Research on real-time object detection based on Yolo algorithm

Yuxin Zhao [a], Shaodong Wang [b]

School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, China

[a] 1186307642@qq.com, [b] 2055698569@qq.com

Abstract. Carrying controlled knives and guns and ammunition in public places is a serious threat to public safety, and the target proportion of images in monitoring is small, and the recognition background is complex. The existing monitoring methods in public places have problems that can not be well recognized automatically. Therefore, the YOLOV5 model is proposed to be applied in Intelligent Security to improve public safety. Focusing on the target detection of contraband in public under the application scenario of Intelligent Security, the characteristics and requirements of the detection task are analyzed, and the viewpoint of applying YOLOV5 algorithm to real-time detection of contraband is put forward. The history and basic principles of YOLO series algorithms and target detection are summarized. YOLOV5 performance test was carried out systematically to verify the feasibility of the method. Then, the detection dataset is made, format conversion and labeling are carried out, and the training data is expanded, and the related detection experiments are carried out on the computer. The results show that the contraband detection method based on YOLOV5 has great potential in practical application.

Keywords: Object detection; YOLO V5; Deep learning; Intelligent Security.

1. Introduction

With the continuous development of society, people's quality of life improves, at the same time, public safety issues are more and more important: a large number of monitoring probes are placed in public places such as subway entrances, railway stations, airports, and so on, forming a complex monitoring network, which provides social and public security protection. However, with the gradual coverage of monitoring scope, monitoring information also rises significantly. Traditional video surveillance requires human resources to monitor it, but due to the increasing scope of surveillance, and considering the actual situation of limited energy of monitors, there may be problems such as inappropriate detection, omission and misdetection. And for large public places with large traffic, it is usually necessary to retrieve the specified target quickly and accurately, but obviously the traditional manual viewing method can not meet its needs. With the development of science and technology, real-time target detection using computer vision, combined with automatic analysis of monitoring information, can greatly improve the detection efficiency and meet the needs of intelligent security at this stage.

Target detection by computer vision is to convert image signals into digital signals based on the information of pixel distribution, brightness and color taken by the monitoring probe, then extract the characteristics of the target by various operations, and output the detection results according to the preset information and other conditions, so as to realize the target detection function of identifying the specified items with high accuracy.

The purpose of analyzing monitoring information through computer vision technology is to identify the visual objects in it. Considering the fact that China does not allow private possession of guns or carrying controlled knives to travel in public, this paper mainly discusses the requirement of computer vision technology for real-time detection of prohibited items in the smart security system. Computer vision for contraband detection focuses on fast identification of contraband, while maintaining a high accuracy. Security assurance in public has been completed.

YOLO is a widely used algorithm, which is known for its characteristics of target detection. In 2015, Redmon et al. developed the first version of YOLO. In the following years, scholars published several subsequent versions of YOLO, called YOLO V2, YOLO V3 and YOLO V5.

YOLOV5 applies a single convolution neural network to the entire image, divides the image into grids, and predicts the class probability and bounding box of each grid. Because YOLO redefines object detection as a regression problem, it does not require complex pipelines for object detection and has good real-time performance.

The YOLO V5 is used to implement a monitoring probe that captures public places, and real-time analysis of the collected pictures can be used to determine whether related personnel are carrying prohibited items such as control knives in places where people are highly mobile. At the same time, because the use of computer automatic detection instead of human resources, it can greatly improve the real-time monitoring, accuracy, and thus improve the detection efficiency.

Target detection technology is the basis of artificial intelligence. In this research paper, we mainly discuss xxx. For this reason, we have carried out relevant simulation experiments. After extensive model and scene training, we propose to use YOLO in combination with target monitoring and apply it to the monitoring probe. Real-time monitoring of prohibited items in public places to achieve high accuracy, in the shortest time to identify the prohibited items, so as to improve public security, improve the building of intelligent security system.

2. Previous work

2.1 Development of Object detection

The earliest target detection system is a series of devices represented by radar. They all process one-dimensional pulse signals.

