Improved YOLOv3 Infrared Image Pedestrian Detection Algorithm

Jianting Shi[1(✉)], Guiqiang Zhang[1], Jie Yuan[2], and Yingtao Zhang[3]

[1] School of Computer and Information Engineering, Heilongjiang University of Science and Technology, Harbin 150022, China
229468764@qq.com
[2] Shanghai Aerospace Electronics Technology Research Institute, Shanghai 201109, China
[3] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China

Abstract. Security surveillance is widely used in daily life. For nighttime or complicated monitoring environments, this article proposes an infrared pedestrian monitoring based on YOLOv3. In the original YOLOv3 network structure, two aspects of optimization were made. One was to optimize the scale in the residual structure, and the rich features of the deconvolution layer were added to the original residual structure. The other was to use the DenseNet network to enhance the features. The optimization of fusion ability and delivery ability effectively improves the detection ability for small targets, and the pedestrian detection performance based on infrared images. After comparative testing, compared with YOLOv3, the overall mean average precision is improved by 4.39% to 78.86%.

Keywords: Infrared · YOLOv3 · DenseNet

1 Introduction

With the development of computer vision technology, the field of pedestrian detection is also becoming more and more popular. Pedestrian detection is to study and judge the given image or whether there is a pedestrian to be detected in each video sequence, and can accurately and quickly find the specific location of the target. In today's era, science and technology are constantly improving, and hardware technology is gradually improving. In response to the call of the times, deep learning is stepping into the global development track step by step. Therefore, in computer vision, the research of pedestrian detection is gradually culminating. In recent years, road safety issues have attracted more and more attention, and people are looking for ways to reduce the occurrence of traffic accidents, and pedestrian detection technology can effectively reduce them. In terms of future driverless technology, pedestrian detection technology is even more important. Therefore, the aspect of pedestrian detection technology is receiving more and more attention from the society.

Traditional visible light equipment cannot meet the requirements of nighttime or unmanned driving. Compared with the traditional situation, infrared thermal imaging is

J. Zeng et al. (Eds.): ICPCSEE 2020, CCIS 1257, pp. 506–517, 2020.
https://doi.org/10.1007/978-981-15-7981-3_37

based on the information of the relative temperature of the object, and is less affected by various additional factors. In intelligent monitoring, vehicle assisted driving, human body Behavior analysis and other fields have broad application prospects [1, 2]. However, the infrared image has no color, and its accuracy is low when detecting pedestrians. Pedestrian detection algorithms can be divided into traditional algorithms and deep learning-based algorithms. Traditional pedestrian detection algorithms include Haar wavelet features [3], HOG-SVM [4], DPM [5], etc. Traditional pedestrian detection mainly uses artificially designed methods to extract image features, combined with machine learning related algorithms, to identify and classify image features, but traditional algorithms are complex to design, sometimes it is difficult to design reasonable methods in complex scenes, weight parameters are difficult to get more accurate values, and generalization ability is not strong.

Convolutional Neural Network (CNN) [6] has made a significant breakthrough in pedestrian detection. CNN can automatically learn the original representation of the target through a large amount of data. Compared with the features designed by hand, it has more advantages strong discrimination and generalization. Then, based on the deep learning algorithm, after the RCNN algorithm was proposed, a new boom was ushered in. The performance of deep learning methods in multiple image processing fields surpassed the traditional algorithms [7, 8], that is, a series of improved algorithms appeared, including Fast RCNN [9], Faster RCNN [10–12], SSD [13], YOLO [12] and other algorithms. There are two-stage and one-stage algorithms for deep learning. Compared with traditional methods, deep learning methods have both improved the efficiency and speed of detection. Among them, before the advent of YOLO, deep learning was not fast in detection speed and could not guarantee real-time performance, especially in the future in driverless technology. Redmon et al. [14] proposed the YOLO (You Only Look Once, Unified, Real-Time Object Detection) algorithm, and thus entered the field of one-stage target detection. In recent years, with the continuous development of deep learning, methods applied to target recognition and model prediction have been continuously introduced [15–17].

