Towards the Classroom of the Future: Improvement Strategies for Automatic Attendance Systems Using Facial Recognition Combining Behavioral Recognition

Sheng Qiang[1]

[1]School of International Education, Guangdong University of Technology, Guangzhou, 511495, China
3121010002@mail2.gdut.edu.cn

Abstract. In a classroom setting, manual roll call consumes a significant portion of the limited available time. To address this issue, automatic attendance systems offer a solution. This paper presents an improved automatic attendance system based on facial recognition through a thorough analysis of previous research, algorithm comparisons, and experimentation. The proposed system enhances efficiency by integrating facial recognition with additional behavioral recognition functions, maintaining a compact size and consistent processing block count. Through optimization efforts, the system is capable of identifying students who arrive late or leave early, reducing real-time processing demands to enable deployment on resource-constrained hardware. Furthermore, the system can adapt to slower computing speeds while maintaining high accuracy, and in case of misidentification, it can still make accurate judgments. Following a series of experiments and tests, the system demonstrates an impressive accuracy level of 96%. Implementation of this system enhances classroom efficiency and simplifies the evaluation process for students' coursework grades.

Keywords: automatic attendance, object detection, facial recognition, behavioral recognition.

1 Introduction

As technology continues to advance, facial recognition has found widespread application across various domains in people's daily lives. As noted by Yassin Kortli [1], faces, being unique to each individual, serve as biometric passwords, effectively representing one's identity. Unlike traditional string-based passwords, facial recognition offers a combination of security and convenience, as there are no risks of leakage and authentication can be performed seamlessly across different environments. Consequently, facial recognition has increasingly supplanted traditional identity cards and has been extensively utilized in various contexts such as business transactions and boarding identification.

Y. Wang (ed.), *Proceedings of the 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024)*, Advances in Intelligent Systems Research 185,
https://doi.org/10.2991/978-94-6463-512-6_66

This convenient technology holds potential for deployment within the education industry. Clearly, class attendance plays a vital role in students' acquisition of academic knowledge and skills. The attendance rate significantly influences a student's participation and overall performance grade. Therefore, it is imperative for professors to accurately record student absences. However, traditional roll call methods are time-consuming, especially in higher education settings where large numbers of students congregate in lecture halls for classes. Fortunately, advancements in scientific techniques have made automatic attendance systems a viable solution.

Previous researchers have explored various algorithms and practical approaches to implement automatic attendance systems. For instance, YOLO [2], developed by Redmon et al., has been widely utilized for target recognition, as demonstrated by Fu [3] in their implementation of an attendance system. Additionally, Wu [4] introduced a novel algorithm tailored specifically for classroom environments to enhance recognition accuracy. Despite some inherent flaws, significant progress has been achieved in this field.

Moreover, additional functionalities such as behavioral recognition can be integrated into the system. Clearly, mere attendance tracking is insufficient to comprehensively evaluate an individual's performance. Behavioral monitoring is also crucial, particularly in addressing multifaceted dimensions such as preventing students from dozing off during class. Therefore, alongside facial recognition, simple action detection should be incorporated into the system.

Previous studies have explored similar avenues. For instance, Liu [5] utilized convolutional neural networks to analyze students' actions in videos, while Huang [6] compared various algorithms for behavior detection and proposed an optimized version based on deep spatiotemporal residuals. While their contributions are commendable, there remains scope for further enhancement.

This paper aims to review the benefits and innovations of both automatic attendance and behavioral recognition systems. Additionally, it will propose strategies for improvement and optimization, culminating in the feasibility of their integration.

2 Design of the system

2.1 System Structure

Most facial and behavioral recognition systems follow a processing structure depicted in the data flow diagram shown in Figure 1. For facial recognition systems (illustrated in the top half), the process unfolds as follows: Firstly, an object detection model such as YOLO [2] or Single Shot MultiBox Detector (SSD) [7] is deployed on the monitoring devices to detect the student. Subsequently, the facial regions within the images are extracted and forwarded to the facial recognition module. Thirdly, the facial images are compared with the pre-recorded faces of students stored in the database using algorithms like FaceNet [8]. Finally, the attendance status is inferred based on a predefined threshold value.

