Multimedia Tools and Applications (2023) 82:44673–44695
https://doi.org/10.1007/s11042-023-15474-y

Intelligent invigilator system based on target detection

Jing Xue[1] · **Wen Wu**[1] · **Qingkai Cheng**[1]

Received: 22 September 2021 / Revised: 10 April 2023 / Accepted: 18 April 2023 /
Published online: 29 April 2023

Abstract
Affected by the COVID-19 epidemic, the final examinations at many universities and the recruitment interviews of enterprises were forced to be transferred to online remote video invigilation, which undoubtedly improves the space and possibility of cheating. To solve these problems, this paper proposes an intelligent invigilation system based on the EfficientDet target detection network model combined with a centroid tracking algorithm. Experiments show that cheating behavior detection model proposed in this paper has good detection, tracking and recognition effects in remote testing scenarios. Taking the EfficientDet network as the detection target, the average detection accuracy of the network is 81%. Experiments with real online test videos show that the cheating behavior detection accuracy can reach 83.1%. In addition, to compensate for the shortage of image detection, we also design an audio detection module to carry out auxiliary detection and forensics. The audio detection module is used to continuously detect the environmental sound of the examination room, save suspicious sounds and provide evidence for judging cheating behavior.

Keywords EfficientDet · Intelligent invigilation · Abnormal behavior detection · Audio analysis · Centroid tracking

1 Introduction

Affected by the COVID-19 epidemic, many universities have conducted online courses and converted final examinations into online video proctoring. [5] Popular remote invigilators fall into two main categories. One is online remote invigilation, in which one invigilator invigilates multiple students at the same time through video monitoring, and the other is offline remote invigilation, in which the examinee records his or her own examination process and sends the recorded video to the invigilator after the examination. These invigilators have some disadvantages, such as the limited number of invigilators cannot achieve full coverage of real-time monitoring, cheating is difficult

✉ Jing Xue
xuejing@njupt.edu.cn

[1] School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, China

to detect, and it is difficult to collect evidence after some cheating stops. Regardless of which of the above remote invigilation methods is adopted, it is necessary to rely on the manual judgment of whether the examinees are cheating, which undoubtedly increases the work intensity of invigilators. Therefore, it is urgent to explore an efficient and accurate intelligent invigilation method.

To solve the problems of remote video invigilation, such as the heavy workload of examiners and inaccurate judgment of cheating behavior, many scholars have explored the use of artificial intelligence technology to reform the effective mode of invigilation (related work will be introduced in the next section). Throughout the research of these scholars, they applied the related technology of computer vision to intelligent invigilation and achieved good results, but the accuracy and time efficiency of detection and the excessive dependence on hardware are still problems to be solved.

To reduce the working intensity of invigilators and improve the accuracy of detecting cheating behaviors, we propose a remote test cheating behavior detection model and audio assistance module based on the EfficientDet [16] framework, which realizes the multidimensional analysis of suspected cheating behaviors of examinees in the remote test environment.

The main contributions of this paper are as follows:

1. A model is trained based on the EfficientDet framework, which realizes the detection and location of the examiners' heads, hands, and examination terminals through the cameras carried by mobile terminals.
2. A target tracking method is proposed to identify the cheating behaviors of candidates by analyzing the moving range of the center of mass of the candidate's head and hands (Section 6.1).
3. An audio auxiliary module was designed to determine abnormalities by analyzing the audio data monitored in the remote examination environment and realized the discrimination and forensics of cheating behavior from the perspective of sound (Section 6.3).
4. The intelligent invigilator system based on the target detection network model combined with the centroid tracking algorithm is proposed. Compared with the YOLOv3 and the target detection model EfficientDet, the accuracy and speed of the data set provided in this paper improved slightly, reaching an average accuracy of 83.1%. The speed can reach 55fps. By using more refined and efficient model structure and optimization techniques, this algorithm can achieve higher accuracy and faster detection speed while maintaining high speed.

The paper is divided into six parts according to the content, and the chapters are organized as follows:

Part one: Introduction. This section describes the research background and significance of this research topic, briefly introduces the current situation of intelligent invigilation, and summarizes the contribution and structure of this paper.
Part two: Related work. This section introduces the research status of remote video invigilation worldwide, compares the advantages and disadvantages of the algorithm model used, and leads to the main research content of this paper.
Part three: EfficientDet advantages and structure. The architecture of the EfficientDet network model and the reasons for its selection are introduced.

Part four: Our proposed method. This paper introduces how to combine the object detection method and centroid tracking algorithm to realize real-time detection and early warning of examinee cheating behavior.
Part five: analysis of experimental results. The experimental environment of the experiment is carried out are introduced. The results of video analysis and audio analysis are shown.
Part six: Summary. This section summarizes the innovative application and practical significance of intelligent invigilation and some imperfections in the system functional design.

2 Related work

Chunmei Li [6] et al. designed an intelligent invigilation assistance system based on video behavior analysis. The system obtains real-time video streams from the camera, which is placed in the regular offline examination room. First, it detects the human body and books in the picture using the YOLO [2] (you only look once) algorithm. Then, it detects the candidate's face posture using the MTCNN [18] (multitask convolutional neural network) algorithm. Finally, it detects and analyzes abnormal behaviors and sends warnings. By using the system, the cheating behaviors of candidates can be detected automatically, cheating behaviors can be automatically warned, and cheating data can be saved.

Miaomiao Ding [3] et al. proposed an anomaly detection system combining a sparse reconstruction model and a motion-connected component model. The sparse reconstruction model can detect abnormal behaviors at a high speed, and the motion-connected component model can detect small abnormal movements, which makes up for the deficiency of the sparse reconstruction model. This hybrid model can be used to identify abnormal behaviors of candidates in fixed sitting posture. The recent research on computer vision and intelligent invigilation are listed in Table 1 and comparison of Precision and Speed for Target Detection Algorithms are listed in Table 2.

Zillur Rahman [14] et al. proposed an automatic wrong-way vehicle detection system from surveillance camera footage on the road in their article. By detecting vehicles from the video frame by using the YOLO algorithm, each vehicle in the target region was tracked using the centroid tracking algorithm. By combining a target detection algorithm with centroid tracking, this set of systems facilitates vehicle tracking on the road.

Liwei Yin [20] et al. designed an SDD [12] algorithm in the paper "Intelligent Invigilation System for Schools Based on Computer Vision". Through the practice of a large number of training sets in the early stage and combined with video processing technology, an intelligent invigilation system suitable for schools was realized. The detection accuracy of the system can reach 84%. However, the system also has some defects, such as a slow video frame and detection speed, and it cannot achieve real-time detection.

By accessing the data and analyzing the above literature, we can see that the current abnormal examination behavior recognition is based on identifying abnormal behaviors that are defined in advance. For example, Chunmei Li judged cheating behavior by detecting prohibited items and the candidate's facial posture, and the candidate detection model proposed by Miaomiao Ding can only detect candidates in a fixed sitting position. However, the position of the candidate in the exam room is complex and variable. When the candidate makes undefined abnormal behaviors, the system may fail to detect and save the abnormal behaviors without predefinition.