Previous methods proposed the method to use principal component analysis: subspace learning-PCA. A general framework for target detection is proposed in static images, which learns features directly from samples without any prior knowledge, models, or motion segmentation [1] [2] [3]. There are lots of comprehension surveys summarized the motion image analysis algorithms and divided them into two categories: the analysis based on optical flow and the analysis based on feature points. The Carnegie Mellon University has developed a visual surveillance project, VSAM, which can monitor ordinary civil scenes and battlefields in real time. Subsequently, the BackgroundSubtractorMOG is proposed for a foreground and background segmentation algorithm based on the Gaussian Mixed Model. Seven methods are commented and on the original classification based on speed, memory requirements, and accuracy. This review allows readers to compare the complexity of different methods and can effectively help them choose the best one for a particular application. Thereafter, visual tracking methods are divided several into non-recursive and recursive techniques. Then the issue of visual tracking is raised and divided it into bottom-up and top-down processing ideas [4]. The Temporal-Difference Method proposed using the pixel-level difference between two to three consecutive frames in an image sequence to extract moving regions.

Shafie et al. proposed optical flow method. This method depends on the apparent velocity distribution of the motion of the brightness pattern in the image and provides information about the spatial arrangement of the observed objects [5]. The optical flow method is computationally complex and suitable for complex dynamic image analysis. As a result, image sequence analysis has entered a new climax of research.

Schmaltz et al. proposed a region-based technology that takes into account changes in the image area corresponding to the moving object [6].

Subsequently, Bouwmans et al. made a comprehensive survey of the statistical background modeling methods used for foreground detection, and classified each method according to the statistical model used [7].

2.2 Development of YOLO

One-Stage is a target detection algorithm based on regression thought proposed in recent years. One of the two typical algorithms is YOLO (You Only Look Once) algorithm. Redmon et al. proposed YOLO algorithm, which satisfies the real-time performance of target detection at a slight sacrifice of detection accuracy [8]. As a viral algorithm, the algorithm can detect 45 frames per second, more than twice as fast as other real-time detection algorithms.

Joseph Redmon et al. proposed Yolo2 using darknet-19 as a classification network. It mainly improves the inaccurate positioning and low recall rate in the first generation of Yolo. The main change is to simplify the network and reduce the image of the input network. At the same time, Joseph Redmon et al. proposed YOLO9000 with Darknet-19 as the backbone. Its main innovation is the use of a variety of computer vision technologies, such as location prediction.

Subsequently, a new version of Yolo, Yolov3, is proposed with Darknet-53 network as the main network. The most striking feature of Yolov3 is that it can detect three different sizes of targets. Yolov3 allows you to create custom bounding boxes for each project and introduces network adaptive features. Similar to FPN, multiscale prediction is performed, and a single neuron network is used to collect and preprocess pictures [9]. Compared to previous versions, Yolov3 replaces softmax loss in Yolov2 with logistic loss, which improves the advantage of small object recognition [10]. At the same time, the processing speed of the algorithm is reduced by increasing the number of predictions [11].

Subsequently, AlexeyAB et al. proposed YOLOV4. The new generation of YoloV4 algorithm still uses CSPDarknet-53 as the main network, and can be trained and tested with the traditional GPU, while obtaining real-time and high-precision detection results [12].

Yolov5 is essentially a structural modification of Yolov3. The current Yolov5 algorithm has good accuracy and real-time detection characteristics.

3. Introduction to Yolo

Yolov5 was proposed in four models: Yolov5s, Yolov5m, Yolov5l, Yolov5x. Among them, Yolov5s has the smallest network and the lowest speed and AP accuracy compared with other versions. Several other versions have improved network and AP accuracy, but recognition speed has decreased.

The structure of Yolov5 is divided into four parts: input, Backbone, Neck, and Prediction. Compared to Yolov4, Yolov5 has improved in each section.The details are as follows：

3.1 Input

3.1.1 Mosaic Data Enhancement

Like YoloV4, the mosaic data enhancement used in Yolov5 is a reference to the CutMix data enhancement proposed at the end of 2019.

The reason for using Mosaic data enhancement is that during normal project training, the AP of small target is generally much lower than that of medium and large target, and the small target (0x0-32x32) accounts for a larger proportion of the coco dataset than medium and large target.

Furthermore, small targets only account for 52.3% of the image location distribution, which is less uniform than medium and large targets.

Compared to CutMix, Mosaic Data Enhancement increases the number of images stitched from two to four, and stitches them with random scaling, random clipping, and random arrangement, which has enriched the detection dataset and making the network more robust. Because the data for four pictures can be calculated directly, the Mini-batch size is not very large and can reduce GPU requirements.

