International Journal of Sensors and Sensor Networks
2021; 9(1): 30-37
http://www.sciencepublishinggroup.com/j/ijssn
doi: 10.11648/j.ijssn.20210901.15
ISSN: 2329-1796 (Print); ISSN: 2329-1788 (Online)

People's Fast Moving Detection Method in Buses Based on YOLOv5

Zhang Xiaoping[1, 2], Ji Jiahui[1, 2], Wang Li[1, 2, *], He Zhonghe[1, 2], Liu Shida[1, 2]

[1]School of Electrical and Control Engineering, North China University of Technology, Beijing, China

[2]Beijing Key Laboratory of Urban Road Intelligent Traffic Control Technology, North China University of Technology, Beijing, China

Email address:
zhangxiaoping369@163.com (Zhang Xiaoping), JiJiaHui10@163.com (Ji Jiahui), wangli939@ncut.edu.cn (Wang Li), zhonghehe@ncut.edu.cn (He Zhonghe), lsdshiwo@hotmail.com (Liu Shida)

[*]Corresponding author

To cite this article:
Zhang Xiaoping, Ji Jiahui, Wang Li, He Zhonghe, Liu Shida. People's Fast Moving Detection Method in Buses Based on YOLOv5. *International Journal of Sensors and Sensor Networks*. Vol. 9, No. 1, 2021, pp. 30-37. doi: 10.11648/j.ijssn.20210901.15

Received: April 26, 2021; **Accepted**: May 10, 2021; **Published**: May 20, 2021

Abstract: To ensure the public's safety such as in buses, it is very important to accurately judge people's behaviors and give early warnings. If by watching the video surveillance manually, the cost will be very high, and it cannot be effectively popularized, so video automatic monitoring is preferred. For buses, its environmental space is closed as well as narrow, and at the same time, it is often in a non-stationary state, so traditional behavior detection methods cannot be used here as they are easily affected by moving environment and difficult to fulfill object behavior identification in real time. Aiming at this problem, for people's fast-moving in buses, a kind of detection method based on YOLOv5 is proposed in this paper. Firstly, the method detects people through one-stage object detection. Secondly, in order to obtain the person's movement trajectory quickly and accurately, an improved two-stage object matching algorithm is designed to track different people. Then, the speed curves of a person during normal activities and fast moving are compared. Finally, an abnormal alarm mechanism is constructed to realize the effective fast movement alarm. Surveillance video in the bus was used to test and evaluate the effectiveness of the method. Results show that the accuracy rate of our method can get 95.4%.

Keywords: Behavior Detection, Fast moving, Video Surveillance, Object Detection, Object Tracking

1. Introduction

People's safety in public places is very worthy of attention. For this reason, intelligent video surveillance system has been widely used in various public places, such as banks, hospitals, campuses [1-3]. For such places, the cameras are usually firmly fixed somewhere, and the environments are relatively simple. Public transportation is another situation that greatly affects people's safety. However, the traditional monitoring methods are usually not useful here for some reasons, such as the bus is moving and not stable, its environment is crowded, the composition of the personnel is complex， and so on. Therefore, we need to find new ways for public security in buses. In a narrow and closed bus, passengers' walking, standing and sitting are considered to be normal behaviors. When there is an emergent, the personal safety of passengers is threatened, and then the passengers will run away because of fear or panic. As a result, passengers' moving speed is a basis for us to judge the emergency. Effective intelligent video monitoring methods can help reduce the accident risk and speed up emergency response, so as to guarantee public's safety.

Human behavior recognition task generally follows these steps [4]. First, detect and locate human bodies in the video. Secondly, track multiple objects and match people in different frames by tracking algorithm. Thirdly, extract the features of human body region to describe the current behavior. Finally, realize the recognition of human behavior. For moving objects in the scene, the traditional detection methods include frame difference, background subtraction and optical flow [5-7]. The characteristics and application environments of these methods are different. Although frame difference and background subtraction have simple principle and good detection effect in most scenes, they are easily affected by light, cameras' movement and change of

background. Therefore, these methods are not suitable for moving scenes as in buses. In dynamic scene, optical flow method can be used to obtain the motion information of the objects because it does not need to pay attention to the background. However, due to the complexity of optical flow field calculation, the algorithm has large amount of calculation, poor real-time performance and is sensitive to light as well as noises.