Table 1 Some recent research on computer vision and intelligent invigilation

Survey	Inference	Year	Algorithm model	Application scenarios
Intelligent invigilator assistant system based on video behavior analysis	Chunmei Li et al. [6]	2019	YOLO [2] and MTCNN [18]	Target detection and facial posture detection of candidates
Research on intelligent invigilation method based on monitoring video of examination room	Miaomiao Ding et al. [3]	2017	Sparse reconstruction model, motion-connected domain model	Test of examinee behavior
Real-time vehicle detection based on YOLO and centroid tracking	Zillur Rahman et al. [14]	2020	YOLO Centroid tracking algorithm	Detect the wrong vehicle
School intelligent invigilation system based on computer vision	Liwei Yin et al. [20]	2020	SSD [12]	Detection of cheating behavior

Table 2 Comparison of Precision and Speed for Target Detection Algorithms

	mAP	AP(IoU=0.5)	AP(0.75)	AP(S)	AP(M)	AP(L)	speed
YOLOv3 [15]	73.6	81.1	70.1	65.7	74.2	76.9	45FPS
EfficientDet	82.9	91.3	82.8	71.3	83.4	84.7	52FPS
EfficientDet-Centroid	83.1	89.8	79.6	76.6	84.3	85.2	55FPS

The above systems, which were designed for standard offline examination rooms, are not suitable for candidates who complete examinations at home alone. The centroid tracking algorithm used by Zillur Rahman gives us a new idea to judge cheating behavior. The research conducted by some scholars in the field of image segmentation also gives us some inspiration. [1] Instead of defining the cheating behavior features in advance, the cheating behaviors are identified by the movement of the candidate's head and hands.

3 Preliminary studies

3.1 Reasons for choosing EfficientDet

Many methods can be used to complete the target detection task. Since the publication of CNN [19] and RCNN [11] (Region-CNN), there have been many algorithms based on convolutional neural networks. RCNN, fast-RCNN [11] and dasher RCNN are two-stage detectors with high accuracy but slow speed. SSD [12] (single-shot multibox detector) and YOLO detectors are first-class detectors, which are fast, but the accuracy is low. Among these methods, YOLO is the most popular in real-time applications. EfficientDet is the latest target detection framework developed by Google Brain. It continues the monopole detection logic represented by YOLO, uses less computing power and parameters, and has better detection performance than YOLO. In the examination room, the situation is complex and changeable, but the speed of model detection is very high. In particular, EfficientDeT-D7 achieves SOTA 52.2 AP results on the COCO dataset on the same model and at the same scale while using only 52 M parameters and 325B FLOPS. Compared with other detectors, our model only needs 4-9 times as many parameters and 13-42 times as many FLOPS.

In general, the EfficientDet series has a simple structure, efficient model expansion, and high performance and development potential. Therefore, the EfficientDet detector with a high detection speed is selected.

3.2 EfficientDet architecture

Figure 1 shows the EfficientDet architecture, which is similar to YOLO and SSD. It uses EfficentNet as the backbone network. Level 3-7 features are extracted from the backbone network and sent to BiFPN [16] (bidirectional feature pyramid network) for weighted feature fusion. The obtained features are input into the category prediction network and location prediction network to generate the category prediction and location prediction of the target.

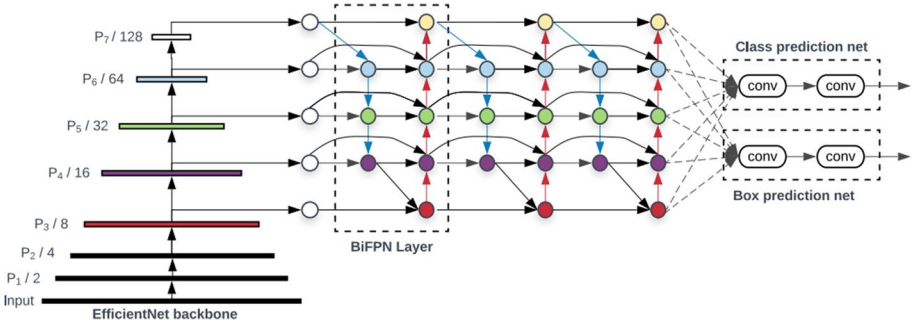

Fig. 1 EfficientDet architecture

A characteristic fusion layer of BiFPN (development history of BiFPN is shown in Fig. 2) is the key to efficiency improvement. There are many featured fusion schemes, from the early FPN [7] (feature pyramid networks), PANet [8] (path aggregation network) to NAS-FPN [4]. The BiFPN used by EfficientDet is a featured fusion scheme that can improve efficiency and accuracy rates by optimizing the network structure for feature fusion. In BiFPN, some optimization methods are used to connect scales, and only one input edge node is deleted. Additional edges are added to input and output nodes at the same level. BiFPN uses many experience-based hyperparameters, which also improves the algorithm accuracy and efficiency.

The confidence calculation formula of the IOU[21] grid prediction framework is shown in Eq. 1.

$$\begin{cases} confidence = \Pr(object) * IOU_{pred}^{truth} \\ IOU_{pred}^{truth} = \frac{area\ (pred \cap truth)}{area\ (pred \cup truth)} \end{cases} \tag{1}$$

When there is a target in the bounding box, $\Pr(object) = 1$; otherwise, $\Pr(object) = 0$. IOU_{pred}^{truth} reflects the positioning accuracy, which is obtained by the intersection of the target prediction and actual bounding box area.

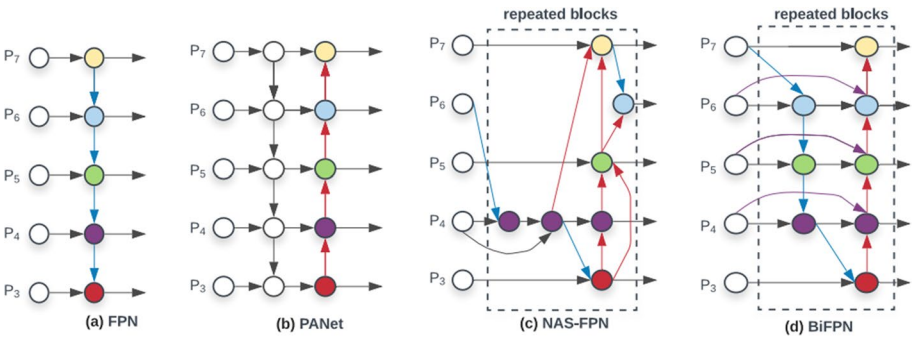

Fig. 2 Development history of BiFPN [16]

Classification loss (cross-entropy loss) adopts focal loss [17], which is used to obtain the prediction results for all species that are not ignored. However, the difference is that there is a margin between positive and negative samples, and the a priori box where IOU is in the margin range will be ignored (shown in Eq. 2).

$$\begin{cases} IOU < 0.4 & , False \\ 0.4 \leq IOU < 0.5 & , Pass \\ IOU \geq 0.5 & , True \end{cases} \tag{2}$$

Regression loss [9], which is similar to $smmoth_{L_1}$ loss (shown in Eq. 3), is used to obtain the prediction results of all positive labeled boxes.

$$smmoth_{L_1}(x) = \begin{cases} 0.5 * x^2, & |x| < 1.0 \\ |x| - 0.5, & |x| \geq 1.0 \end{cases} \tag{3}$$

4 Proposed method

To implement the proposed cheating behavior detection model, it is necessary to identify the candidate's head, hands and exam terminal contained in the video frame and to judge whether there is cheating based on the location changes in these targets. Based on the EfficientDet framework, a model is trained to detect and locate the head, hands and exam terminal, and then the centroid tracking method is used to judge cheating behaviors. To compensate for the image recognition deficiency, an audio detection module is designed to acquire and analyze the detected audio data and the saved audio data containing evidence of suspected cheating after the analysis.

The system architecture is shown in Fig. 3. This section contains the detailed operation process of the system, and details of each stage are described in the following sections.

First, we separate the video data into image data and audio data, which are input into the image analysis module and the audio analysis module, respectively. If the audio or image analysis module detects that cheating has occurred, the system issues an alert and intercepts the video or audio for forensics.where D represents the proctor video dataset, S(D) represents the audio and video separation of data D, and then the image stream V and audio stream A are obtained. E(V) indicates that the video frame target quantity information Q and the target position information L of the video frame are obtained by object detection on the image. C represents the image analysis result R_1 produced by the centroid tracking algorithm. G denotes the audio analysis module and the parameter audio analysis result R_2.