For behavioral recognition systems (depicted in the bottom half), the initial stage mirrors that of facial recognition, employing an object detection model. Subsequently,

actions are captured utilizing various methods such as time-space graphs, skeletons, etc. Once the student's action is confirmed, the result is logged accordingly.

Fig. 1 provides a clear overview of these two whole systems. Those three algorithms are worth focusing on and can be improved or optimized depending on the practical situation. Moreover, these two systems have a high overlap that uses consistent computing blocks, inspiring the possibility of combining. With simplification, the ultimate system will contain two functions to provide comprehensive assessment while the time expense or hardware resource remains relatively low.

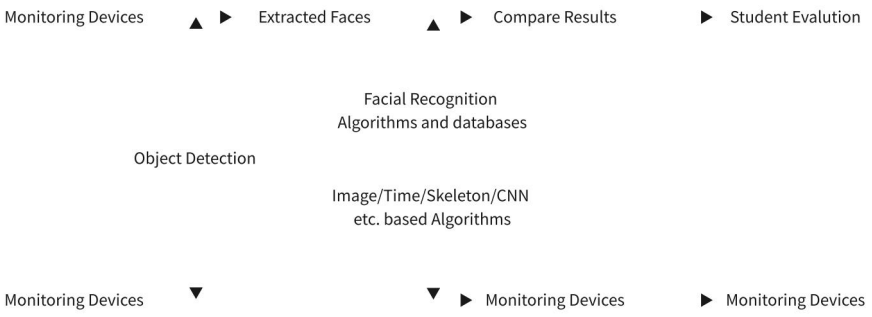

Fig. 1. Data flow diagram (Picture credit: Original)

2.2 Analysis and improvement strategies

Former researchers have already developed many innovative and practical design schemes, but there is still room for improvement in terms of accuracy and adaptability. Meanwhile, some related algorithms have been updated to the latest version with higher performance. This chapter aims to analyze the previous research and propose an improved system.

Analysis of the current systems. The analysis of the two systems will be separated, containing both advantages and disadvantages. The analysis of the related algorithm will be illustrated by listing graphs and comparing different versions.

For facial recognition systems. Fu's system [3] applied YOLOv5 and FaceNet in face detection and recognition tasks with an acceptable accuracy rate on average occasions. Their main idea, including the detecting methods, Choice of the algorithm, and data flow structure, deeply inspired this paper's improved version of the system. The high real-time performance of YOLOv5 perfectly satisfied the attendance process requirements: When YOLOv5n6 is applied and computed on the CPU, each 1280-pixel image takes only 153ms to finish recognizing [9]. Moreover, FaceNet is also well-performed and attains 99% accuracy on the LFW dataset [8]. In Wu's system

[4], they added a backlight compensation function and optimized the recording angle, allowing the model to be more adaptive.

Although the accuracy of the facial recognition module is high, the accuracy of the face extraction (object detection) module is unsatisfactory. In particular, those two adopt one-time attendance checks, leading to low precision overall. Besides, some of the extreme situations have not been considered. For example, if a student goes to the bathroom during the attendance check or leaves early after the attendance check, the system will judge mistakenly. Even if one steps back and assumes the accuracy rate reaches 100%, the system still cannot avoid misjudgments. Consequently, their systems need to be improved.

For behavioral recognition systems. Liu's system [5] uses a 4K resolution rate camera with an extensive lens aperture range. Hence, this system can be applied in large classrooms. The application of ResNet allows for deeper training than a regular CNN, enabling recognition of subtle differences in actions, such as writing, sleeping, staring, playing on phones, etc. Training and inference time expense is excellent too, below 30s and 1s respectively. Huang's system [6] judges based on time-space contrast. Their paper aims to analyze the video frames, enabling the detection of dynamic actions like turning, raising hands, etc. Using video frames to conclude is very enlightening in dealing with the lack of accuracy in the object-detecting module. It will be mentioned afterward.