Traditional behavior detection methods can achieve good results in the environment with simple background, but it is often not effective in the case of real-time individual behavior detection with crowded personnel and many interference factors. In order to solve the problems of behavior recognition in such special scenes, this paper proposes a new kind of people's fast moving detection method in buses based on YOLOv5. The overall structure of the method is shown in Figure 1. The object detection algorithm based on YOLOv5 is used to detect people in the surveillance video and obtain the accurate position information and image information [8]. Based on the deep learning model and the tracking strategy, a two-stage object matching algorithm is constructed, which combines the advantages of different algorithms to meet the application requirements [9-11]. In this part, the bounding box obtained by the detection algorithm is matched to achieve multi-object tracking, so as to accurately obtain the real-time changes of people's position. The method in this paper analyzes and measures the internal spatial structure of the bus, obtains the relationship between the image pixel and the actual distance, calculates the moving speed of the people objects, sets the trigger alarm mechanism, and finally outputs alarm signals.

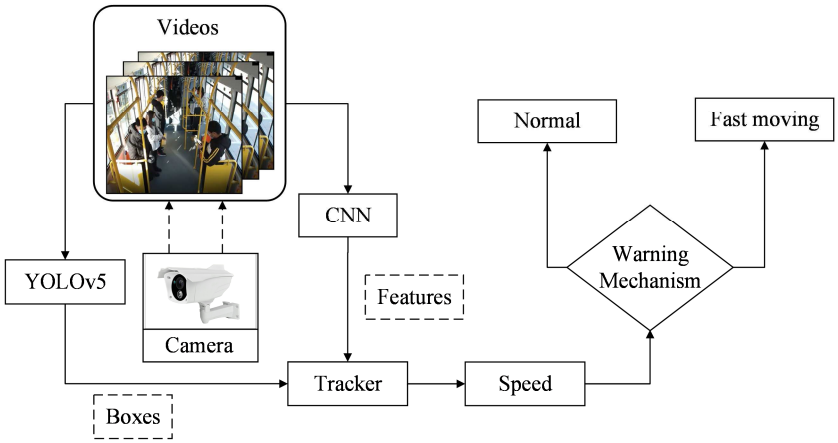


Figure 1. *The overall structure of the method.*

2. People's Detection and Tracking

The main problems of behavior recognition in buses are as follows:

a) It is difficult to obtain accurate information of people in the bus as it is crowded.

b) Most tracking algorithms are difficult to meet the requirements of rapidity and accuracy simultaneously in buses.

c) There are many interference factors in the detection process, such as the light or the cameras' jitter.

The scenes in buses are complex and variable. Existing methods have different limitations. Therefore, it is necessary to design appropriate methods to ensure the accurate recognition of behaviors. In this paper, object detection methods based on deep learning are used, which can automatically extract the depth features of images, and achieve relatively accurate object classification and positioning.

2.1. People's Detection Based on YOLOv5

At present, object detection algorithms based on deep learning are mainly divided into two categories: two-stage detection algorithm and one-stage detection algorithm. In contrast, two-stage algorithms often perform better in the accuracy of recognition and location, while one-stage algorithms often have more advantages in the detection speed. The detection task in the bus has higher requirements for the real-time performance of detection algorithm, so one-stage detection algorithms are the first choice. As a representative one-stage object detection algorithm, YOLO series of algorithms have been widely used [12, 13]. These methods directly extract features from the original images and predict the categories and positions of objects by regression analysis. After many times of optimization and improvement, YOLOv5 now has high real-time performance and accuracy. The structure of YOLOv5 is composed of four parts: input, backbone, neck and prediction [8].

a) The input includes Mosaic data enhancement, adaptive anchor box calculation and adaptive images scaling. Mosaic combines four training images, which makes each training sample more diverse and reduces the demand for large batch size. In the series of YOLO algorithms, there are anchor boxes with different initial length and width for different data sets. In the training phase, the network predicts the bounding box based on the initial anchor box, compares the prediction with the ground truth, calculates the gap between them, updates the network in reverse, and adjusts the network parameters

through iteration. YOLOv5 embeds the function of adaptive anchor box calculation into the code, sets the initial anchor box according to the COCO data set, and automatically adjusts the best anchor box in different data sets through training. In practical application, the length-width ratio of the input image is different. According to traditional ways, the original image is uniformly scaled and filled to a standard size, resulting in different border sizes around the image. If the image is filled in a large area, there will be information redundancy, which will affect the reasoning speed. Therefore, YOLOv5 is modified to adaptively add the least black edges to the original image, so as to reduce the amount of calculation in reasoning, and improve the detection speed.