4.1 Target tracking

To implement the proposed cheating behavior detection model, it is necessary to identify the candidate's heads, hands and test terminals contained in the video frames and judge whether there are cheating behaviors according to position changes in these targets.

To determine whether there are suspected cheating behaviors in the detection screen, the system uses the centroid comparison method. This method uses the bounding box as input and EfficientDet to generate bounding boxes and then sends these bounding boxes

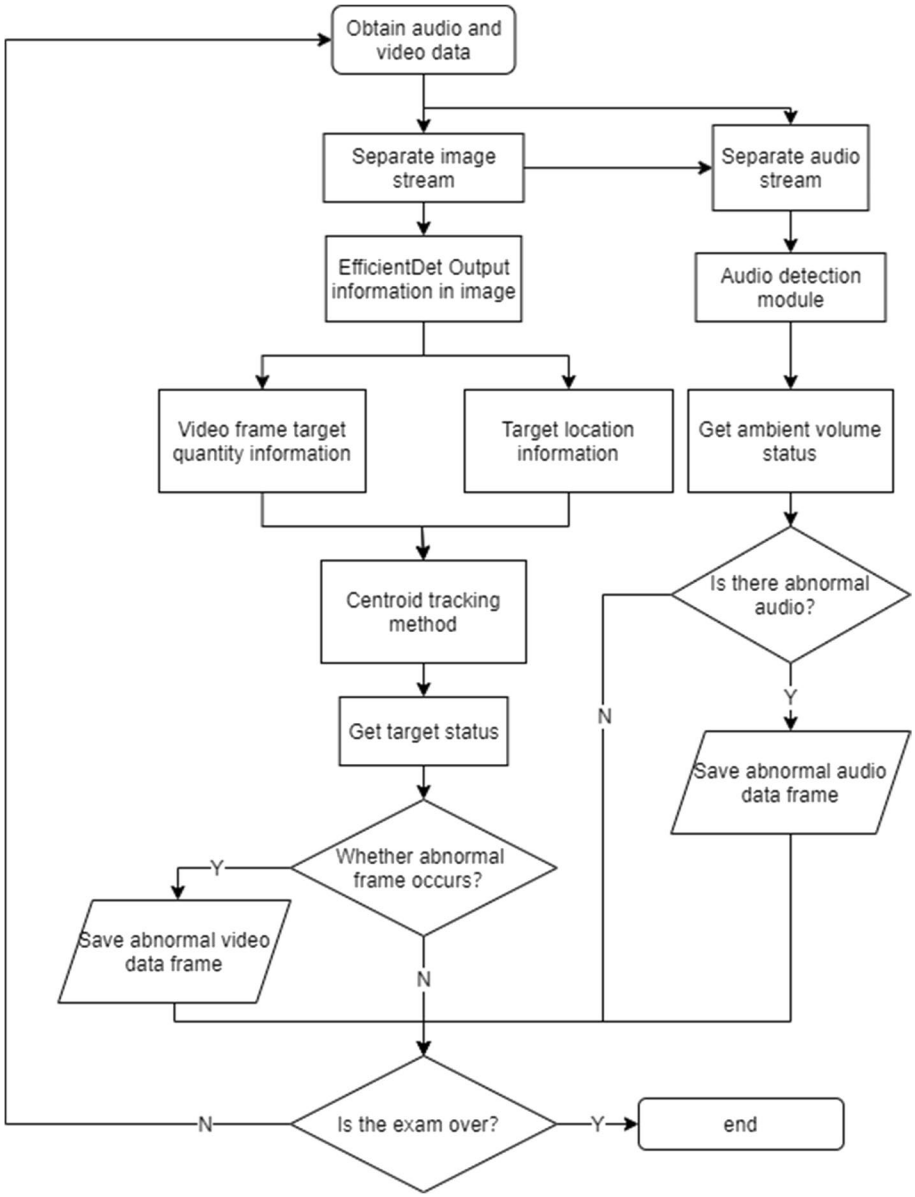

Fig. 3 The system flowchart

to the centroid container. The coordinates of each object detected in each frame are saved in memory. Taking n frames as the period, if all the position information of target a in the previous m frames exists, the average position (x_k, y_k) of target a in the previous m frames is taken as the reference position; otherwise, the average position (x_{k-1}, y_{k-1}) of target a in the previous period is taken as the reference position. The Euclidean distance [13] between the position of the target a (x_i, y_i) and the reference point (x_k, y_k) must be calculated in each subsequent cycle:

$$x_k = \begin{cases} \frac{1}{m}\sum_{i=1}^{m} x_i, & x \geq 0 \\ x_{k-1}, & \textit{missing target} \end{cases} \tag{4}$$

$$y_k = \begin{cases} \frac{1}{m}\sum_{i=1}^{m} y_i, & y \geq 0 \\ y_{k-1}, & \textit{missing target} \end{cases} \tag{5}$$

$$d_{(k,i)} = \sqrt{\left(x_i - x_k\right)^2 + \left(y_i - y_k\right)^2} \tag{6}$$

where x_k and y_k are the horizontal and vertical coordinates of the reference position, respectively, $d_{(k,\,i)}$ is the offset Euclidean distance of the target centroid in frame i, and a missing target indicates one or more missing targets in m frames.

If the position of target a in frame $I \in [I, n]$ exceeds the allowable offset threshold at the reference position (as shown in Fig. 4), the system determines that the student's examination status is abnormal. If the coordinate information of target a in the previous m frames is incomplete, for example, the coordinate information of the target in a certain frame is missing, the last reference position is taken, and the current frame state is set as target missing. As shown in Fig. 5, if the number of targets in the current frame I is fewer than that of the previous frame or there is no target, which indicates that one or more targets disappear from the screen, the examinee's examination status is judged as abnormal.

Through the above moves, we can judge whether there is a suspected cheating state in the detected video stream based on the large target movement and missing target.

4.2 Record video containing suspected cheating

Generally, cheating in a real exam is not instantaneous, and the cheating process lasts for a while. To save the detected pictures containing the candidate's suspected cheating continuously for a period of time to record continuous behaviors, the system uses the following methods.

Taking t seconds as a period, all frames within t seconds are detected successively, and the number of frames is p. If more than 85% of the frames are suspected of cheating, that is, the number of frames suspected of cheating within t seconds is $q \geq p \times 85\%$, the video recorded during t seconds is saved as a small clipping.

The variable *TIME_JUDGE* is set to record the end time of each saved video. If a saved video starts before the recorded *time_judge* time, there is no need to save it

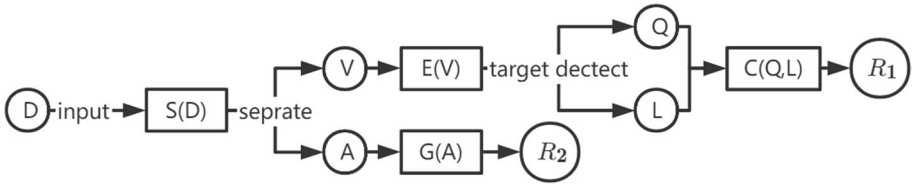

Fig. 4 System architecture diagram

Fig. 5 Target displacement

twice, since suspected cheating has been detected before that time and the video has been saved. As shown in Fig. 6.

Finally, the system synthesizes the saved independent videos into a video file for the convenience of invigilators.

In summary, this part of our work allows us to automate access to videos of candidates suspected of cheating.