Since the aim is to combine behavioral recognition to evaluate one's class performance, some functions are redundant. For instance, actions like standing up, turning around, and raising hands are unnecessary to be detected because they are not decisive. In addition, computing overhead is worth considering, as analyzing videos frame by frame will cost enormous resources. Therefore, their systems should be simplified.

For related models/algorithms. The CNN-based method and YOLO are widely used for object detection. Lee and Kim [10] argued that images must be scanned across in the CNN-based method, and then classifying and locating operations must be executed. In contrast, YOLO transformed object detection into a regression problem, leading to a faster execution speed. However, on average, the CNN's accuracy is 10% higher than that of YOLO. As shown in Table 1.

Table 1. Comparison of CNN and YOLOs [11]

	Model	mAP^{Val}	Frames per second	mAP x FPS
CNN based	Faster R-CNN ResNet	76.4	5	382
	Faster R-CNN VGG-16	73.2	7	512
	Faster R-CNN ZF	62.1	18	1118
YOLOs	YOLO	63.4	45	2853
	Fast YOLO	52.7	155	8169
	YOLOv2 288 288	69.0	91	6279

YOLOs possess real-time features, making up for the defect in accuracy, but CNN is more adaptive in backlight and dark environments that will not be influenced by the weather outside the classroom.

Universally acknowledged, YOLOv5 and YOLOv8 are the two most well-performed versions of this model. The latest version is YOLOv9. Table 2 shows the operating situation of those versions trained on COCO. In general, as the version is updated, the level of accuracy also increases.

Table 2. The performance of different versions of YOLO [9,12,13]

Model	Size (MB)	Params (M)	FLOPs (G)	mAPVal 50	mAPVal 50-95	Speed (ms)	hardware
YOLOv5n	3.87	1.9	4.5	45.7	28.0	45	AWS p3.2xlarge CPU
YOLOv5m	40.8	21.2	49.0	64.1	45.4	224	
YOLOv5x	166	86.7	205.7	68.9	50.7	430	
YOLOv8n	6.2	3.2	8.7	52.6	37.3	80.4	Amazon EC2 P4d CPU
YOLOv8m	49.7	25.9	78.9	67.2	50.2	234.7	
YOLOv8x	131	68.2	257.8	71.0	53.9	479.1	
YOLOv9t	-	2.0	7.7	53.1	38.3	-	-
YOLOv9c	49.1	25.3	102.1	70.2	53.0	-	-
YOLOv9e	112	57.3	189.0	72.8	55.6	-	-

The authors did not provide the inference speed for YOLOv9. Finding representative classroom monitor videos is challenging since most of them are private, making it hard to observe real classroom situations. However, the similar person-detection tasks can differentiate YOLOv8 and YOLOv9's performance as well. Table 2 shows that YOLOv8m and YOLOv9c has similar size, parameters and FLOPs; and both has the intermediate performance. Hence, these two versions are chosen to be tested. Fig. 2 is the screenshot of the experimentation that recognizing walking persons. Although YOLOv9c's accuracy is slightly higher, the inference time shows a stark contrast, with more than 50ms per frame, while YOLOv8m can deal with about 18ms per frame (computed on NVIDIA RTX 2060M). Moreover, the lighter versions, YOLOv9t and YOLOv9s, have not been released, indicating that YOLOv9 is still immature.

Fig. 2. Pedestrians recognizing task (Picture credit: Original)

When it comes to facial recognition frameworks, both of the existing ones maintain a high degree of accuracy, as shown in Table 3.

Table 3. High accuracies on LFW verification [14]

Model	Train data	LFW(%)
DeepFace	Privatedataset	97.35
VGGFace	VGGFace	98.95
FaceNet	Privatedataset	99.63
ArcFace	ms1m	99.83

Each model has its drawbacks: DeepFace and VGGFace have deep layers and numerous parameters, causing high Calculation and maintenance costs; FaceNet is sensitive to the training set, and the triple loss function makes the training process unstable.

improvement strategies. At this point, improvement strategies can be proposed.