b) Backbone includes Focus and CSP (Cross-Stage-Partial connections) [14]. Slicing is an important part of Focus. For example, the original image size of the input is 608 × 608 × 3, the slice operation is to obtain the feature map, whose size is 304 × 304 × 12. After a convolution operation with 32 cores, it becomes 304 × 304 × 32. In order to reduce the amount of computation, YOLOv5 designs two structures based on CSP, where in YOLOv5s, CSP1_X is applied to backbone, and CSP2_ X is used in the neck.

c) Neck adopts the structure of FPN (Feature Pyramid Networks) and PAN (Path Aggregation Network) to enhance the ability of feature fusion [15, 16]. FPN uses a top-down path to connect high-level features with low resolution as well as rich semantic information, and low-level features with high resolution as well as less semantic information. It can improve the feature extraction ability of the network, and fuse multi-level features without affecting the speed, so that each level of features has rich semantic information. In traditional serial CNN models, the shallow network is mainly used to extract the local texture and pattern information, and transmit it to the deep network, so as to obtain the global semantic information. With the deepening of the network, local information is likely to be lost. Therefore, PAN adopts a bottom-up pyramid and uses element-wise max to fuse the information of all layers.

d) Prediction includes the loss function of bounding box and the non-maximum suppression (NMS). Generalized Intersection over Union (GIOU) is the loss function of the bounding box. In the post-processing of object detection, the method of weighted NMS is used to filter many bounding boxes in YOLOv5.

There are four network models in YOLOv5. The depth and width of various models are different. Among them, YOLOv5s is the network with the smallest depth as well as width of feature map, and has the fastest speed. At the same time, it is suitable for real-time detection of large targets. In the task of identifying abnormal behavior of people in buses, YOLOv5s can effectively detect people. Using a large number of people images to train the deep neural network model can obtain a better people object detection model. The model can quickly detect the position coordinates of people's body in the image, and the confidence of people's category in the application stage. Due to the large number of people in the bus, there will be many detected objects, so the object detection confidence screening mechanism is set here. Multiple thresholds are set to filter the detected bounding boxes, so that the qualified boxes can be listed, which will be matched for people tracking.

2.2. Object Tracking

In the fast-moving behavior recognition task, most of the recognition errors are caused by the object matching errors, so fast and accurate matching algorithm is very important for effective tracking. A two-stage object matching algorithm is designed in this paper to meet above requirements.

In this section, based on the improved IOU (intersection-over-union) algorithm, and combined with feature extraction, a two-stage object matching algorithm is constructed to realize the bounding boxes matching. In the first stage, calculate the IOU of two adjacent frames and compare their coincidence degree. If the discrimination requirements are met, the two objects can be regarded as the same. After matching, the position information and number of each object can be obtained. However, if the coincidence degree does not reach the threshold that set in the first stage, the object matching fails and the process enters the second stage. This tracking algorithm is based on the information between adjacent frames. If the object changes too fast between frames, it will affect the reliability. Higher frame rate can weaken the influence of objects' changes in inter frames to a certain extent. But it is difficult to obtain high frame rate video in real-time applications. At the same time, the feature differences between tracking objects are ignored. These all lead to the decline of accuracy and practicability of IOU tracker. Therefore, it is necessary to use feature measure to match the object to make up for the shortcomings of the IOU algorithm. In the second stage, we use the algorithm to extract pedestrian features from video, and measure their similarity to achieve accurate object matching [9]. The details of the algorithm are as follows.