4.3 Record audio with abnormal sound intensity

The sound intensity level of audio is a judgment index. There are two options for the audio sensitivity index. One is the amplitude of the sound wave (as shown in Fig. 7), and the other is the sound intensity level, which is the common DB index (as shown in Fig. 8). As

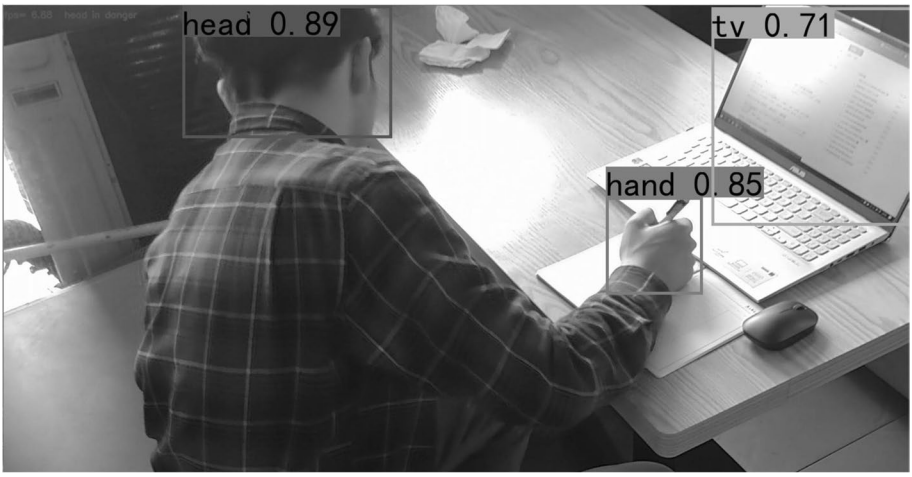

Fig. 6 Target loss

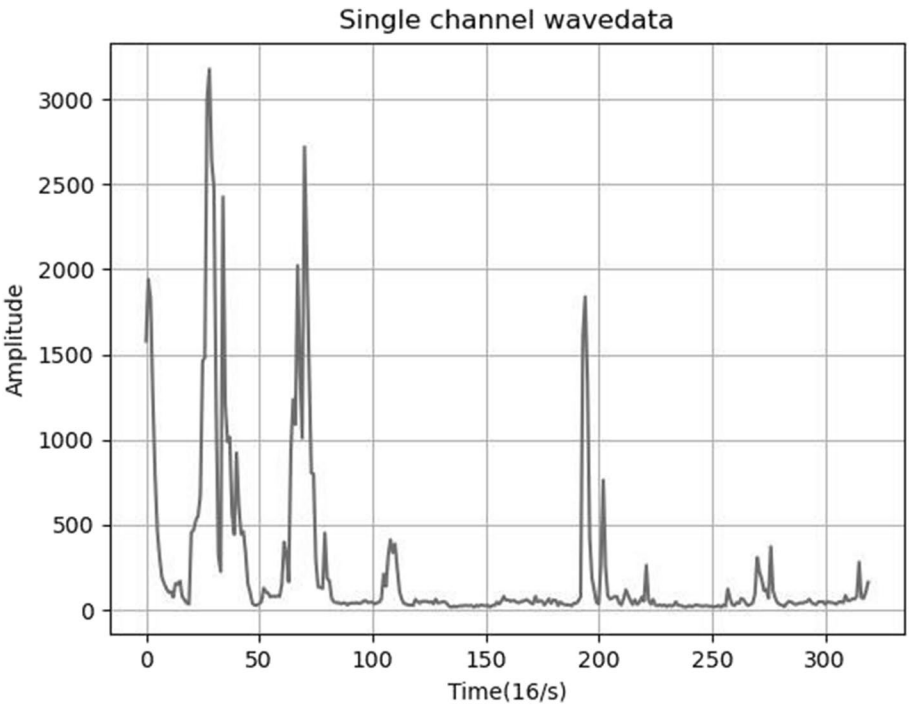

Fig. 7 Amplitude image

the voice volume of candidates is relatively small in an exam, the decibel index sensitive to bass is used. Additionally, the reference value of the sound intensity level can be easily obtained.

By comparing the two indicators, it is found that the sound intensity level is more sensitive to low volume and easier to meet design needs.

The recognition and capture of low sound can be achieved through continuous microphone detection. After repeated tests, it is found that noise in the test environment is usually higher than 30 dB, and the normal AC is higher than 50 dB. Communication during cheating is usually lower than that of normal communication. Therefore, when the surrounding volume in the test is higher than 40 dB, it is considered that the test environment is not "quiet" at this time, and capturing sound will not start until it is quiet.

The workflow of the module is shown in Fig. 9, which is the scope of the examination time. If so, return to the detection state state, the audio features captured by the microphone are continuously analyzed, and the audio data are not saved at this time. When the audio is at a high volume at a specific time, it enters the capture state. At this time, the audio data is recorded in the memory, and the recorded data is analyzed. The indicator for the end of the capture state is a blank that lasts for five seconds. As there are temporary pauses in a conversation, we cannot judge whether the audio segment ends based on the sound intensity level of the current audio. When it is detected that the audio contains bass that lasts for five consecutive seconds, the capture state of the module ends. After the capture state is over, it is time to judge whether the current time is within the scope of the examination time. If so, return to the detection state; otherwise, it ends.

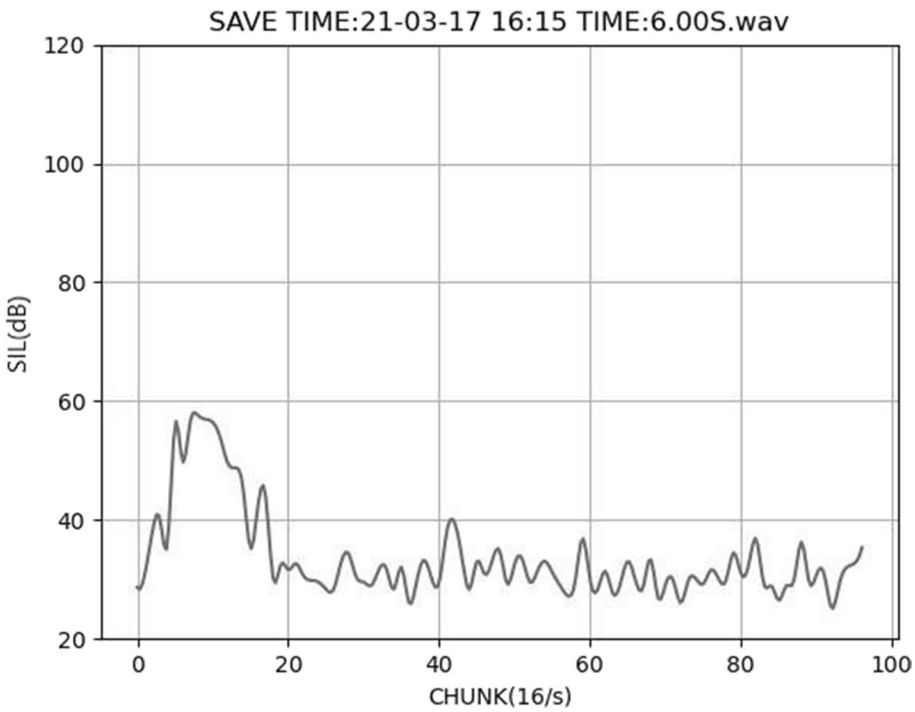

Fig. 8 Sound intensity level image

Save audio data as a file and visualize it. An image similar to Fig. 7 or Fig. 8 is obtained, which contains the general information of the audio and plays a supporting role in cheating judgment. From video to audio, the system detects whether there is cheating behavior in the same scene from multiple perspectives, which further improves the accuracy of cheating behavior detection.