Function rejection and model selection in Behavioral Recognition. In action detecting, dynamic ones like standing, raising hands, raising head, turning, and stretching should be abandoned. These actions require analyzing the consistent frames dynamically, causing high resource demands while they are not significant. Similarly, analyzing whether students are looking at the platform or blackboard should be discontinued, as a high-resolution monitor is required to predict their sight accurately. On the contrary, the key point determining students' performance is simple to recognize: actions like sleeping, playing on mobiles, or regularly sitting are relatively static when captured on camera. Hence, high-density frame analysis becomes unnecessary.

Consequently, model selection becomes more diverse. YOLO is no longer essential that the demands in real-time reduced. With a smaller size and higher accuracy, a CNN-based model may be a better choice, depending on the interval between each inspection.

Judgment optimization in attendance system. As mentioned in 2.2, the previous system relied on a single check, leading to instances of misjudgment. In contrast, the proposed system utilizes multiple checks: facial recognition is conducted intermittently, and the results are recorded accordingly.

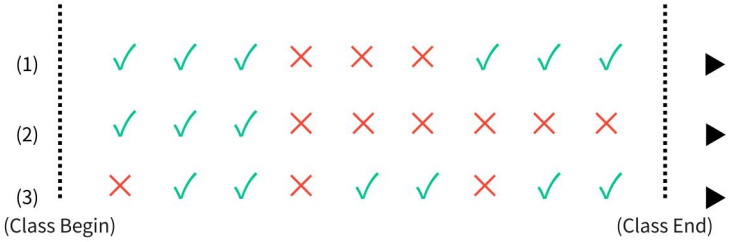

Fig. 3. Three situations during the attendance check (Picture credit: Original)

The ticks and crosses in Fig. 3 represent the check result of a student during one class.

In case (1), when a student goes to the bathroom in the middle of class and returns later, the system will judge him as attended.

In case (2), when a student attends the class at the beginning and then leaves early, the system will judge him as not attending.

In case (3), some mistakes were made during several one-checks (caused by the constrained algorithm's low accuracy/the student was blocked by something in the front), and the system will judge him as attended.

This optimized judgment strategy can prevent extreme situations and solve low-accuracy problems.

Combination of the two blocks. To make the system size smaller, only one algorithm is selected to serve students' face and action detection functions. Then, apply a facial recognition algorithm. The process is shown in Fig. 4.

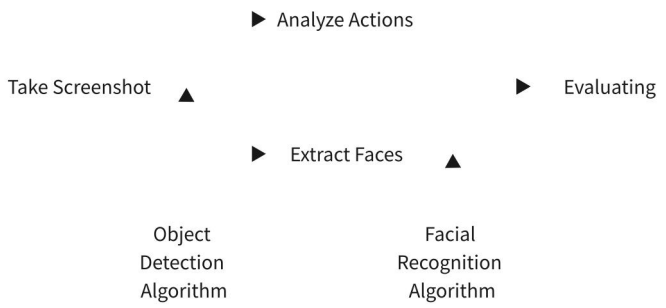

Fig. 4. The process of each check of the combined system (Picture credit: Original)

For example, use YOLO or CNN for detecting students' faces and static action at the same time. Then sperate this two kinds of results and flow into different processing blocks. In this way, the combined system will use the same number of algorithms as the original one, but with additional functions.

3 Experimental Results

The system successfully extracted the students' faces and detected the behaviors of sleeping and playing on phones in Fig. 5. Even with a large number of students in Fig. 6, the system still achieves a high recognition rate and accuracy.

The overview and accuracy of the model are shown in Table 4. Since the dataset of faces is private, this system cannot be tested in an actual situation. However, with the statistics in Table 3, the results of face matching and final judgment can be estimated. Although the recall rate and the precision rate are above 90%, with the improved judgment strategy, the final judgment accuracy reached more than 96%.

Table 4. Recognition results

Model Size	Inference speed	Confidence	Recall	Precision	Final Judgement Accuracy
49.6MB	70ms/image	0.30	84.5%	89.2%	96.8%

Even if the datasets are screenshots of the classroom monitor rather than consistent videos, the optimized judgment approach can still produce relatively accurate results.