The principle of the first stage tracking algorithm is matching according to the IOU coincidence degree of the boxes. A total of F frames of image are input each time, and each frame contains N detections. First, initialize the parameters, set the tracking flag and eliminate the detections with low scores. Then, detect each activated detection in the previous frame, and find the detection corresponding to the largest IOU in the current frame. σ_{IOU} is used to limit whether one detection can be added to the trajectory of the target or not. The unmatched detections of the current frame will wait for the next processing. Finally, the trajectory must satisfy two conditions: one detection score is higher than σ_h, and the trajectory time is not less than $t_{\min}$. Where σ_h is used to ensure the practicability of trajectories, and $t_{\min}$ is used to filter out short trajectories. The calculation formula of IOU is as follows:

$$IOU(x,y)=\frac{Area(x)\cap Area(y)}{Area(x)\cup Area(y)} \tag{1}$$

$Area(x)$ represents the area where the object x is located, and $Area(y)$ represents the area where the object y is located. The IOU tracker is a fast and efficient multi-object tracker. This tracking algorithm based on detection can achieve object tracking with the location information of the object area and without the original image. This way can achieve good tracking effect in the case of high frame rate video and high-precision detector. Although the method works well in most cases, there are still a few cases of object loss and tracking error. When the object matching cannot be achieved in the first stage, the algorithm enters the second stage. In this case, the tracking can be achieved by measuring the similarity between the features of the current objects and those objects to be matched. The depth feature of each passenger is obtained by convolution neural network, and cosine distance is used to measure the similarity of features. The similarity value is between 0 and 1. The closer the value is to 1, the more similar the two objects are. If the similarity is greater than the similarity threshold σ_{sim}, the current object and the object to be matched are identified as the same one. If no one in the list can match the activated object, it will be regarded as a new object and added to the next object list. After many times of comparison, if there is still an object that is not matched, the object will be considered lost, and will be deleted from the list that to be matched. The feature similarity used in our approach is defined as

$$\text{sim}(V_o, V_m) = \frac{V_o \cdot V_m}{|V_o| \times |V_m|} \tag{2}$$

Where V_m is the feature vector of the object to be matched, and V_o is the feature vector of the current object. Compared with IOU tracker, our tracking algorithm adopts highly modular design and adds feature measurement mechanism. The feature extraction model can support a variety of training methods, which makes the algorithm more flexible and scalable. The design can be applied to engineering practice quickly after refactoring code.

3. People's Fast-moving Detection and Warning in Buses

3.1. The Movement Speed of People

In the process of detection, it is necessary to record the time stamp of each frame when input the video. For five consecutive frames of input images, the objects in each frame are detected and tracked, so as to obtain the actual moving trajectory of each object in the current image sequence. When the moving speed is greater than the threshold that set, the behavior of the target is regarded as moving fast. The movement speed is defined as

$$S = \frac{\sqrt{(P_{n_x} - P_{(n-k)_x})^2 + (P_{n_y} - P_{(n-k)_y})^2}}{t_n - t_{n-k}} \tag{3}$$

Where t_n is the time of the n^{th} frame, t_{n-k} is the time of the k^{th} frame before the n^{th} frame, P_n is the position of the object in the n^{th} frame, and P_{n-k} is the position of the object in the k^{th} frame before the n^{th} frame. The interior of the bus is narrow and deep, so the horizontal moving distance of passengers can be ignored, and only the moving distance in the depth direction is calculated. When calculating the speed, the plane of the camera is taken as the reference, and the position of the object can be regarded as the distance between its plane and the plane of the camera.

3.2. Distance Measurement

At present, the cameras that mostly used in buses are still monocular cameras, with which it is difficult to obtain the spatial information of the scene directly. Therefore, it is necessary to calibrate the camera, so as to establish the relationship between the camera image pixel position and the actual spatial position. According to the past experience, it is found that the position relationship between the pixel and the distance of the collected calibration point is similar to quadratic function. The formula of pixel and distance curve can be regarded as follow.

$$y = ax^2 + bx + c \tag{4}$$

Where x is the image pixel of the calibration point and y is the actual distance between the point and the camera. The parameters a, b and c can be obtained by least squares fitting. According to the actual situation of the bus interior space, the least square method is used to fit the curve, and the parameters are obtained. In order to reduce the accidental error and improve the detection accuracy, changes of distance and time interval of 1 frame, 2 frames,..., k frames are recorded. These can be used to calculate the current moving speed of the object. Then equation (3) can be rewritten as

$$S_a = \frac{1}{k+1} \sum_{i=0}^{k} \frac{y_{n-i} - y_{n-i-1}}{t_{n-i} - t_{n-i-1}} \tag{5}$$

Where y_n is the position of the actual distance between the object and the camera in the n^{th} frame, and y_{n-i} is the distance in the i^{th} frame before the n^{th} frame. The value of k can be determined by debugging.

