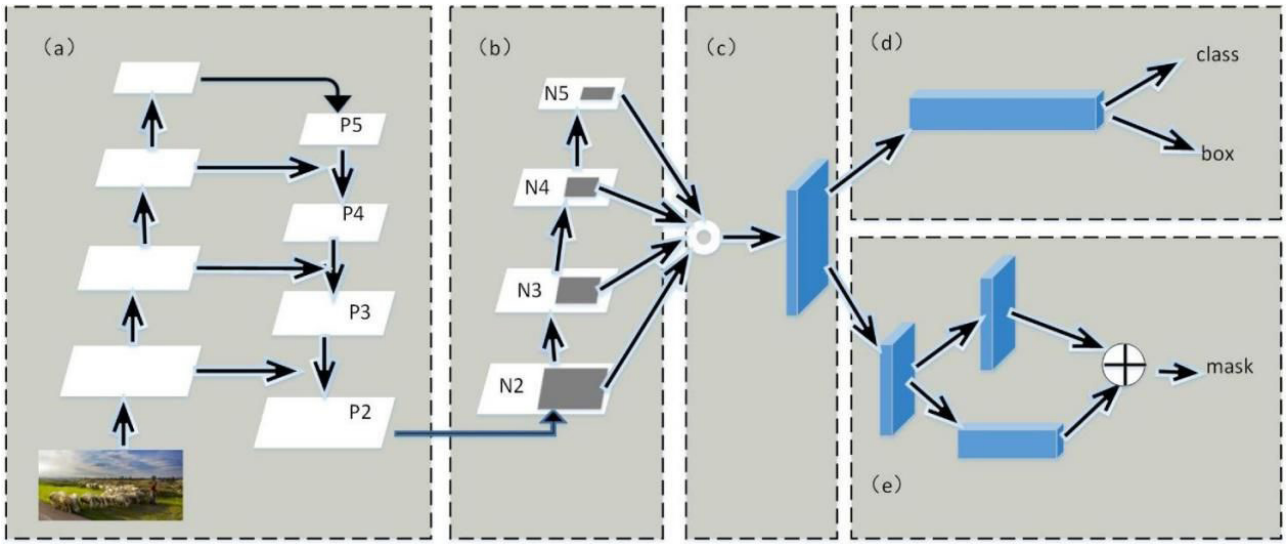

Figure 2. Path aggregation network.

ensures a balanced trade-off between accuracy and speed, contributing to a further enhancement in the performance of the target detection algorithm.

3.2 Path Aggregation Network (PAN)

In contrast to PP-YOLO, which employs the Feature Pyramid Network for constructing a feature pyramid, our model utilizes the PAN. Our network leverages low-level positional cues to reinforce the feature hierarchy through the bottom-up pathway, thus shortening the information transfer distance between the lower and upper layers (as illustrated by the green line in Figure 2). Additionally, PAN introduces an adaptive feature pool to establish connections between the feature grid and all feature layers for top-down feature aggregation. This innovative approach enables the direct propagation of valuable information from each feature layer to the proposal subnet, facilitating a comprehensive aggregation of feature information.

Activity function is a function mapping the output of the neuron input to the output in the neural network. The currently widely used activation functions include ReLU, Tanh, Sigmoid, Leaky ReLU, and Swish. This article uses a new activation function Mish for PP-YOLOv2.

The advantage of the Mish function is that it has no upper limit, which can reach any height and avoid saturation. For negative values, Mish allows a mild gradient to flow,