The one-stage concept solves the problem of speed in object detection, while ensuring a certain accuracy, greatly improving real-time performance. Although the speed is improved, compared with other algorithms, the accuracy is not very high. Then came YOLOv2, YOLO9000, YOLOv3. Among them, YOLOv3 has a simple and efficient network structure, which makes it easy to deploy and has a wide range of application scenarios. It is one of the preferred algorithms in many commercial fields. Combined with our actual application scenarios, it is applied to large outdoor surveillance to detect areas where pedestrians are prohibited. And for small object detection and in the case of pedestrian detection in infrared images, YOLOv3 has great application prospects. It not only uses better backbone networks, such as classifiers from backbone networks such as DarkNet or ResNet, but also can detect quickly. The main thing is that the setup environment is simple, the background detection error rate is low, and the versatility is strong. Although the YOLOv3 network uses multi-scale prediction and combines with better classifiers, it has great advantages. However, YOLOv3 still has the following disadvantages: compared with other RCNN series object detection algorithms, the accuracy of identifying objects is poor, and the recall rate is low.

In view of the above problems, this paper improves the YOLOv3 network framework and borrows the ideas of the DenseNet network. Through the improvement, compared to the original YOLOv3 detection effect on pedestrians in infrared images, the (accuracy) MAP is increased by 4.39%. Compared with the network before the improvement, the improved network also has an increase of 2.36% over the network intersection ratio (IOU) before improvement.

2 YOLOV3 Network Algorithm Structure

YOLOv3 is the beginning of one-stage detection. It is a single neural network-based object detection system proposed by Joseph Redmon and Ali Farhadi and others in 2015. In 2017, CVPR Joseph Redmon and Ali Farhadi published YOLOv2, which further improved the accuracy and speed. After further improvement appeared YOLOv3 algorithm. YOLOv3 network algorithm structure Fig. 1 is as follows: YOLOv3 is mainly divided into three aspects, namely: network input, structure, output where defined.

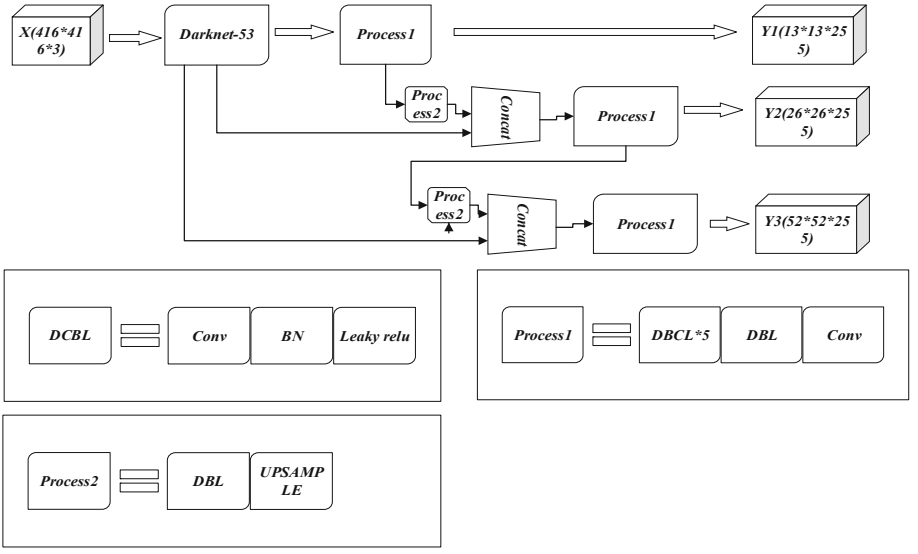

Fig. 1. YOLOv3 network algorithm structure

2.1 Network Input

The general input size of YOLOv3 network is 320 * 320, 416 * 416, 608 * 608, this article introduces 416 * 416, the size must be an integer multiple of 32, so as to facilitate the subsequent training test and analysis. The YOLOv3 network mainly uses 5 downsampling. The YOLOv3 network is based on the DarkNet-53 backbone network, and the step size of each sampling is 2, so the largest size of the backbone network is.

2.2 Network Structure

The general input size of YOLOv3 network is 320 * 320, 416 * 416, 608 * 608, this article introduces 416 * 416, the size must be an integer multiple of 32, so as to facilitate the subsequent training test and analysis. The YOLOv3 network mainly uses 5 downsampling. The YOLOv3 network is based on the DarkNet-53 backbone network, and the step size of each sampling is 2, so the largest size of the backbone network is.