3.3. Warning Mechanism

In order to ensure that the system can give stable alarm information, the fast-moving behavior warning mechanism is used. When the people's moving speed is greater than the threshold TH_h, the warning mechanism is triggered. When the speed is lower than TH_l, the alarm is released. TH_h is the trigger alarm speed, and TH_l is the release alarm speed. A is the alarm signal, the initial state is A=0. When the alarm is triggered, A=1; when the alarm is released, A=0. The principle of people's fast-moving warning is shown in Figure 2. It can effectively reduce the repeated and false alarm.

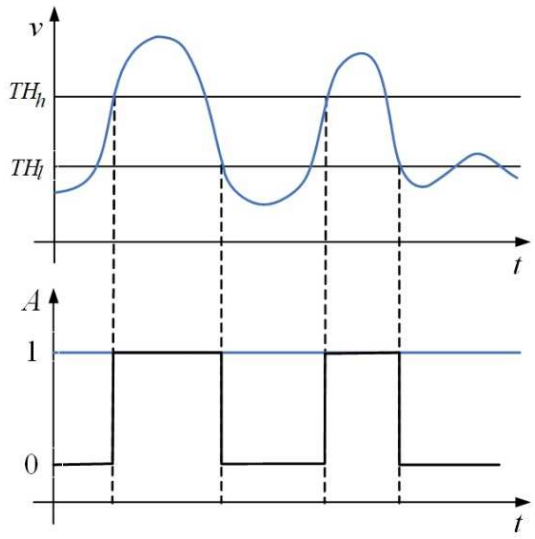
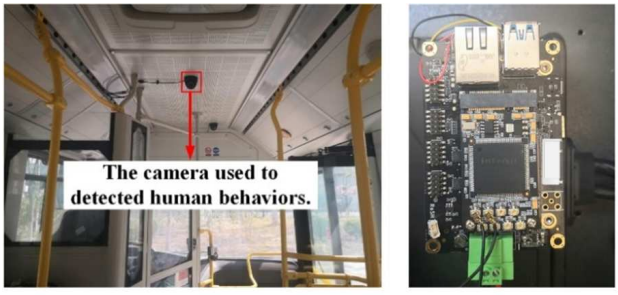


Figure 2. *Warning principle of people's fast moving.*

4. Experimental Results and Analysis

4.1. Experimental Environment

a) To evaluate our approach, two cameras are installed in the front and the rear of the bus for us to collect videos and images of people's activities. The detection equipment in the bus is shown in Figure 3.
b) During training, the computer is equipped with Ubuntu 18.04 operating system, and PyTorch is used to train the model.
c) In testing phase, intelligent computing module equipped with NVIDIA Jetson Xavier NX is used to detect the fast-moving behavior in the bus. The specific configuration is as follows: Jetson version 4.4, cuda 10.2, cudnn 8.0, tensorRT 7.1.0. The C++ Programming Language is used to implement algorithms.



(a) Camera in the bus (b) Intelligent computing module

Figure 3. *The detection equipment in the bus.*

4.2. Experimental Results

This experiment collects real data of people's activities in the bus, including fast moving and normal moving video. After screening and processing, the detection algorithm is used to obtain the position of people in the bus, so as to obtain the speed change curve, as shown in Figure 4. x represents the number of frames of the video, and y represents the calculated average speed of the person. Similarly, the speed curve of people moving normally in the bus is shown in Figure 5. The internal space of the bus is closed and narrow, so it is difficult for people to move at normal walking speed, and the change of people's speed is irregular. But through many experiments and speed calculations, the method can still distinguish between normal movement and fast movement.