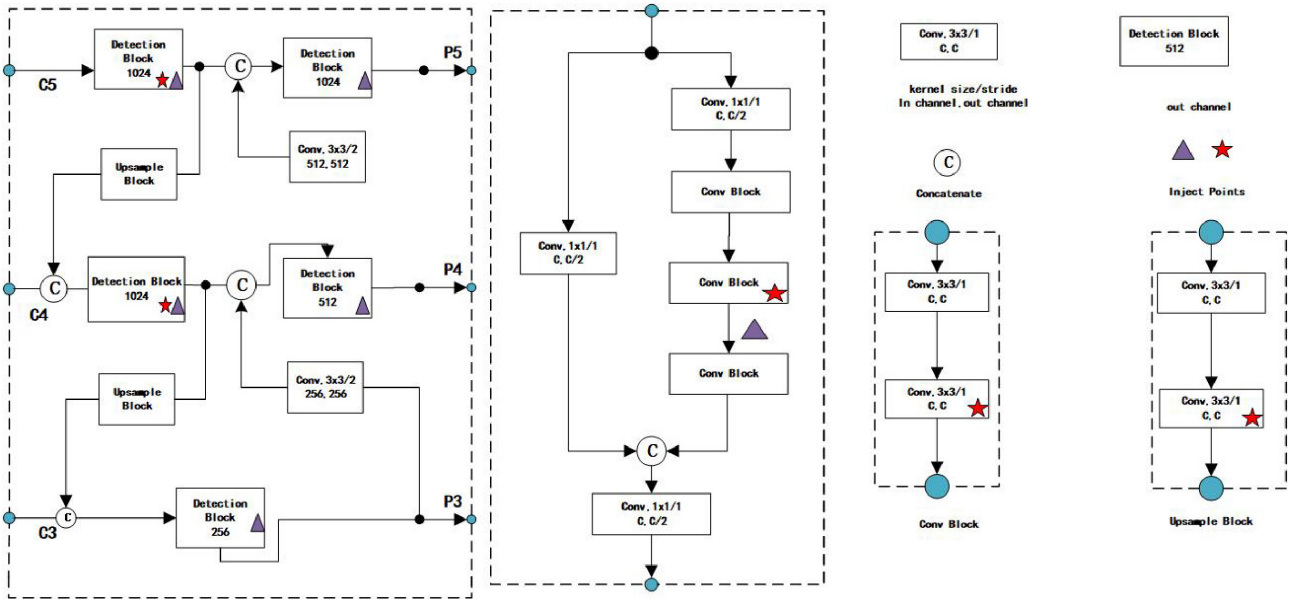

Figure 3. PP-YOLOv2 detection of the neck architecture.

in contrast to the hard zero borders of ReLU. The smooth activation function enables the information transmission to the neural network, thereby improving accuracy and generalized performance. According to verification experiments, Mish's accuracy was increased by 0.494% compared to SWISH, which was 1.671% compared with ReLU[20].

The Mish function has a lower limit under the protection of information flow, so that the function can be regulated regularized. The expression of the Mish function is as shown in (1):

$$Mish = \mathrm{x} * \tanh\left(\ln\left(1 + e^{\mathrm{x}}\right)\right) \quad (1)$$

This activation function, being non-monotonic, introduces updates to most neurons, providing enhanced stability that surpasses other activation functions, such as Swish and ReLU.

Similar to many other models, PP-YOLOv2 also incorporates the Mish activation function. However, a distinctive approach is taken as Mish activation is not applied to the backbone network. The pre-trained parameters of the original backbone network have achieved an impressive TOP-1 accuracy of up to 82.4% on ImageNet. Consequently, this model continues to leverage the original backbone network and introduces the Mish activation function in the Detection module to further enhance its performance (see Figure 3)[21].

3.3 Improvement of Loss Function

IOU AWARE Branch: In PP-YOLO, we noticed that the calculation of IOU Aware Loss, where the Soft Weight Format was used was found to be inconsistent with the original intention. To improve this problem, we introduced the incorporation of the Soft Label Format. Equation (2) illustrates the IOU Aware Loss.

$$loss = -t * \log\left(\sigma(p)\right) - (1 - t) * \log\left(1 - \sigma(p)\right) \quad (2)$$

In the above formula, t is the IOU between the anchor point and the grounding wire junction box. p is the raw output of the IOU Aware Branch, subject to the SIGMOID activation function. The IOU Aware Loss specifically addresses positive samples, contributing to a notable enhancement in model performance by replacing the loss function. Compared to before, this improvement significantly improves the performance of the model.

4 EXPERIMENT

4.1 Experimental Description

Research on monitoring students' classroom behavior typically requires careful consideration of measures to protect privacy, especially in the experimental stage before large-scale implementation. Potential methods and strategies to protect the privacy of classroom behavior monitoring are outlined below:

Anonymity and data deidentification: Researchers can ensure that the collected data is anonymous, devoid of personal information that can identify students. In addition, for any data that may identify students, data deidentification methods can be used to eliminate any potential personal identification information.

Informed consent: Before starting research, researchers can obtain informed consent from students, parents, or educational institutions. This means that they clearly inform relevant parties about the nature, purpose, and data collection method of the study, obtaining written consent.