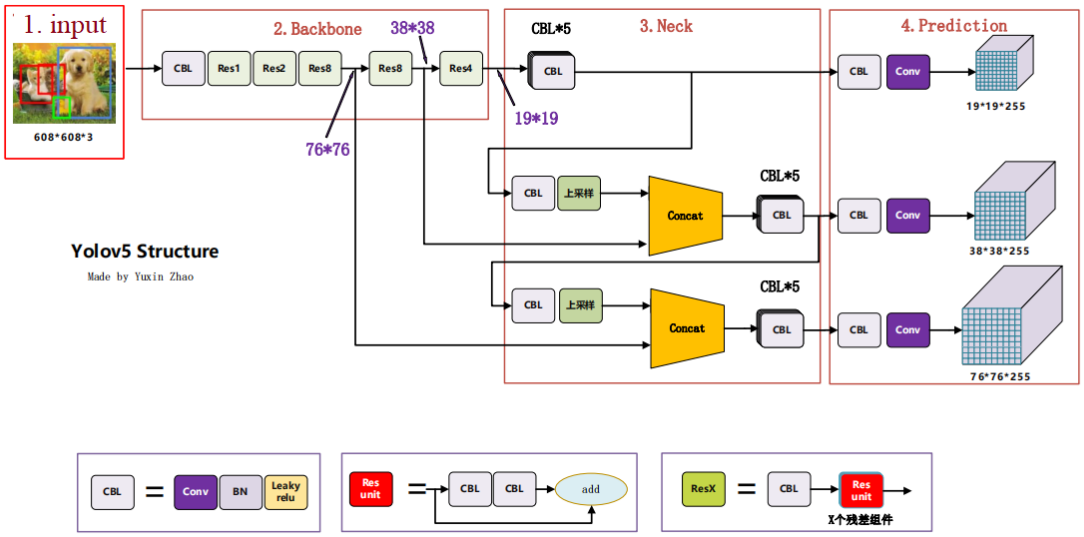

Fig. 1 Mosaic Data Enhancement

3.1.2 Adaptive anchor frame calculation

In the Yolo algorithm, there are anchor frames with initial lengths and widths for different datasets. In the normal network training, Yolov3 and Yolov4 use the following method:first calculating the initial anchor frame separately, then outputting the prediction box on the basis of the initial anchor frame, then comparing with the real box groundtrue, calculating the deviation, and then updating the parameters of the iterated network in reverse.

However, Yolov5 embeds this feature in the code so that each training session, the algorithm adaptively calculates the value of the best anchor frame in different training sets. Of course, Yolov5 also supports turning this off.

3.1.3 Adaptive picture scaling

In Yolo algorithm, different detection pictures may have different lengths and widths. Therefore, Yolo scales the detected pictures to a uniform format for detection (416*416 or 608*608).

In Yolov5, when zooming in and out of a picture, the original image is first adaptively added with a minimum of gray edges to reduce information redundancy, in order to increase the reasoning speed.

3.2 Backbone

3.2.1 Focus structure

The Focus structure which is not found in Yolov4 was added in Yolov5. In Yolov5s, 32 convolution cores are used in this structure. The most critical one is the slicing operation: after the original 608*608*3 image is input into the Focus structure, the image is sliced into a 304*304*12 feature map, then a convolution operation of 32 convolution cores turns to a 304*304*32 feature map.

3.2.2 CSP structure

The Yolov4 version only designs the CSPDarknet53 structure in the backbone network, which is designed to enhance the learning ability of CNN while reducing computing bottlenecks and memory costs, so that the model remains lightweight with good accuracy. The convolution core in front of each CSP module has a size of 3*3 and a stride of two, so it can act as a downsampling. In Yolov5, two CSP structures were designed. Take Yolov5s for example, CSP1_X structure applies to the backbone network, while CSP2_ The X structure is used for Neck.

3.3 Neck

3.3.1 FPN+PAN structure

Both Yolov5 and Yolov4 adopt the FAN+PAN structure. FPN transmits and fuses high-level feature information by up-sampling from top to bottom to get a predicted feature map. FPN conveys strong semantic features downwards, while the two PAN structures behind the FPN layer convey strong positioning features from the bottom to the top, which aggregates parameters from different backbone layers for different detection layers to further improve the ability of feature extraction.

3.4 Prediction

3.4.1 CIOU_ Loss

The loss function for target detection tasks generally consists of two functions, Classificition Loss and Bounding Box Regeression Loss.

Yolov5 continues the CIOU_Loss regression method that has been used in Yolov4, which takes into account the scale information of the aspect ratio of the bounding box on the basis of DIOU, makes the prediction box regression faster and more accurate.