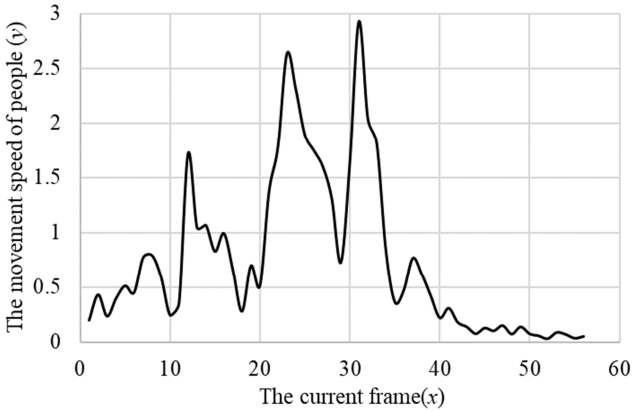


Figure 4. *Speed curve of people's fast moving in the bus.*

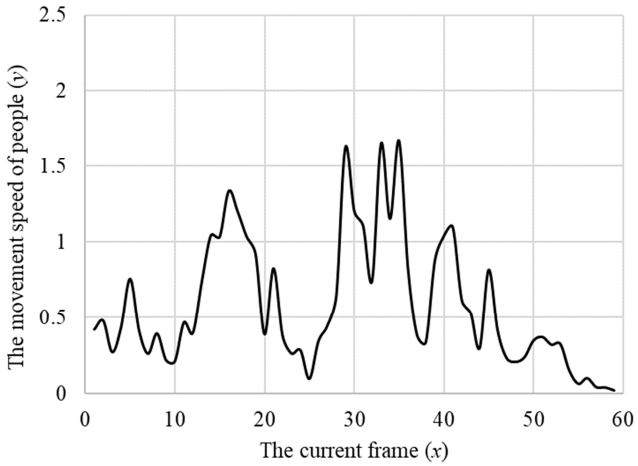


Figure 5. *Speed curve of people's normally moving in the bus.*

Through the above analysis, we can see that the speed curve of people in different states will show obvious differences, and the speed fluctuation range of people in normal movement is small. Therefore, we can set the speed threshold to judge whether the person is in fast moving. Figure 6 shows the test results of the algorithm in the bus.

After many times of debugging, the parameters are determined as $k = 4$, $TH_h = 1.9$, $TH_l = 1.6$. When someone moves quickly, the system will give an alarm. Through visualization, abnormal objects will be distinguished. The results of the algorithm running in the bus are shown in Figure 7.

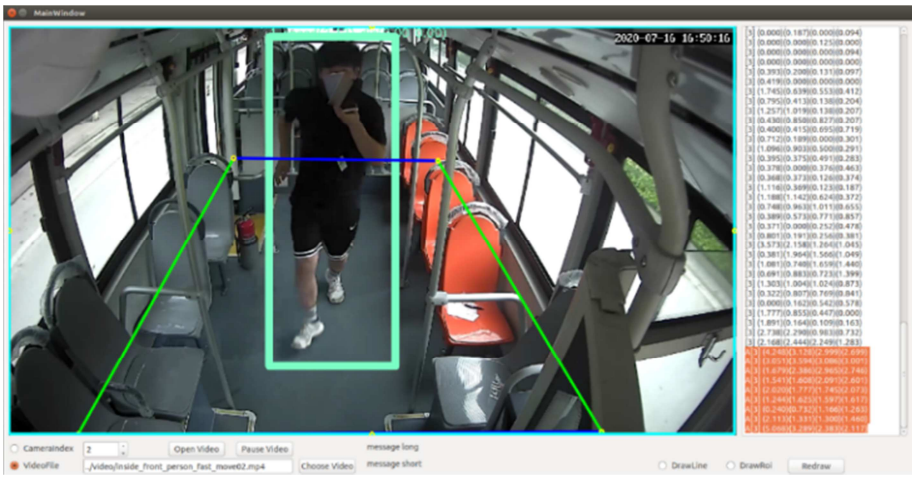


Figure 6. *Test results of fast moving detection algorithm in the bus.*

Figure 7. *Implementation of the fast moving detection algorithm in bus.*

4.3. Test Results

In order to apply YOLOv5 in the bus, the trained model is transformed into onnx model, and later transformed into the engine model which can adapt to tensorRT. After that, initialize the engine, and specify the network's input and output. Finally, the real-time detection of people in the bus is realized by using intelligent computing module.