Data desensitization and aggregation: The collected data can be desensitized before analysis, such as by deleting

https://doi.org/10.53964/jmer.2023015

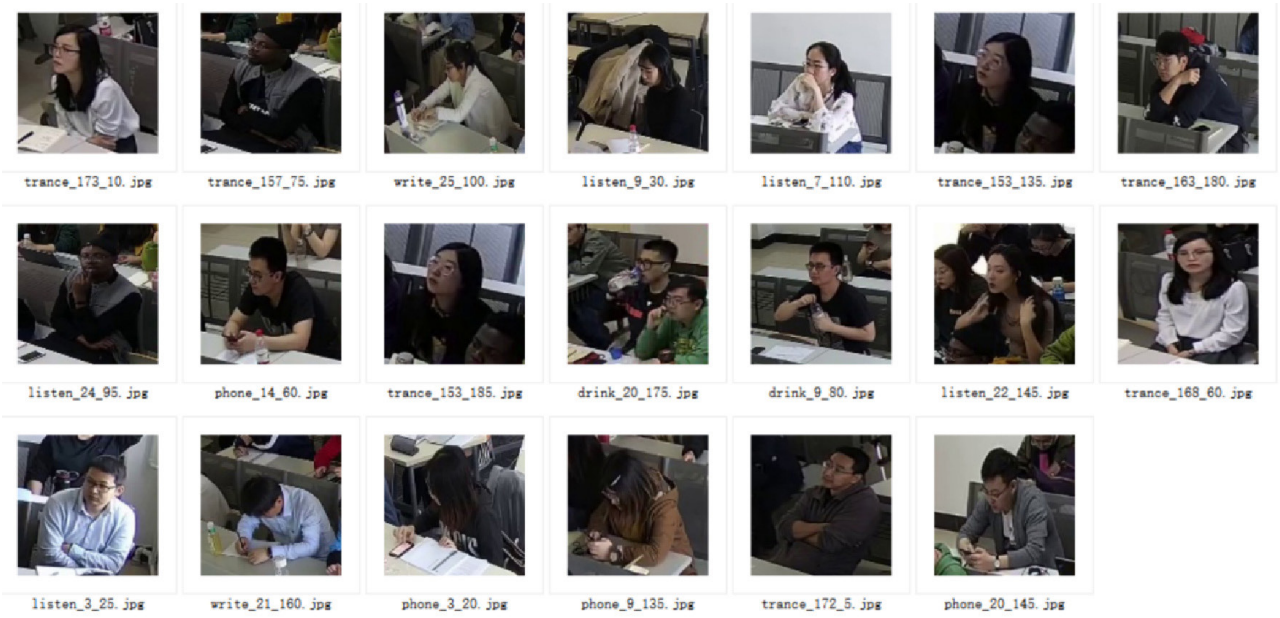

Figure 4. Experimental data set.

or replacing specific personal information before analysis. Aggregate data to summarize and analyze overall trends, rather than focusing on individual student behavior.

Data security and storage: The collected data needs to be stored and processed in a secure manner to prevent unauthorized access or data leakage. This may include encrypting data, restricting access permissions, and using secure storage and transmission methods.

The principle of minimum data: Researchers should only collect the minimum necessary data related to their research objectives to reduce potential privacy risks. Unneeded information should be excluded from the scope of data collection.

Ethical review and compliance: Prior to conducting a study, researchers may need to submit a research plan for ethical review to ensure that the study complies with ethical standards and regulatory requirements, especially when involving human participants.

Data access control: For the data involved in the study, access control measures need to be implemented to ensure that only authorized personnel can access and process the data.

Data retention and destruction: Data retention and destruction are necessary after the completion of the research to avoid unnecessary retention and ensure compliant destruction.

These methods and strategies aim to strike a balance between the needs of educational research and the protection of student privacy. Researchers should consistently adhere to ethical standards, complying with relevant regulations and policies to ensure proper protection of students' privacy throughout the research process.

4.2 Experimental Environment

In the experiment, the computer configuration and the system environment are as follows: CPU is 12th Gen Intel (R) Core (TM) I9-12900HX; GPU is GeForce RTX 3090; operating memory is 24G; the operating system is Window 10; Python version is 3.9; The learning framework is PyTorch 1.13.1.

4.3 Experimental Data Set

The total number of data sets in this experiment is 5012. The five typical students' behavioral status analyzed in this article includes: watching mobile phones, drinking water, writing, trancing, and listening. A total of 2506 (about 50%) of the semi-automatic labeling part, as shown in Figures 4 and 5.

The dataset for this experiment has a total of 5012 pieces of data, each data contains a picture and its corresponding label, divided into the following five categories: listen, write, phone, drink, and trance. It is further divided into training, validation, and test sets based on a ratio of 7:2:1, as shown in Figure 6.