3.4.2 NMS non-maximum suppression

In post-target detection processing, NMS operations are usually required for filtering many target frames. Compared with the DIOU_NMS approach used by Yolov4 on the basis of DIOU_Loss, the weighted NMS approach is used in Yolov5. This change improves the recognition of some overlapping targets.

4. Experimental process and results

In this part, we will show the experimental process and the results of this experiment.

4.1 Experimental environment and data

Table 1. Experimental environment

Environment	Edition	Function
Pytorch	1.10	In-depth learning framework
CUDA	11.3	Parallel computing architecture
Python	3.8	Writing Code Language
Coco	2017	Training dataset
cuDNN	8	In-depth learning framework
YOLOV5	V6.1	Test Target
Jupyter lab	3.0.16	Online experimental platform

Table 2. Hardware environment

Hardware environment	Model/ capacity
Operating system	Windows 10
CPU	AMD Ryzen 9 5900HX
Memory	32G
Hard Disk	1T
Graphics card	NVIDIA Tesla K80

4.2 Test data

This training test collected 4000 pictures containing aircraft, crowd and tool information from the Internet through crawler technology, including data sets in different environments and angles. At the same time, the data set is expanded by scaling the data set. The ratio of training set to test set is 9:1, that is, 3600 training sets and 400 test sets.

4.3 Evaluation criteria

The model uses P, R, F1 as the evaluation parameters of network training results. P represents the correct proportion of all targets predicted by the model; R represents the correct target proportion predicted by the model in all real targets. The calculation formulas of P and R are as follows:

$$P=[A/(A+B)] * 100\% \quad (1)$$

$$R=[A/(A+C)] *100\% \quad (2)$$

Note: A is the number of correct prediction targets in the prediction results; B is the number of wrong prediction targets in the prediction results; C is the number of targets that have not been predicted. The higher the accuracy rate P, the higher the proportion of correct result samples in the prediction results and the lower the false detection. The higher the recall rate r value, the more positive samples are correctly detected in the prediction results, and the lower the missed detection.