5 Analysis of experimental results

5.1 Target selection and data processing

5.1.1 Target selection

By analyzing the behavior characteristics of cheating, we know that under normal circumstances, examinees should put their hands on the table and face the test paper or the test terminal. However, when an examinee exhibits cheating behavior, the examinee's hands and head move in different ranges. The purpose of this movement is easy to determine: moving the hands to obtain the cheat sheet or prepared cheat device and moving the head to find the answer on the cheat sheet or cheat device.

According to the above conclusion, we determined the detection target of the system: head, hand and test terminal.

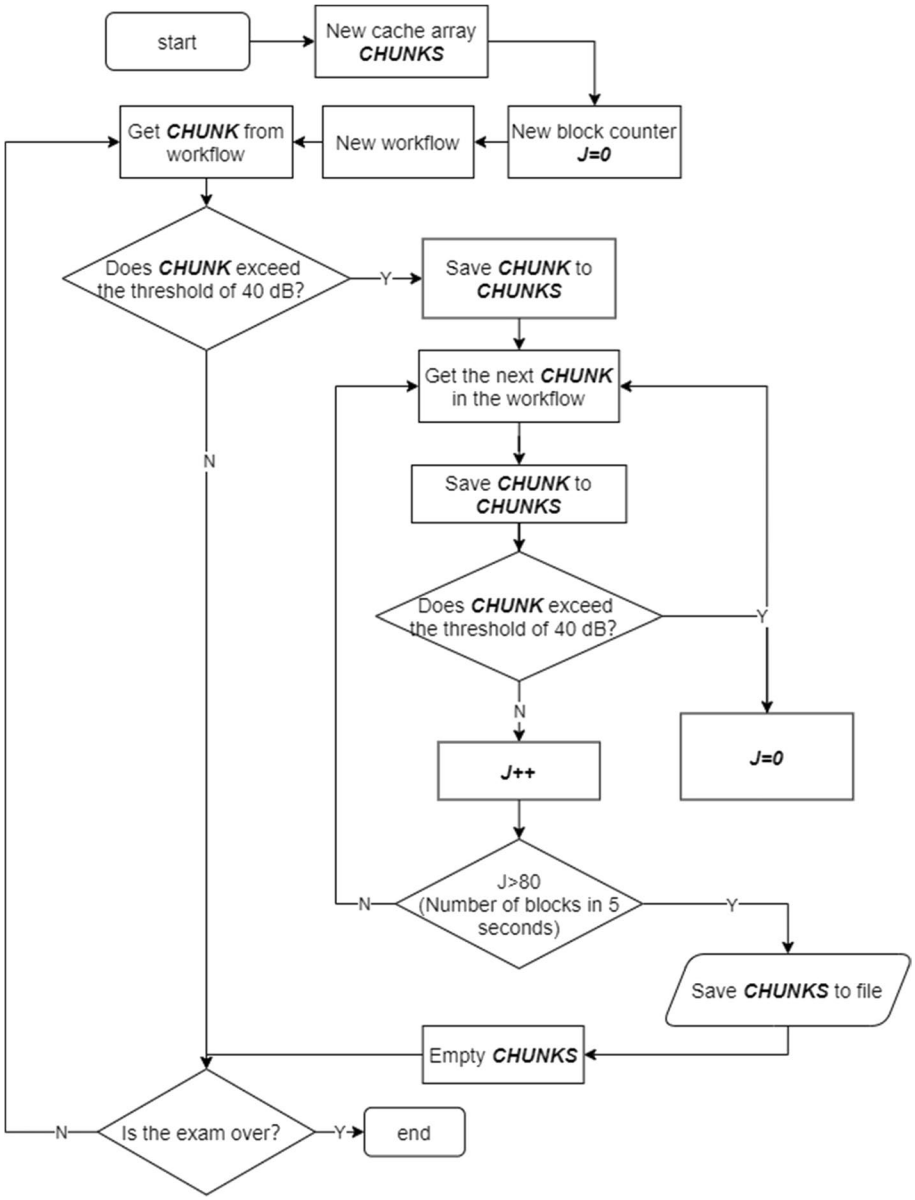

Fig. 9 Audio analysis flowchart

5.1.2 Training dataset

The Coco2017 dataset is an open source dataset. We extract images including head, mobile phone and computer using scripting code, then use LabelMe to manually annotate the relevant targets in these images, then use the script to process the generated JSON file, and finally carry out training.

To realize the target detection of the head, hands and exam terminal, it is necessary to establish the image sample library. Since there is no open source dataset labeled by both the head and hands, the dataset needs to be labeled manually. More than 2,000 pictures were selected from the COCO2017 dataset and labeled with the LabelMe tool to ensure that there were more than 1,000 object frames in each category, as shown in Fig. 10. Among them, the samples include the target sample images in various scenes.

5.1.3 Testing dataset

Due to epidemic prevention and control reasons, the final exams of some subjects in the second half semester of the 2022 academic year of Nanjing University of Posts and Telecommunications were converted to online remote video invigilation, which provided a very good test opportunity for our intelligent invigilation system. We obtained invigilation videos of 5 video examination rooms as shown in Fig. 11. Each examination room had 15 candidates. The examination length was 110 minutes, and the paper was not submitted in advance. We obtained over 8,000 minutes of video data.

Obtaining video data and labeling are very important. To test the system performance index parameter, we must manually annotate the video data; for convenience and to save labor costs and system implementation, we adopted the following solutions: artificially view the video, whenever suspected cheating behavior occurs, record the number of minutes, annotate the file generated time node.

5.2 Experimental environment

5.2.1 Training environment

Colab and Linux are the model training platform. This platform is preinstalled with the TensorFlow and Python machine learning frameworks. It has 12.72 GB ram and a 68.4 GB hard disk. The CPU is a dual-core Xeon CPU. The GPU is a Tesla P4 with 7.6 GB of video memory. The SIM of the GPU driver is 418.67, and the CUDA version is 10.1.

Fig. 10 Image library: head library; hand library; exam terminal library

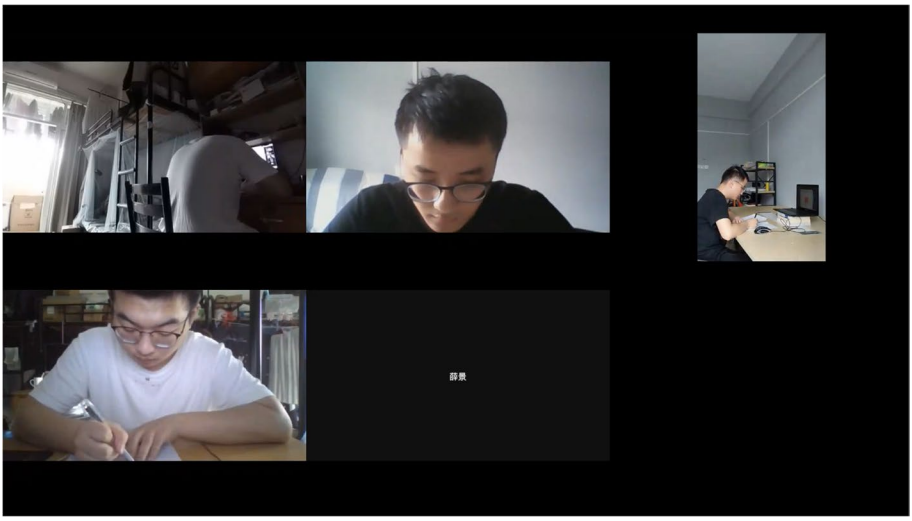

Fig. 11 A screenshot of the invigilation video

5.2.2 Testing environment

The OpenCV deep learning library and PyTorch learning framework in the Python environment are used. The model detects a frame rate of 11 FPS on a Windows system, Intel i7-1065G7 CPU, and MX350 GPU device.