In our experiment, 63 groups of fast moving videos and 70 groups of normal moving videos are used as test data to verify the accuracy of the algorithm. The test data includes single person moving and multi person moving, which can be divided as less occlusion, partial occlusion and severe occlusion. Among them, there are two directions for people's movement, which are the direction close to the camera (from the rear of the bus to the front of the bus) and the direction away from the camera (from the front of the bus to the rear of the bus). Meanwhile, this experiment measures the actual effect of the algorithm through three indicators: false alarm rate, missed alarm rate and accuracy, so as to verify the rationality and practicability of the detection system. False alarm refers to the situation that the system alarms when the people move normally and there is no fast movement in the video segments. The number of false alarm samples is recorded as FP, and the false alarm rate is defined as Fal. Missed alarm refers to the situation that the system does not give an alarm when people move fast. The number of missed samples is recorded as FN, and the missed alarm rate is defined as Mis. Therefore, the correct identification of the system includes two situations: one is that the system alarms when the people move fast (the number of such samples is recorded as TP), and the other is that no alarm when the people move at normal speed (the number of such samples is recorded as TN), then the system will maintain normal operation. The accuracy is defined as Acc. The calculation method of each indicator is as follows.

The false alarm rate is expressed as

$$Fal = \frac{FP}{TN + FP} \times 100\% \tag{6}$$

The missed alarm rate is expressed as

$$Mis = \frac{FN}{FN + TP} \times 100\% \tag{7}$$

The accuracy is expressed as

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{8}$$

The test results of the approach are shown in the Figure 8. TP=61, TN=66, FP=4, FN=2. Therefore, the false alarm rate is Fal=5.7%, the missed alarm rate is Mis=3.1%, and the accuracy is Acc=95.4%. Through the analysis of the video segments of error recognition, it shows that when the bus is extremely crowded, or the object person is far away from the camera, people will be severely obscured, resulting in the loss of the object, and makes it difficult for the system to detect the moving speed accurately. In the follow-up research, if multiple cameras can be used for joint detection, this problem will be effectively solved, and the detection ability of the system will also be improved. Overall, the accuracy of the algorithm is higher than 95%. In addition, online real-time test is carried out in this experiment. The algorithm can process

about 20 frames per second, and the running speed can meet the real-time requirements.