4.4 Experimental Explanation

The experiment selected the PP-YOLOv2 ResNet50 pre-training model as the target detection model. The image input size is set to 128*px*×128*px*, with an average image value of 0.485, 0.456, 0.406, and an image variance of 0.229, 0.224, and 0.225. The experiment runs for 300 iterations with a learning rate of 0.0001, a batch size of 30, and a storage interval of 30.

https://doi.org/10.53964/jmer.2023015

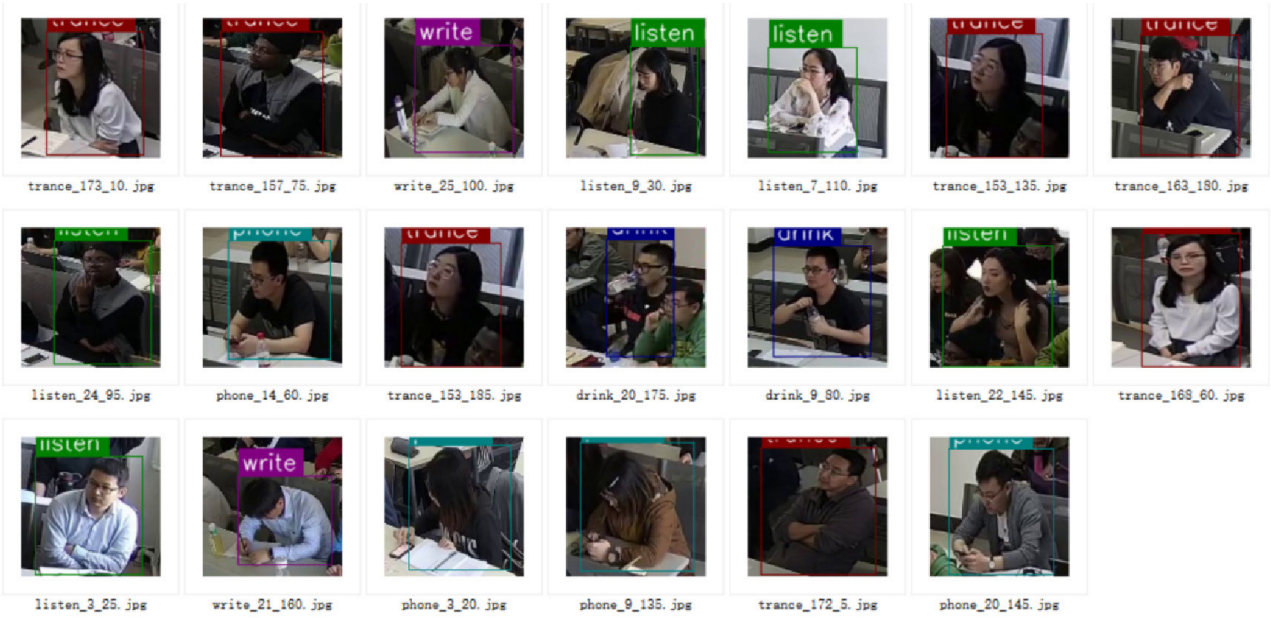

Figure 5. Data annotation.

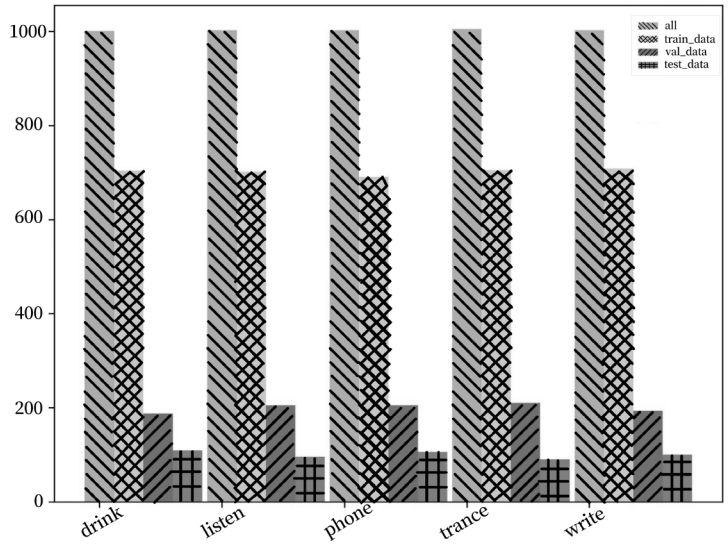

Figure 6. Experimental data set.