5.3 Results of video analysis

After obtaining the position of the head and hand targets, it is necessary to analyze the examinee's behavior to judge whether the examinee is suspected of cheating. In the saved video data, it is found that it is impossible to judge whether there is cheating by using a fixed method due to the examinee's changeable actions during the examination. Taking Fig. 12 as an example, the target displacement amplitude is used to judge whether there is suspected cheating.

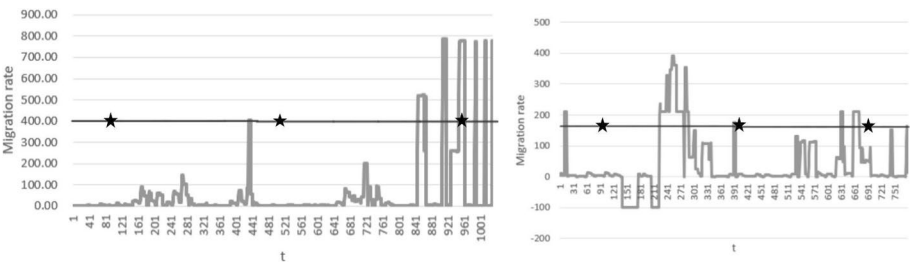

Fig. 12 Target displacement amplitude

The horizontal axis is time t, the vertical axis is the offset degree $migration\_rate$ of target a relative to the benchmark position, and the red line represents the threshold range of allowable deviation d_{max}. When the offset size of target a exceeds the threshold, the candidate is suspected of cheating.

Take the candidate's head as an example. The coordinates of the candidate's head center point si (x_s, y_s). When the candidate answers questions normally, the coordinate offset of the head center point is small. When the candidate displays cheating behavior, the coordinate value of the head center point may have a large offset.

Taking the head as an example, we define a box; if the leader of the examinee moves in this box, it is considered normal movement; otherwise, it is judged as suspected cheating. L and R are the left limit and the right limit of this box, and U and D are the upper limit and the lower limit of this box.

The criteria are as follows:

$$\begin{cases} \mathrm{L} < x_{s,t} < \mathrm{R} \\ \mathrm{U} < y_{s,t} < \mathrm{D} \end{cases} \tag{7}$$

where t is the head, hands or the exam terminal, s is the frame number, $x_{s,\,t}$ is the abscissa of the target t and $y_{s,\,t}$ is the ordinate of the target. L and R are the left limit and the right limit of this box, and U and D are the upper limit and the lower limit of this box.

To judge suspected cheating in the examination, take the hand as an example, as shown in Fig. 14, and judge whether there is suspected cheating by obtaining the coordinate position of the center point of the examinee's hands:

When the examinee's hand coordinates are basically unchanged or the movement range is very small and kept within the allowable offset range, the examinee's examination state is considered normal;

$$\sqrt{\left[\left(x_{s,t} - \frac{1}{m}\sum_{i=1}^{m} x_{i,t}\right) + \left(y_{s,t} - \frac{1}{m}\sum_{i=1}^{m} y_{i,t}\right)\right]^2} < d_{max} \tag{8}$$

where s is the frame number, $x_{s,\,t}$ is the abscissa of the target t and $y_{s,\,t}$ is the ordinate of the target.

Of course, if the target disappears in the surveillance video, the system determines that the candidate is suspected of cheating.

The function flow of the module is shown in Fig. 13.

Several test materials are prepared for the video model experiment. The video is shot at the right rear of the examinee, and the resolution is 1920*1080. Two videos showing cheating were captured from the actual exam video. Video 1 contains the cheating behavior of an examinee who has taken out the prepared note. Among them, the displacement of one hand of the examinee first exceeds the threshold, disappears briefly and then appears in the picture. Video 2 contains cheating by an examinee who bent to pick up an item on the ground. The examinee's hands disappear in the detection screen for a period of time, and the head displacement exceeds the threshold.

As shown in Fig. 14, the system successfully detects the key detection targets in the normal examination state. After detecting and processing of the test video by the system, two videos with suspected cheating are obtained, as shown in Figs. 15 and 16. The cheating behavior in the video is successfully detected.