The YOLOv3 network is a fully convolutional network that uses the first 52 layers of DarkNet-53, but does not use a fully connected layer and a pooling layer, and uses many residual structures for layer hopping connections. The use of the residual structure helps the network structure to stay in a deep situation and maintain convergence, so that the training continues, and the deeper the network, the better the training result, the better the results obtained by classification and detection, and the 1 * 1 convolution reduces the amount of calculation to a certain extent. YOLOv3 uses three types of downsampling, which are 32× downsampling, 16× downsampling, and 8× downsampling. In order to ensure that the deeper the characteristics of the network, the better the effect is displayed. Sampling, this can be used for target detection of deep features. YOLOv3 performs shallow features generated by downsampling. YOLOv3 wants to make use of shallow features, so it has a route layer. Tensor stitching (concat) is then performed, and the upsampling of the DarkNet middle layer and a later layer is stitched.

2.3 Output

Taking the input 416 * 416 * 3 as a reference, the output in Fig. 1 above has three scales, which are 52 * 52, 26 * 26, and 13 * 13. This gives the grid a center position. Can detect targets of different sizes. The network output needs to predict the anchor box, and predict three bounding boxes for each cell of the feature map. For each bounding box, it will predict the bounding box. Three aspects, (1) the position of each box. (2) Detect objectness prediction (3) M categories. The bounding box coordinate prediction formula is as follows:

$$b_x = \sigma(t_x) + c_x \tag{1}$$

$$b_y = \sigma(t_y) + c_y \tag{2}$$

$$b_w = p_w e^{t_w} \tag{3}$$

$$b_h = p_h e^{t_h} \tag{4}$$

among them t_x, t_y, t_w, t_h Predicted output representing the bounding box, c_x, c_y Represents the coordinates of the grid cell, p_w, p_h Represents the size of the bounding box before prediction, that is, the width and height of the anchor box. b_x, b_y, b_w, b_h Represents the center coordinates and size of the obtained bounding box.

The confidence formula of the YOLOv3 algorithm is as follows:

$$C_i^j = P_r(Object) * IOU_{pred}^{truth} \tag{5}$$

among them $P_r(Object)$ The probability that the bounding box has an object, IOU_{pred}^{truth} When representing the boundary and the object, the value of the intersection of its predicted boundary and the true boundary of the detected object, C_i^j Confidence of the j-th bounding box representing i grids cell.

Suppose a picture is divided into S * S grids and B anchor boxes, that is, each grid has B bounding boxes, each bounding box has four position parameters, 1 confidence level, and one is set. The parameters are: when the confidence degree indicates that there is an object in the current bounding box, the probability of the class is, and there are classes of class probability, then the output dimension formula of the final output layer of the model is:

$$S * S * [B * (4 + 1 + classes)] \tag{6}$$

Where S represents the length and width of a grid, B represents the number of anchor boxes, and classes represents the class probability.

3 Improved YOLOV3 Infrared Pedestrian Detection Algorithm

3.1 Problems

Under the conditions of good lighting conditions and high imaging quality, the YOLO algorithm can detect pedestrians with an accuracy of more than 96.5%. However, when the lighting conditions at night are insufficient and pedestrians and the background are mixed, the detection accuracy of the YOLO algorithm is only 68.4%. There are more missed inspections. At present, a common solution is to use an infrared camera for shooting. Depending on the principle of thermal imaging, the infrared camera can effectively separate pedestrians from the background, so that pedestrians can be more intuitively identified in monitoring. Although the use of infrared camera can improve the detection efficiency, it still has the following four problems: The first point is that the brightness of the object in the infrared image is related to the surface temperature of the object. When there is more clothing wrapped in winter, the imaging result is poor. The second point is that the infrared image has no color information and the target does not have detailed features such as texture. At the same time, the infrared image has low resolution and many noises, which has a certain impact on the convolution operation. In addition, the subject of this experiment is infrared small target pedestrian detection. Compared with scene pedestrian detection, the third point is that the picture taken by the infrared camera only retains the boundary information of the heating source, but if the pedestrian is wearing thick clothes. Or when there is interference from other heat sources, the boundary between pedestrians and the surrounding environment will not be too obvious. At the same time, due to the characteristics of infrared imaging,

additional noise will be introduced, and the effect of noise on the convolutional neural network increases with the number of network layers. The fourth point and the main problem is that the pedestrian target to be detected has a small pixel area in the image and is not easy to detect.

3.2 Improvement Plan

In order to solve the above problems, this paper optimizes and improves the network based on YOLOv3 algorithm. The improvement work mainly includes the following two measures.