4.4 Specific process

4.4.1 Flow-process diagram

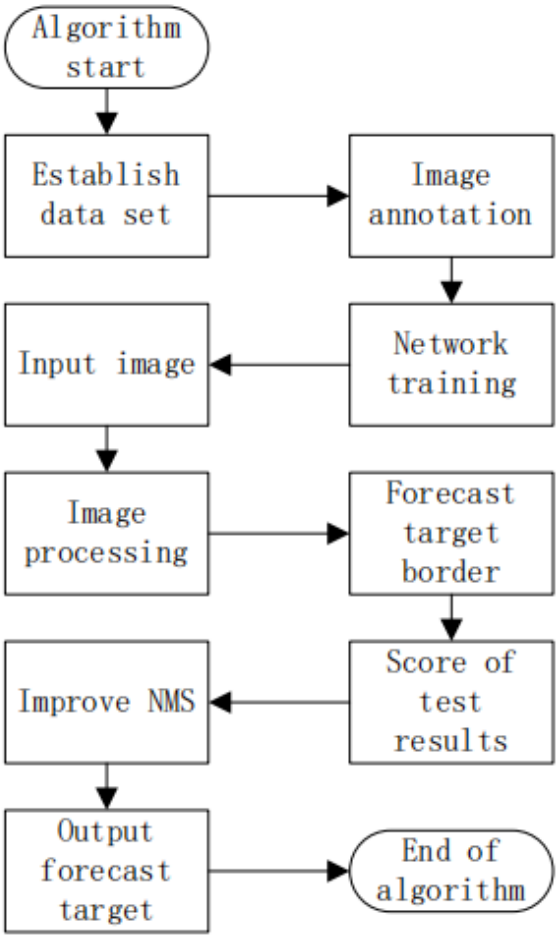

Fig. 2 Flow-process diagram

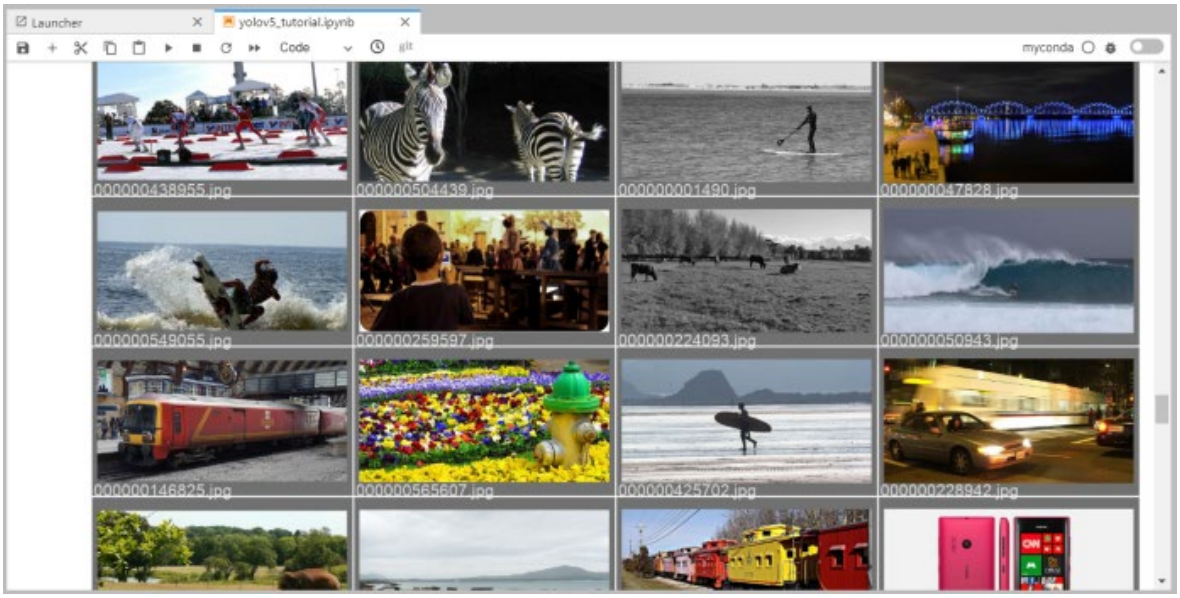

Fig. 3 legend

4.4.2 Training data set process

1) Create a configuration file under the data folder

Create 1make_txt.py files under the root directory

Continue to create 1voc_label.py files under the root directory

2) Execute the above two .py files in sequence

The execution results show as follows. Firstly,we generate txt file under labels folder(display the specific annotation data of the dataset). Then the four txt files are generated under imagesets folder. Last but not least, we generate three txt files under data folder (path with image)

3) Modify .yaml file. Configure a yaml file belonging to your data set and modify it in three parts. Firstly, we modify the path of train, val and test files to the path just generated. Then, the number in the NC file represents the category of the data set. Lastly, the class name marked for your dataset in the names file

4) Modify models file. There are four models under models. The training time of smlx increases in turn, and you can select a file to modify according to your needs.

5) Training train.py file

6) Test detect.py file

4.5 Experimental result

Table 3. Experimental results

Algorithm	Average detection time of single graph/ms	Accuracy rate/% P	Recall% R
YOLOv4	41.62	73.1	71.9
YOLOv5	38.97	84.7	83.5

The results from the table above show that the YOLOv5 model has stronger multi-scale target detection ability than the YOLOv4 model, can detect more targets, and improves the detection accuracy of overlapping and occluding targets.

4.6 Discussion

For crowded places, test data show that YOLOV5 can still be more accurately identified. At the same time, it can distinguish the relationship between the front and back of the population, and will not cause problems such as misidentification, wrong number, etc. due to population occlusion.

4.7 Recognition effect display

The practical results of our system are demonstrated below in Fig. 4.