Upon completion of the dataset, it is divided into two folders. The JPEGIMAGE folder stores images in JPG format, while the Annotations folder contains labels in XML format. The file distribution is shown in Figure 7.

4.5 Experimental Results

Evaluation criteria: In order to evaluate the performance of the model, the experiment used five labels: listening, writing, making phone calls, drinking, and trance. Three key indicators are measured for evaluation: accuracy, recall, and average accuracy (APALAGE precision). The evaluation results are shown in Table 1 to gauge the effectiveness of the model.

The "five typical student behavior states" and "five categories" mentioned in the previous paragraph are divided based on the different types of behavior that students may exhibit in a classroom setting. These categorizations draw from observations, research by relevant researchers and educational psychologists' understanding of various activities and states students may exhibit in the classroom.

The division of these behavioral states is based on academic research, educational psychology theory, and experimental design needs. The specific theories are as follows:

Observation and experiential foundation: Observe and record students' actual behavior in the classroom to identify five typical behavioral states. For example, students often look at their phones, drink water, take notes, feel distracted, or concentrate on listening during class. These observations are based on years of educational experience

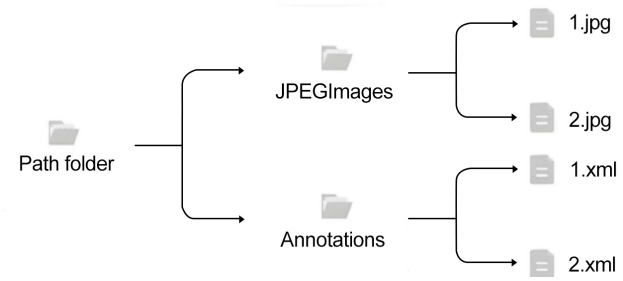

Figure 7. Experimental file distribution.

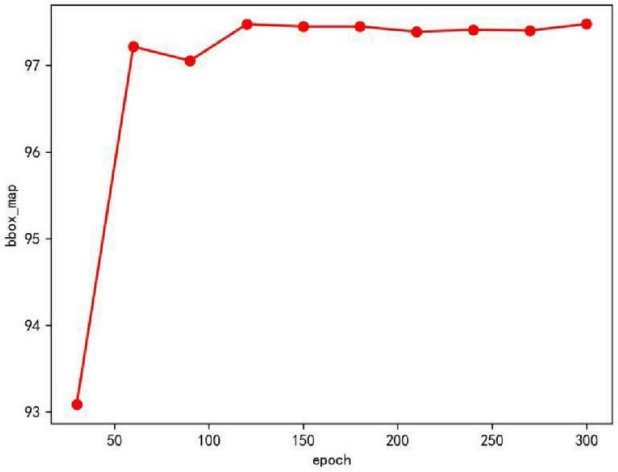

Figure 8. Experimental data set.

and research. The relevant theoretical foundation is attention and cognitive theory, exploring how humans choose, maintain, and allocate attention resources. Related concepts include multitasking, distraction, and working memory. Specific theories may include Broadbent's "Selective Attention Theory", Treisman's "Feature Integration Theory", Baddeley's "Working Memory Model", etc. You can search for these theories and related research to learn more.

Educational psychology theory: The division of these behavioral states may also be based on educational psychology theories, such as learning and attention theory. For example, listening to classes, taking notes, distracting attention, and paying attention to lectures are all concepts related to students' attention and learning behavior. The relevant theoretical foundation is educational psychology, which studies psychological phenomena in the process of learning and education. Related concepts include learning theory, educational evaluation, and educational psychological measurement. Some famous educational psychology theories include Piaget's cognitive development theory, Vygotsky's socio-cultural theory, and Baroque's achievement motivation theory, and so on. You can search for these theories to learn more about relevant literature.

Experimental design purpose: Based on research questions and practical needs, that is, the detection of students' classroom behavior. The experimental design defines categories according to research purposes, ensuring accuracy and operability. The theoretical basis for experimental design criteria can be found in classic literature on experimental design and methodology, including Campbell and Stanley's experiments and experimental design, Creswell's research design, Cohen, Manion, and Morrison's research methods, to obtain a theoretical basis for selecting and evaluating experimental design criteria.

Table 1. Evaluation Result

Class	Precision	Recall	Average Precision
Drink	0.9176	1.0000	1.0000
Listen	0.9124	0.9952	0.9649
Phone	0.8904	1.0000	1.0000
Trance	0.9045	1.0000	1.0000
Write	0.9208	1.0000	1.0000

The division of these actions can help better understand students' behavior in the classroom and evaluate the performance of the model. The labels used in the evaluation criteria are also used to measure the accuracy of the model's classification of these different behavioral states, in order to evaluate the model's performance in this task.