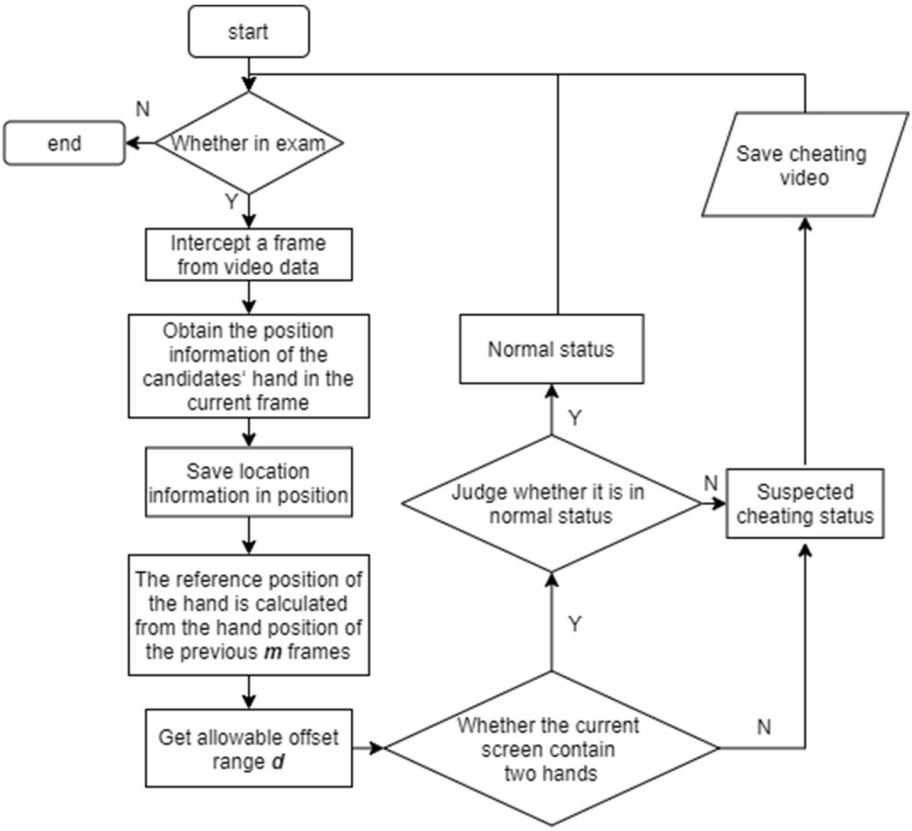

Fig. 13 Video analysis module

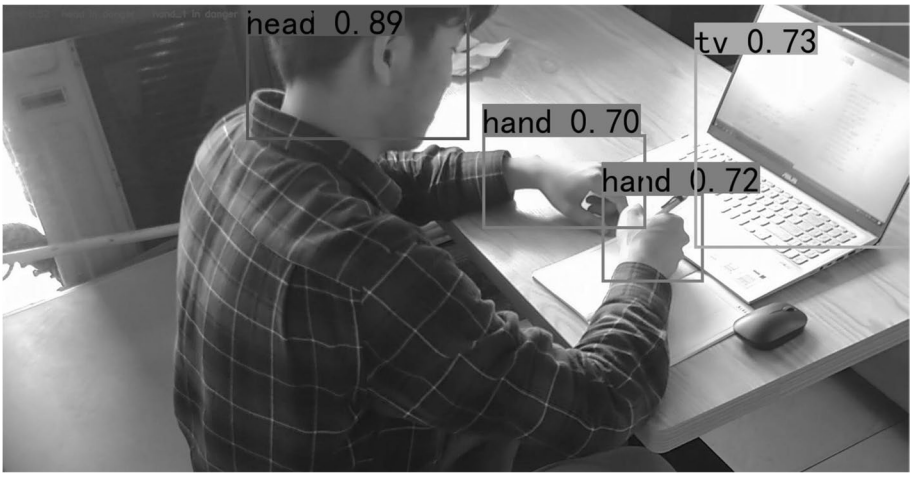

Fig. 14 Normal examination

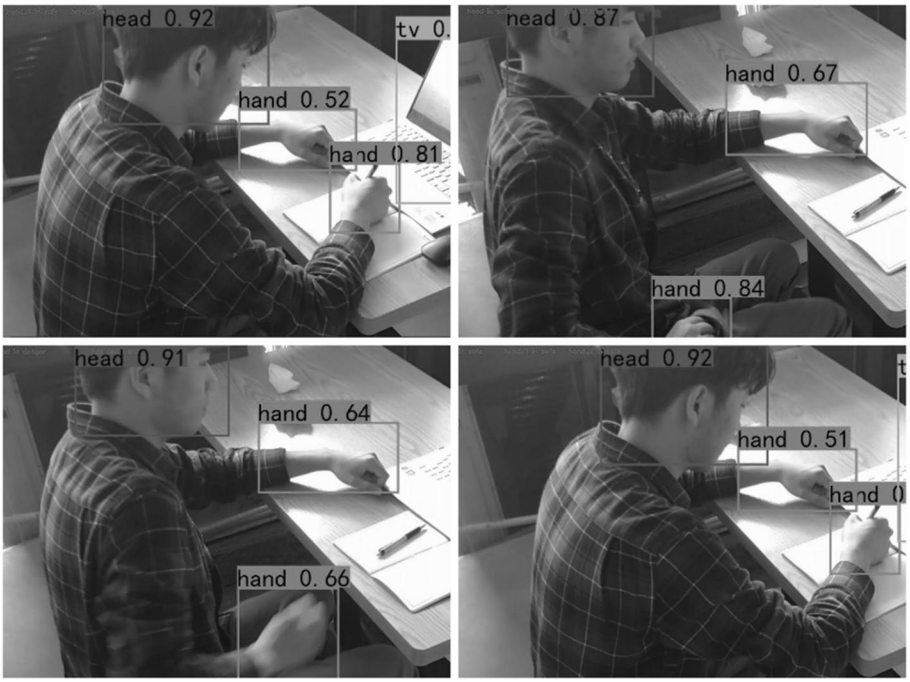

Fig. 15 Target displacement

Fig. 16 Target loss

5.4 Results of audio analysis

A separate audio detection module is designed to detect the cheating behavior of candidates communicating with people outside the picture while their limbs are still. The open source framework PyAudio is used to obtain audio data. The audio sampling rate is 32,000 Hz, 16 blocks per second. Monosampling is adopted with the sampling format of paint16. The data information is encapsulated into a WAV format file.

There are several abnormal data feature images in the saved audio data as shown in Figs. 17 and 18.

$h1$ is a single peak audio segment with abnormal sound intensity, which needs to retain the data from the beginning to the period of five seconds after the end. $h2$ is a multipeak audio segment with abnormal sound intensity, which needs to save the data from the beginning to the period of five seconds after the end of the last peak. L is the threshold value of sound intensity selected through the saved audio data. The audio segment exceeding this value is regarded as an abnormal sound intensity, which needs to be monitored continuously and saved.

The sound intensity formula is used for audio analysis:[10]

$$\frac{S}{L} = 10 * \log_{10}\left(\frac{I}{I_0}\right) \tag{10}$$

In the formula, I is sound intensity, $I_0 = 10^{-12}$ watts/square meter is the reference sound intensity, and the common unit of sound intensity is a decibel (dB).

Conditions for entering the capture state:

$$t < t_{end} \ \& \ data_{chunk} > data_{min} \tag{11}$$

Conditions for entering the detection state:

$$t < t_{end} \ \& \ num_{chunk} > 16 * s \tag{12}$$

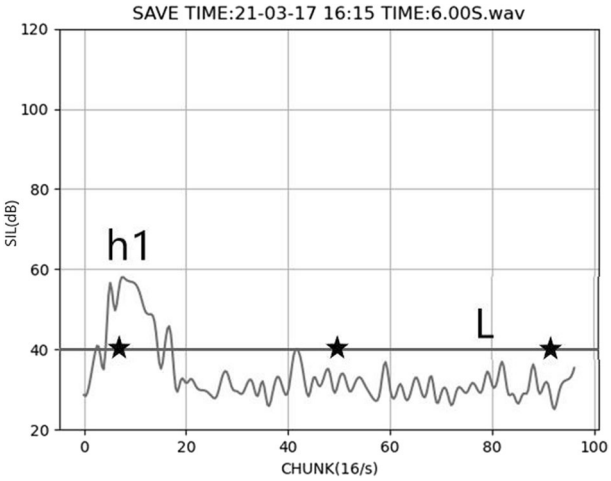

Fig. 17 Abnormal data feature A

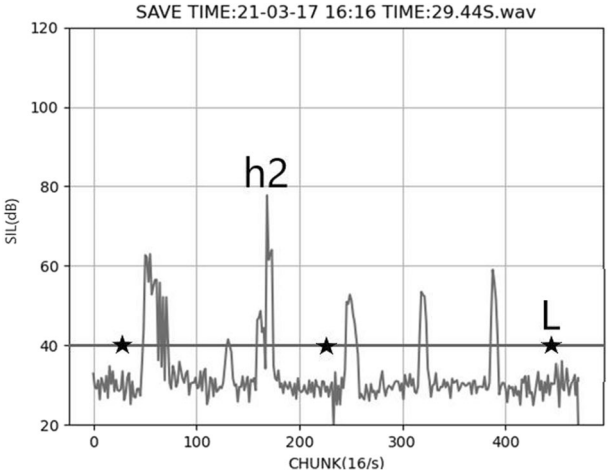

Fig. 18 Abnormal data feature B

In the formula, t is the current time, and t_{end} is the end time of the exam. $datd_{chunk}$ is the sound intensity of the block, and $data_{min}$ is the detection threshold. $datd_{chunk}$ represents the block counter, and s represents the length of time the low volume lasts.

Several audio materials of different scenes are prepared. Among them, audio material 1 is a continuous dialog. Audio material 2 is a quiet period of time, followed by a conversation.

As shown in the figure, two visual audio images and two audio data files are obtained. Figure 19 is the analysis result of audio 1. An audio segment is intercepted from audio 1. Figure 20 is the analysis result of audio 2. The data of quiet sound in the beginning are not recorded. The introduction of audio analysis makes up for the deficiency of video cheating behavior detection and improves the accuracy of the system.