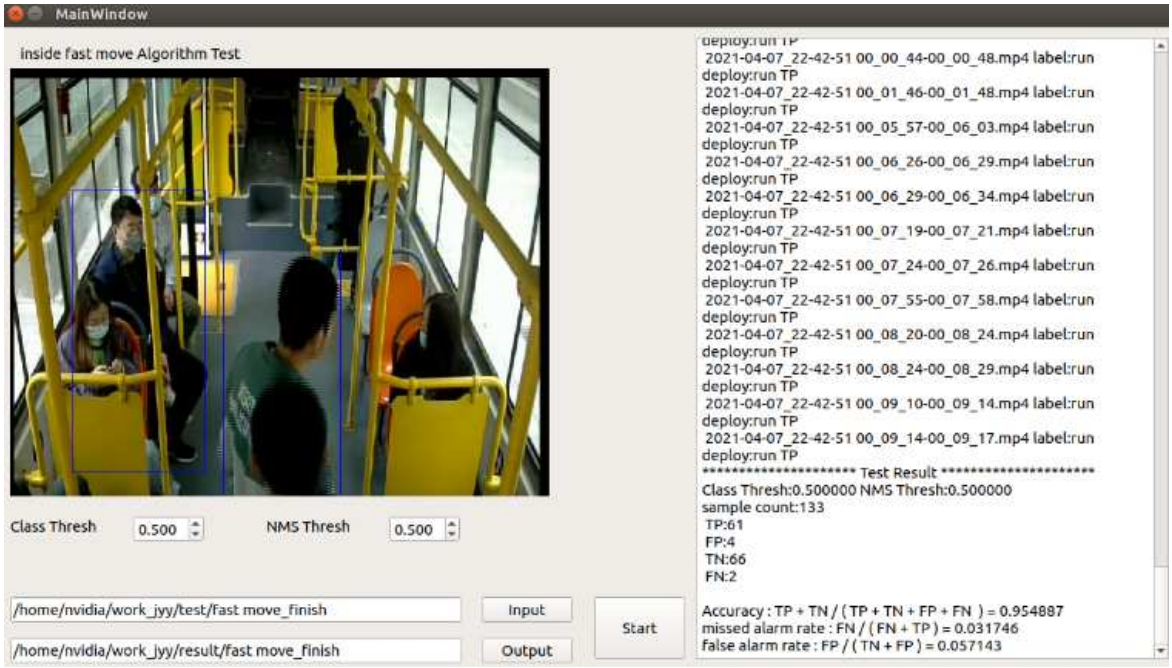


Figure 8. *Indicators results of fast moving detection algorithm in the bus.*

5. Conclusions

In general, the proposed method can achieve good results without serious obstruction and can meet the requirements of real-time detection. However, when the crowd is too dense, the accuracy of the object detection will decline to a certain extent in the actual monitoring. Next, YOLOv5s and two-stage tracking strategy are introduced to the method, which are flexible, practical and extensible, and can adapt to a variety of recognition tasks in complex scenes. Meanwhile, the intelligent computing module used in our method can execute algorithms efficiently. It is small and light, and can adapt to any mobile scene. People's fast moving is one of the abnormal behaviors in the bus. In addition, there are many other behaviors and actions in the bus that require people's attention, such as fighting, abnormal gathering, and fall behavior. How to find more effective methods to identify these abnormal behaviors and improve the efficiency of algorithms will be the next step of our research.

Acknowledgements

This work was supported in part by Beijing Natural Science Foundation (grant no. 4204096, 4212035), National Natural Science Foundation of China (grant no. 61903006), National Key R&D Program of China under Grant (grant no. 2017YFC0821102, 2017YFC0822504), Beijing Municipal Great Wall Scholar Program (grant no. CIT&TCD 20190304), Basic Scientific Research Projects of Beijing Municipal Education Commission, Youth Yuyou Talent Project of North China University of Technology, Research Initial Foundation of North China University of Technology.

References

[1] Hu, Y., Lu, M., & Lu, X. (2019). Feature refinement for image-based driver action recognition via multi-scale attention convolutional neural network. Signal Processing Image Communication, 81.

[2] Hatirnaz, E., Sah, M., & Direkoglu, C. (2020). A novel framework and concept-based semantic search Interface for abnormal crowd behaviour analysis in surveillance videos. Multimedia Tools and Applications, 1-39.

[3] Tripathi, V., Mittal, A., Gangodkar, D., & Kanth, V. (2019). Real time security framework for detecting abnormal events at ATM installations. Journal of Real-time image processing, 16 (2), 535-545.

[4] Dhiman, C., & Vishwakarma, D. K. (2019). A review of state-of-the-art techniques for abnormal human activity recognition. Engineering Applications of Artificial Intelligence, 77, 21-45.

[5] Zhan, C., Duan, X., Xu, S., Song, Z., & Luo, M. (2007). An Improved Moving Object Detection Algorithm Based on Frame Difference and Edge Detection. In Proceedings of the Fourth International Conference on Image and Graphics, 519-523.

[6] Tsai, D. M., & Lai, S. C. (2008). Independent component analysis-based background subtraction for indoor surveillance. IEEE Transactions on image processing, 18 (1), 158-167.

[7] Aslani, S., & Mahdavi-Nasab, H. (2013). Optical flow based moving object detection and tracking for traffic surveillance. International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering, 7 (9), 1252-1256.

[8] Jocher, G. (2020). Yolov5. Code repository https://github. com/ ultralytics/ yolov5.

[9] He, L., Liao, X., Liu, W., Liu, X., Cheng, P., & Mei, T. (2020). FastReID: a Pytorch toolbox for real-world person re-identification. arXiv preprint arXiv: 2006.02631.

[10] Bochinski, E., Eiselein, V., & Sikora, T. (2017). High-Speed tracking-by-detection without using image information. In 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance, 1-6.

[11] Bochinski, E., Senst, T., & Sikora, T. (2018). Extending IOU based multi-object tracking by visual information. In 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance, 1-6.

[12] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, 779-788.

[13] Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv: 1804. 02767.

[14] Wang, C. Y., Liao, H. Y. M., Wu, Y. H., Chen, P. Y., Hsieh, J. W., & Yeh, I. H. (2020). CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 390-391.

[15] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2117-2125.

[16] Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path aggregation network for instance segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, 8759-8768.