In the experiment, the PP-YOLOv2 pre-training model achieved a Mean Average Precision (MAP) value of 99.3% after 300 iterations, as shown in Figure 8. The results indicate high performance and accuracy, with the overall MAP reaching 99.3%. Each category MAP falls between 96% and 100%, suggesting balanced performance in identifying different categories.

The results and accuracy of the action test in the experiment are shown in Figure 9.

4.6 Experiment Analysis

The experimental results, obtained after 300 iterations, highlight the exceptional performance and accuracy of the pre-trained PP-YOLOv2 model when evaluated on the 5012 dataset. This achievement holds significant implications for educational applications. The overall MAP value impressively reaches 99.3%, underscoring the model's exceptional target detection capabilities. Furthermore, it's noteworthy that each category's MAP score falls within the range of 96% to 100%, indicating a well-balanced performance across different categories of objects. This balance suggests that the model can effectively identify various categories, a feature of paramount importance in educational settings where diverse classroom behaviors need to be detected and analyzed.

The notable accuracy achieved in our experiments underscores the model's robust generalization capabilities

https://doi.org/10.53964/jmer.2023015

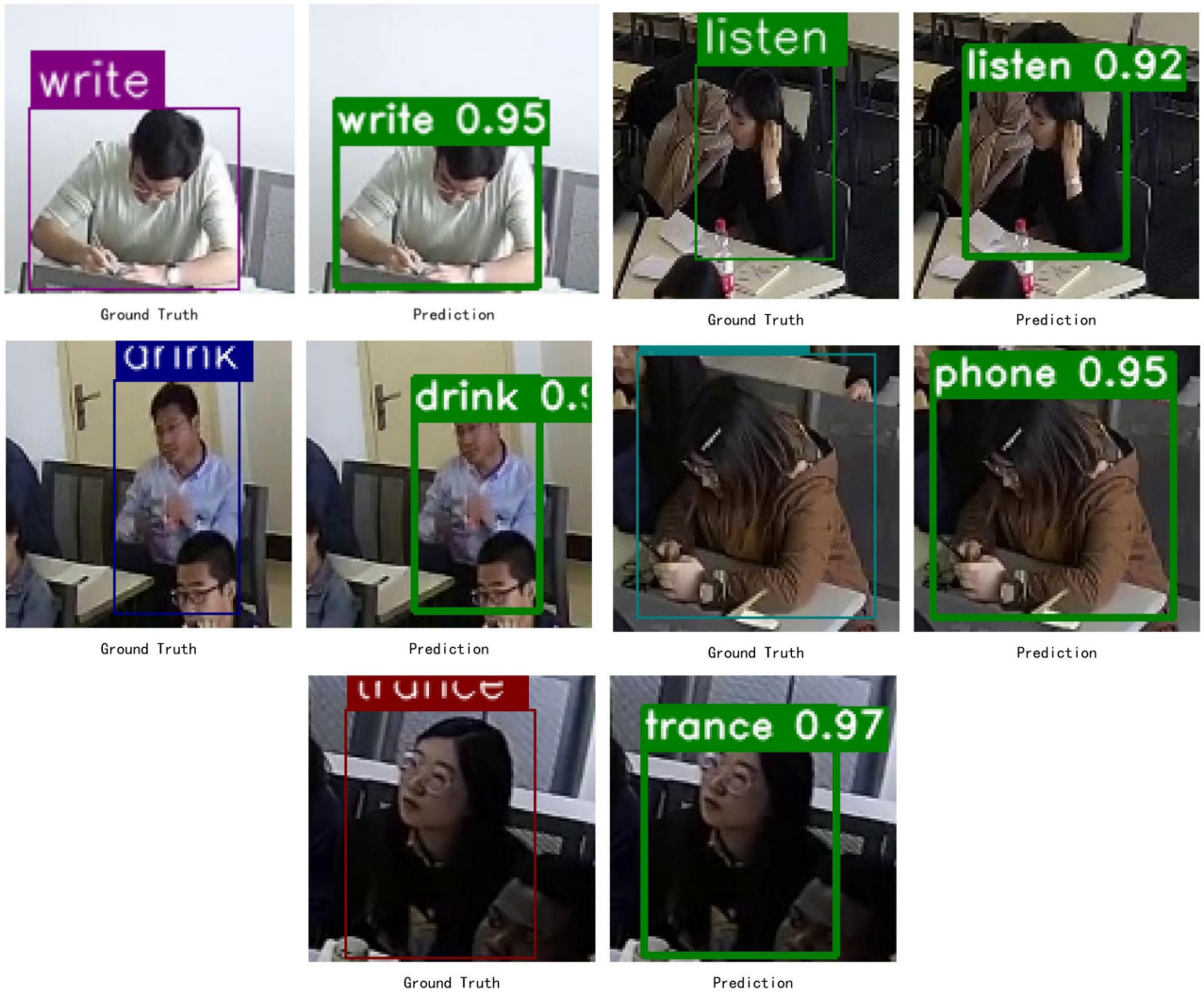

Figure 9. Experimental prediction results.

within the current dataset. However, it is essential to recognize that our experiments were limited to 300 iterations, allowing for potential further optimization. In practical educational contexts, there is room for improvement by considering an increase in the number of iterations or adjustments to training parameters. Furthermore, to comprehensively evaluate the model's generalization, conducting experiments on larger-scale datasets becomes imperative.

Increase the Number of Iterations: Exploring an increase in the number of iterations, ranging from 500 to 1000 or beyond, is crucial to assess whether the model's performance continues to improve with extended training. This investigation aligns with the educational objective of refining detection accuracy and efficiency.

Adjust Training Parameters: Experimenting with variations in training parameters such as learning rates, weight attenuation, and optimizer configurations can shed light on how the model behaves under different settings. This adaptive approach aims to discover the optimal parameter combination that further enhances the model's accuracy, aligning with the educational goals of precise behavior detection.

Cross validation: The use of K-fold cross-validation method is beneficial for evaluating the model's robustness. By dividing the dataset into K subsets, with each subset iteratively serving as both the validation and training sets, the model's performance across various training and validation sets can be comprehensively evaluated. This approach minimizes the risk of overfitting and helps achieve consistent detection results, which is essential in educational environments.