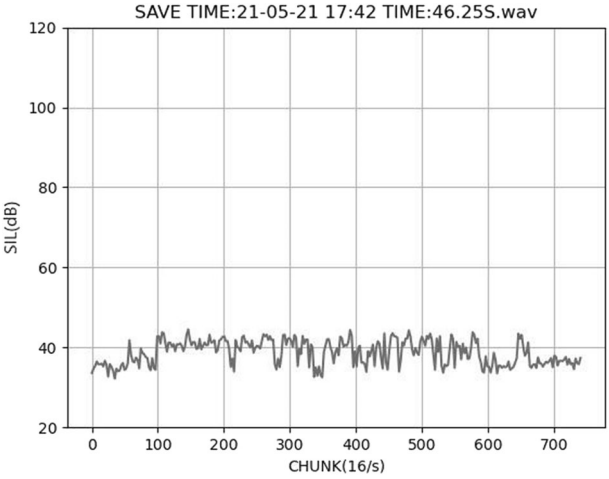

Fig. 19 result of audio 1

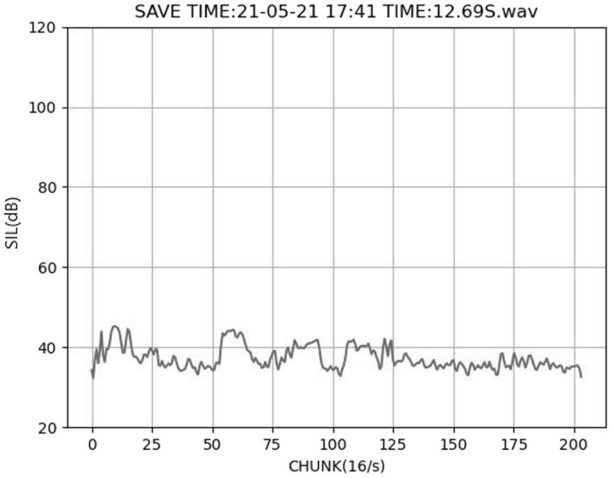

Fig. 20 result of audio 2

5.5 Experimental results

Overall, the detection accuracy and recall rate of this system are relatively ideal and can basically meet the requirements of intelligent invigilators in schools at this stage.

The final experimental results show that the detection accuracy of the system can reach 83.1%, and the recall rate can reach 87.4%.

The dataset annotation for the above experiments was done by Wen Wu, Qingkai Cheng.

The experimental results of this study are based on a self-annotated dataset and demonstrate good performance. To showcase the excellent performance of the EfficientDet target detection network model combined with a centroid tracking algorithm in the context of educational exam proctoring, we replaced the target detection algorithm in the model with YOLOv3 and EfficientDet, respectively, and compared the average precision and detection speed. The results are shown in the table below, and we refer to the improved algorithm proposed in this study as EfficientDet-Centroid.

According to the data in the table, it can be seen that the proposed model in this study achieved slightly better precision and speed compared to the YOLOv3 and EfficientDet algorithms in the specific context of educational exam proctoring, indicating the success of our improvements.

6 Conclusions and future work

This paper proposes a cheating behavior detection model and audio assistance module based on the EfficientDet framework. First, the model is trained using the EfficientDet framework to realize the detection and location of the head, hand and test terminal. Second, the centroid tracking method is used to track the target, the displacement is used to judge whether there is suspected cheating, and a warning is made. In the audio auxiliary module, abnormal sounds can be identified and captured by continuous detection of the microphone. When the volume of the test environment is greater than 40 dB, the test environment is not quiet, and the sound is recorded until the environment becomes quiet. In a real test environment, the cheating behavior detection precision rate of our intelligent

invigilation system reached 83.1%, and the recall rate reached 87.4. It can realize real-time cheating behavior detection and reduce the work intensity of invigilators.

However, our cheating behavior detection system is still inadequate. There is a certain probability of misjudgment, and the small movement range of cheating behavior detection is not sensitive and cannot capture the change in the candidate's microexpressions and eyes when cheating. We hope that in future research, we can adopt an updated algorithm to train the model with more real data to achieve a better cheating behavior detection effect.

Acknowledgment This work is supported by Instructional Reform Item of Nanjing University of Posts and Telecommunications (Grant No. JG00421JX75) and Innovative Research Group Project of the National Natural Science Foundation of China (Grand No. 61906098).

The authors would like to appreciate all anonymous reviewers for their insightful comments and constructive suggestions to polish this paper in high quality. (Haowei Zhang collected and processed the data; Haowei Zhang and Wen Wu performed the experiments and conducted the analyses; All authors agree with the above contribution details.)

Data availability The raw data required to reproduce these findings cannot be shared at this time as the data may contain privacy infomation of students in Nanjing University of Posts and Telecommunications.

Declarations

Conflict of interests The authors declare that there is no conflict of interests regarding the publication of this article.

References

1. Chakraborty R, Verma G, Namasudra S (2021) Ifodpso-based multi-level image segmentation scheme aided with masi entropy[J]. J Ambient Int Human Comput 12(1):1–19
2. Deepa R, Tamilselvan E, Abrar ES, Sampath S (2019) Comparison of Yolo, SSD, Faster RCNN for Real Time Tennis Ball Tracking for Action Decision Networks[J]. Int Conf Adv Comput Comm Eng
3. Ding M (2017) Research on Intelligent Monitoring Method Based on Examination Surveillance Video [D]. University of Science and Technology of China
4. Ghiasi G, Lin T, Le Q (2019) NAS-FPN: Learning Scalable Feature Pyramid Architecture for Object Detection[C]. Proceed IEEE/CVF Conf Comput Vision Patt Recogn:7036–7045
5. Gopal R, Singh V, Aggarwal A (2021) Impact of online classes on the satisfaction and performance of students during the pandemic period of COVID 19[J]. Educ Inform Technol:1–25
6. Li C, Shao X, Liu L (2019) Intelligent Invigilation Auxiliary System Based on Video Behavior Analysis[J]. Technol Innov Appl 18:8-10
7. Lin T, Dollar P, P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature Pyramid Networks for Object Detection[J]. Computer Vision and Pattern Recognition[C]. Proceed IEEE Conf Comput Vision Patt Recogn:2117–2125
8. Liu S, Qi L, Qin H, Shi J, Jia J (2018) Path Aggregation Network for Instance Segmentation[C]. Proceed IEEE Conf Comput Vision Patt Recog:8759–8768
9. Liu H, Yang X, Liu H, Kong T, Sun F (2020) Near-duplicated Loss for Accurate Object Localization[J]. IEEE 7th Int Conf Data Sci Adv Anal
10. Makarewicz R, Gołębiewski R (2016) Estimation of the long term average sound level from hourly average sound levels[J]. Applied Acoustics:116–120
11. Manana M, Tu C, Owolawi PA (2018) Preprocessed Faster RCNN for Vehicle Detection[J]. Int Conf Intell Innov Comput Appl:416–419
12. Morera Á, Sánchez Á, Moreno AB, Sappa ÁD, Vélez JF (2020) SSD vs. YOLO for Detection of Outdoor Urban Advertising Panels under Multiple Variabilities[J]. Image Sensors: Syst Appl
13. Morin L, Gilormini P, Derrien K (2020) Generalized Euclidean Distances for Elasticity Tensors[J]. J Elastic 138:221–232
14. Rahman Z, Ami AM, Ullah MA (2020) A Real-Time Wrong-Way Vehicle Detection Based on YOLO and Centroid Tracking[J]. IEEE. Region 10 Symposium (TENSYMP)

15. Redmon J, Farhadi A (2018) YOLOv3: An Incremental Improvement. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7263-7271)
16. Tan M, Pang R, Le Q (2020) EfficientDet: Scalable and Efficient Object Detection[C]. Proceed IEEE/CVF Conf Comp Vision Patt Recog:10781–10790
17. Weber M, Fürst M, Zöllner JM (2020) Automated Focal Loss for Image based Object Detection[J]. IEEE Intell Vehicles Symp:1423–1429
18. Xiang J, Zhu G (2017) Joint Face Detection and Facial Expression Recognition with MTCNN[J]. Int Conf Inform Sci Control Eng
19. Yanagisawa H, Yamashita T, Watanabe H (2018) A Study on Object Detection Method from Manga Images using CNN[J]. Int Workshop Adv Image Technol
20. Yin LW, Wang H, Lei Y (2020) Computer vision-based school intelligent invigilation system [J]. Int Things Technol 10(12):3
21. Zhao J, Li C, Xu Z, Jiao L, Zhao Z, Wang Z (2021) Detection of passenger flow on and off buses based on video images and YOLO algorithm[J]. Multimedia Tools Appl