Fig. 5. Three behaviors extracted (Picture credit: Original)

Fig. 6. A large number of people scenario (Picture credit: Original)

4 Conclusion

Through a comprehensive evaluation of existing automatic attendance and behavioral recognition systems, this study has proposed novel improvement strategies and analyzed their pros and cons. For instance, YOLOs can process video frames faster, while CNNs demonstrate greater adaptability at lower speeds. Additionally, the evaluation of various facial recognition models indicates that each performs well, with the choice depending on specific situational requirements and trade-offs.

After careful consideration of practical needs, this paper has refined and updated algorithms, leading to the development of a novel automatic attendance system integrated with behavioral recognition capabilities. In this paper's version, dynamic action detection has been omitted, maintaining a similar system size to its predecessor while incorporating an additional behavioral recognition function. Furthermore, through optimization of judgment methods, exceptional cases like students arriving late to class and leaving early have been addressed. The use of multiple detection methods has effectively mitigated the issue of lower accuracy and reduced real-time processing demands, enabling the selection of not only YOLO but also more accurate but slower algorithms like CNN. Ultimately, this system was tested on the image dataset captured by the classroom monitor, with the objective of accurately recognizing students' identities and detecting instances of students sleeping or using their phones during class. Overall, the system achieved an impressive accuracy rate exceeding 96%, while maintaining a manageable size of over 50MB.

As a result, manual class attendance processes can be streamlined, leading to more efficient classes and simplifying the task of evaluating students' coursework grades for professors. However, constraints on the confidentiality of classroom monitoring videos have limited access to datasets, restricting experiments to image-based testing only. Future research directions include recognizing faces obscured by masks or glasses and developing new encryption and anonymization technologies to safeguard collected and processed facial data. Crucially, future studies should also consider the adaptability and sensitivity of facial recognition technology across diverse cultural and social contexts worldwide.

References

1. Kortli, Y., Jridi, M., Al Falou, A., Atri, M.: Face Recognition Systems: A Survey. Sensors 20(2), 342 (2020).
2. Redmon, J., Divvala, S., Girshick, R., et al.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788 (2016).
3. Fu, Y. Y., Lin, Z., Wu, C. X., et al.: A classroom attendance intelligent recognition system combining object detection and face recognition. Electronics World 31(1), 67–70 (2024).
4. Wu, C. L., Feng, Z. W., Zhong, S. H., et al.: Automatic attendance management system for college classrooms based on face recognition technology. Intelligent Buildings and Smart Cities (2020) (5), 20–24.
5. Liu, L.: Design of student classroom behavior recognition system based on convolutional neural network. Modern Electronics Technology 47(6), 142–146 (2024).
6. Huang, Y., Liang, M., Wang, X., et al.: Multi-person classroom behavior recognition in classroom teaching videos based on deep spatiotemporal residual convolutional neural network. Computer Applications 42(3), 736–742 (2022).
7. Liu, W., Anguelov, D., Erhan, D., et al.: SSD: Single shot multibox detector. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pp. 21–37. Springer International Publishing (2016).
8. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 815–823 (2015).
9. Jocher, G.: YOLOv5 by Ultralytics (Version 7.0). https://doi.org/10.5281/zenodo.3908559 (2020), last accessed 2022/11/22
10. Lee, Y. H., Kim, Y.: Comparison of CNN and YOLO for Object Detection. Journal of the Semiconductor & Display Technology 19(1), 85–92 (2020).
11. Du, J.: Understanding of object detection based on CNN family and YOLO. In: Journal of Physics: Conference Series. IOP Publishing, 1004, 012029 (2018).
12. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLO (Version 8.0.0). https://github.com/ultralytics/ultralytics (2023), last accessed 2024/5/19
13. Wang, C. Y., Yeh, I. H., Liao, H. Y. M.: YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. arXiv preprint arXiv:2402.13616 (2024).
14. Wickrama Arachchilage, S. P., Izquierdo, E.: Deep-learned faces: a survey. EURASIP Journal on Image and Video Processing 2020(1), 25 (2020).