Real-World Scenario Testing: Finally, transitioning from controlled experiments to real-world educational scenarios, such as online classrooms or physical classrooms, holds substantial promise. Observing how the model performs in these authentic contexts provides valuable insights into its practical applicability, allowing for the validation of its effectiveness in real-time behavior detection. This also helps identify potential challenges and

avenues for further refinement.

5 CONCLUSION

Classroom behavior detection is critically important in the field of education, serving as a powerful tool to assess student engagement and concentration, directly impacting the quality of learning experiences. Despite its significance, traditional behavioral detection methods have often been intricate and time-consuming, hindering their practical application in educational settings.

In response to this challenge, this paper introduces an innovative classroom behavior detection method based on PP-YOLOv2, harnessing the capabilities of computer vision and deep learning technology. By meticulously curating student sample datasets and comprehensive data preprocessing, this method offers an efficient and precise means of behavior detection.

Notably, our research incorporates the use of the Mish activation function, an inventive approach that enhances the model's learning capacity and augments the accuracy of behavioral detection. The results of our experiments clearly demonstrate that this method excels in real classroom environments, presenting educators with a reliable, real-time classroom behavior monitoring tool.

This research carries profound implications for the education sector by enabling continuous, real-time monitoring and evaluation of student behavior. Such capabilities are invaluable in improving teaching outcomes and promoting personalized learning journeys. Educators gain access to timely insights into students' learning statuses, enabling tailored feedback and guidance, ultimately enhancing teaching quality and fostering students' self-directed learning capabilities.

Looking ahead, future research endeavors should explore the broader application of this method across diverse grade levels, subject areas, and teaching scenarios. Further optimization of the algorithm will be essential to enhance model performance and efficiency, aligning it with the demands of real-time behavior detection within the education field. Through relentless research and innovation, we aim to push the boundaries of educational technology, ultimately shaping a more conducive and adaptive learning environment for both students and educators.

Acknowledgements

This work is supported by the National Undergraduate Innovation Training Project of China under Grant (No. 202314278014), the Collaborative Project for the Development of Guangzhou Philosophy and Social Science in 14th Five-Year Plan (No. 2023GZGJ171), the Educational Science Planning Project of Guangdong Province (No. 2022GXJK073, No. 2023GXJK125), the Science and Technology Plan Project of Guangzhou (No. 202002030232, No. 202103010004), the Natural Science Foundation of Guangdong Province (No. 2022A1515010485), the Teaching Quality and Teaching Reform Project of Guangdong University of Education (No. 2022jxgg33) and the Special Support Program for Cultivating High-Level Talents of Guangdong University of Education (2022 Outstanding Young Teacher Cultivation Object: Wanyi Li).