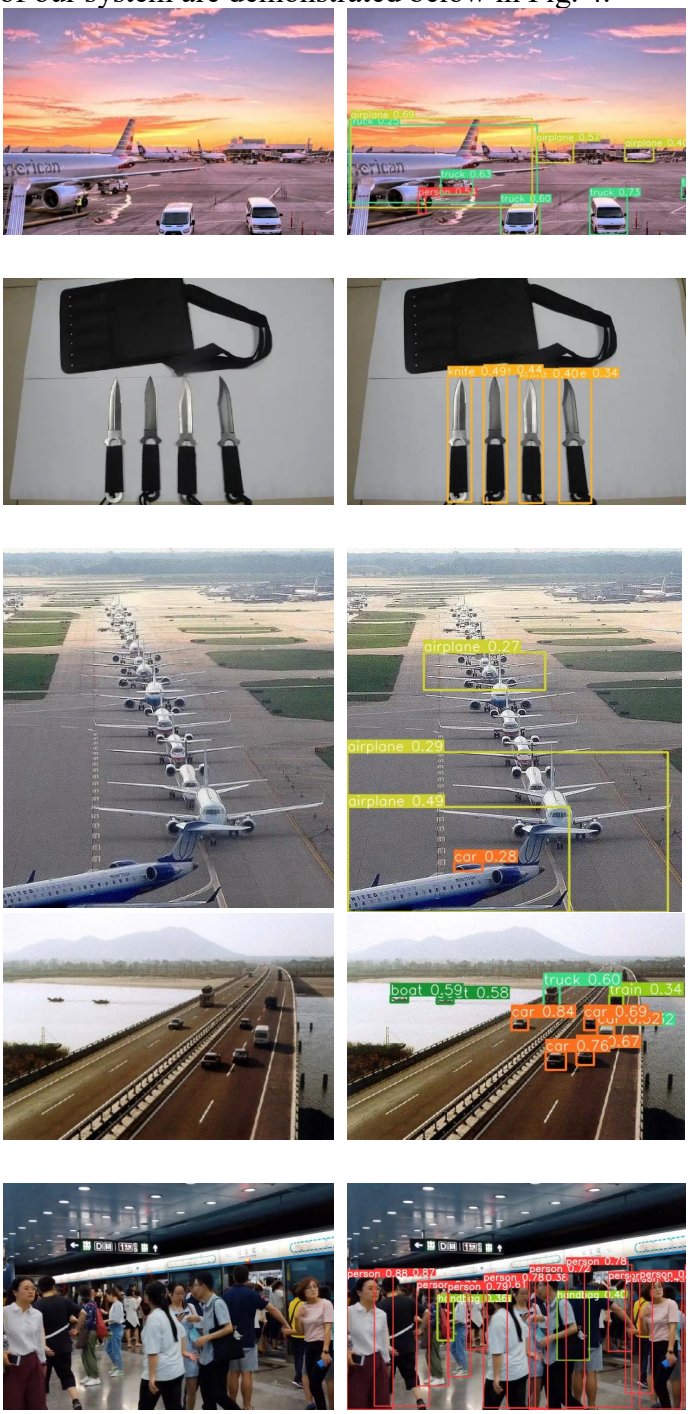

Fig. 4 Result in Real scenarios

5. Conclusions

In this paper, the YOLOV5 algorithm is applied to the real-time detection of contraband in public. Through validation on target datasets of cutters and crowds, this method has good detection performance for target cutters in complex crowded environment in public places, and provides strong technical support for existing regulatory means of Intelligent Security. The experimental results show that the detection accuracy of YOLOV5 is 11.6% higher than that of YOLOV4, and the recognition speed reaches the requirement of real-time.

References

[1] Papageorgiou C P, Oren M, Poggio T. A general framework for object detection[C]//Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271). IEEE, 1998: 555-562.

[2] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks [C] //European conference on computer vision. Springer, Cham, 2014: 818-833.

[3] Zhao Z Q, Bian H, Hu D, et al. Pedestrian detection based on fast R-CNN and batch normalization[C] //International Conference on Intelligent Computing. Springer, Cham, 2017: 735-746.

[4] Grauman K, Leibe B. Visual object recognition (synthesis lectures on artificial intelligence and machine learning)[J]. Morgan & Claypool, 2011, 3.

[5] Yaseen Z M, El-Shafie A, Jaafar O, et al. Artificial intelligence based models for stream-flow forecasting: 2000–2015[J]. Journal of Hydrology, 2015, 530: 829-844.

[6] Schmaltz C, Rosenhahn B, Brox T, et al. Region-based pose tracking[C] //Iberian Conference on Pattern Recognition and Image Analysis. Springer, Berlin, Heidelberg, 2007: 56-63.

[7] Bouwmans T, El Baf F, Vachon B. Background modeling using mixture of gaussians for foreground detection-a survey[J]. Recent patents on computer science, 2008, 1(3): 219-237.

[8] Papageorgiou C, Poggio T. A trainable system for object detection[J]. International journal of computer vision, 2000, 38(1): 15-33.

[9] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1904-1916.

[10] Chen D, Hua G, Wen F, et al. Supervised transformer network for efficient face detection[C]//European Conference on Computer Vision. Springer, Cham, 2016: 122-138.

[11] Najibi M, Rastegari M, Davis L S. G-cnn: an iterative grid based object detector[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2369-2377.

[12] Wu D, Lv S, Jiang M, et al. Using channel pruning-based YOLO v4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments[J]. Computers and Electronics in Agriculture, 2020, 178: 105742.