Conflicts of Interest

The authors declared no conflict of interest.

Author Contribution

Chen Y, Qian Z, and Li W designed the experiment. Li W, Chen Y, and Huang J supervised the work. Qian Z performed the data analysis. Chen Y, Weng H, and Zheng J drafted the manuscript. All the authors contributed to writing the article, read and approved its submission.

Abbreviation List

CNN, Convolutional Neural Network
DCN, Deformable Convolutional Layer
MAP, Mean Average Precision
PAN, Path Aggregation Network

References

[1] Zhen J, Fang Q, Sun J et al. SMAP: Single-Shot Multi-person Absolute 3D Pose Estimation: Proceedings of the Computer Vision - ECCV 2020: 16th European Conference. Glasgow, UK, 23-28 August 2020.[DOI]

[2] Benzine A, Luvison B, Pham QC et al. Single-shot 3D multiperson pose estimation in complex images. *Pattern Recogn*, 2021; 112: 107534.[DOI]

[3] Ionescu C, Papava D, Olaru V et al. Human 3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE T Pattern Anal Mach Intell*, 2013; 36: 1325-1339.[DOI]

[4] Peña-Tapia E, Hachiuma R, Pasquali A et al. LCR-SMPL: toward real-time human detection and 3D reconstruction from a single RGB image: 2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct. Recife, Brazil, 9-13 November 2020.[DOI]

[5] Bogo F, Kanazawa A, Lassner C et al. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image: Computer Vision - ECCV 2016: 14th European Conference. Amsterdam, The Netherlands, 11-14 October 2016.[DOI]

[6] Bagautdinov T, Alahi A, Fleuret F et al. Social scene understanding: end-to-end multi-person action localization and collective activity recognition: Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu, HI, USA, 2017.

[7] Zhu K, Wang R, Cheng J et al. A novel skeleton-based action recognition method via cuboid rearranging: 2018 IEEE International Conference on Information and Automation (ICIA). Wuyishan, China, 11-13 August 2018.[DOI]

[8] Deng YN, Luo JX, Jin FL. Overview of human pose estimation

methods based on deep learning. *Comput Eng Appl*, 2019; 55: 22-42.

[9] Zhang J, Mao X, Chen T. A Comprehensive Review of Motion Object Tracking Algorithms [In Chinese]. *Appl Res Comput*, 2009; 26: 4407-4412.[DOI]

[10] Zhang S, Gong Y, Wang J. Development of Deep Convolutional Neural Networks and Their Applications in Computer Vision [In Chinese]. *Chinese J Comput*, 2019; 42: 45-53.[DOI]

[11] Zhai J, Zhao W, Wang X. Research on Image Feature Extraction [In Chinese]. *J Hebei Univ (Nat Sci Edit)*, 2009; 29: 106-112. [DOI]

[12] Fang LP, He HJ, Zhou GM. Research overview of object detection methods. *Comput Eng Appl*, 2018; 54: 11-18.

[13] Li K, Chen Y, Liu J et al. Review of object detection algorithms based on deep learning [In Chinese]. *Comput Eng*, 2022; 48: 1-12.[DOI]

[14] Lin H, Ma Y, Song T. Research on Target Tracking Algorithm Based on SIFT Feature [In Chinese]. *Acta Automatica Sinica*, 2010; 36: 1204-1210.

[15] Girshick R. Fast R-CNN: Proceedings of the IEEE international conference on computer vision. Santiago, Chile, 2015.

[16] He K, Zhang X, Ren S et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE T Pattern Anal Mach Intell*, 2015, 37: 1904-1916.[DOI]

[17] Girshick R. Fast R-CNN: Proceedings of the IEEE International Conference on Computer Vision, 2015

[18] Ren S, He K, Girshick R et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *NeurIPS*, 2015.

[19] Huang X, Wang X, Lv W et al. PP-YOLOv2: A practical object detector. *arXiv preprint arXiv:2104.10419*, 2021.[DOI]

[20] Mish MD. A Self Regularized Non-Monotonic Activation Function. *arXiv preprint arXiv:1908.08681*, 2019.

[21] Liu Q, He HY, Wu LJ et al. Classroom Teaching Behavior Analysis Method Based on Artificial Intelligence and Its Application. *China Electrochem Educ*, 2019: 13-21.