The first one point is that this article has replaced the residual module of YOLOv3. At the same time, this article also optimized the structure in the residual block. The target of this test belongs to the category of small targets. We add a deconvolution layer in the residual module to expand the input feature map to twice the original size and fuse it with the output of the previous scale module. The optimized residual membrane block structure is shown in Fig. 2 as follows:

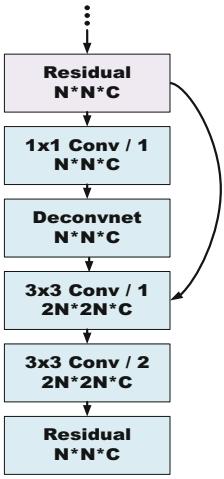

Fig. 2. Structure of the optimized residual module

Second point: The residual structure used in the YOLOv3 feature extraction network is designed with reference to the ResNet network. The network structure is simple and has good feature extraction effects, but for images with less feature information such as infrared images, its feature extraction capability obviously insufficient. Analyzing the data set samples used in this experiment, it can be found that in the image pedestrians currently only have appearance contours, and because of the characteristics of thermal imaging, features such as texture structure cannot be extracted. Convolutional neural networks can only be discriminated by the contours of pedestrians. The simple upper-lower network connection relationship of the ResNet network cannot be applied to this scenario. Therefore, we have borrowed the ideas of DenseNet and

strengthened the connection between the shallow network and the deep network. For neural networks, the types of feature information extracted from the deep layer and the shallow layer are different. This improvement is that the features that can be extracted in the shallow network are mostly simple features, such as the edge structure and texture color of the object. With the deepening of the network, the stronger the learning ability of the network, the features contained in the deep network have richer semantic information. But for infrared images, simply deepening the number of network layers does not have a good effect, so this article borrowed the DenseNet network and input the shallow feature information into the deep network in turn. In the YOLOv3 network, there will be five downsamplings, and the pixel area occupied by pedestrians in the image is originally small. After five times of zooming, the feature discrimination ability of the target in the feature map will be greatly reduced. The feature map is expanded by deconvolution. This measure can improve the saliency of the target feature in the feature map, so that the extraction operation performed on the feature map residual block of the original 13 * 13 size is also 26 * 26-size feature map is performed, which improves the ability to extract small target features.

The optimized convolutional neural network structure of YOLOv3 is shown in Fig. 3.

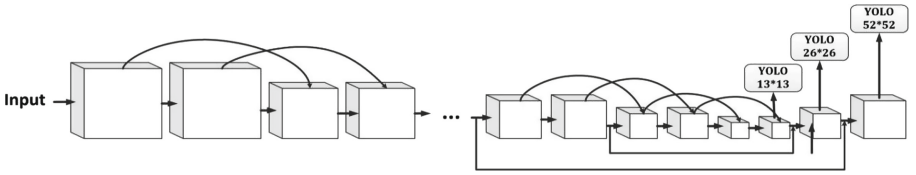

Fig. 3. Convolutional neural network structure of YOLOv3

Based on the above network optimization strategy, we named the optimized network YOLO-I (intimate).

3.3 Experimental Results and Analysis

The data set of the experiments in this paper is from the public data set. First, the data set is cleaned, and 1500 ordinary samples, 400 difficult samples, and 200 negative samples are selected to form a training set of 2100. And a 300 test set with 200 ordinary samples and 100 difficult samples. YOLOv3 introduced the anchor mechanism. The default anchor size is not suitable for this dataset. The K-means clustering method is utilize to recalculate the nine anchors suitable for small target pedestrian datasets.

The loss convergence curve during the network training process is as shown in Fig. 4. The abscissa represents the number of iterations, around 79000. When the network iterates around 60,000 times, it tends to be stable, the parameters change basically stable, and finally the loss value drops to 0.29. Judging from the convergence of this parameter, the network training results are ideal.

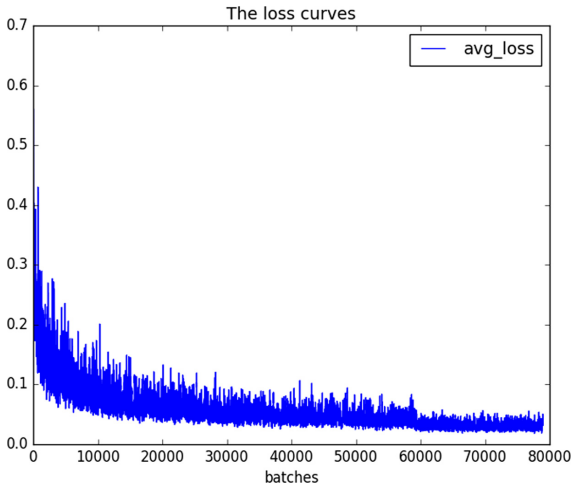

Fig. 4. Loss curve of YOLO-I

As shown in Fig. 5, the test set is tested to obtain the Precision-Recall curve corresponding to the improved model:

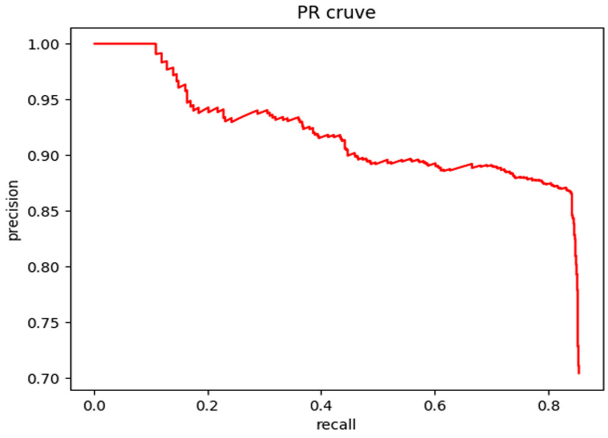

Fig. 5. Precision-recall curve of YOLO-I

It can be seen from Fig. 5 that after the recall rate is 0.8, precision shows a rapid downward trend.

After training the YOLOv3 network before improvement and YOLO-I after improvement, this article then conducted a horizontal comparison test of YOLO-I and YOLOv3. The test environment is the same. In order to reflect the robustness of the network after optimization in more detail, To improve, this article have added test indicators for a more comprehensive comparison. The test results are shown in Table 1.

Table 1. Horizontal comparison table

Contrast situation	Comparison of YOLO-I and YOLOv3				
	Precision	Recall	IOU (%)	F1-score	mAP (%)
YOLOv3	0.86	0.68	60.75	0.76	74.47
YOLO-I	0.89	0.73	63.11	0.80	78.86

The full name of IOU is Intersection-over-Union. A concept used in object detection is the overlap ratio of the generated candidate bound and the ground truth bound, that is, the ratio of their intersection to union. The full name of mAP is Mean Average Precision, which is the average precision value. Is to average the accuracy value of multiple validation set individuals.

Among them: the formula for calculating F1score in Table 1: 2 * Precision * Recall/(Precision + Recall).

From the comparison of the F1-score column in Table 1 above, we can see that the overall robustness of YOLO-I is better than that of YOLOv3. The IOU crossover ratio also reflects the multi-level nesting and fusion of the positional characteristics of the shallow network into the deep Networks are of great help in improving the accuracy of predicting bounding boxes.

Finally, we compared the YOLO-I with the SSD-ResNet and Faster RCNN networks. The comparison results are shown in Table 2.

Table 2. Vertical comparison table

Contrast situation	Comparison of YOLO-I with SSD-ResNet and Faster RCNN			
	Precision	Recall	IOU	F-score
Faster RCNN	0.87	0.72	63.21%	0.79
SSD-ResNet	0.81	0.70	61.66%	0.75
YOLO-I	0.89	0.73	63.11%	0.80

A Subsection Sample It can be known from Table 2 that the overall data of YOLO-I is better than that of SSD-ResNet, and the SSD-ResNet algorithm adopted by SSD. For Faster RCNN network, due to its two-stage network structure, the candidate frame area is extracted beforehand for target classification detection, which reduces the influence of invalid background images. The accuracy of the network and the intersection ratio are better than SSD-ResNet Network, slightly lower than YOLO-I. The actual test comparison chart for the three networks are as shown in Fig. 6, 7 and 8.

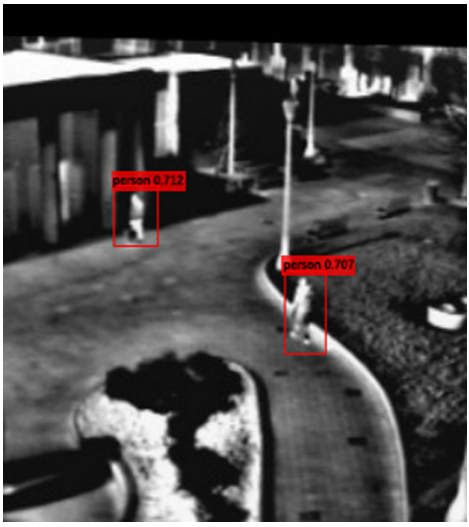

Fig. 6. Faster RCNN

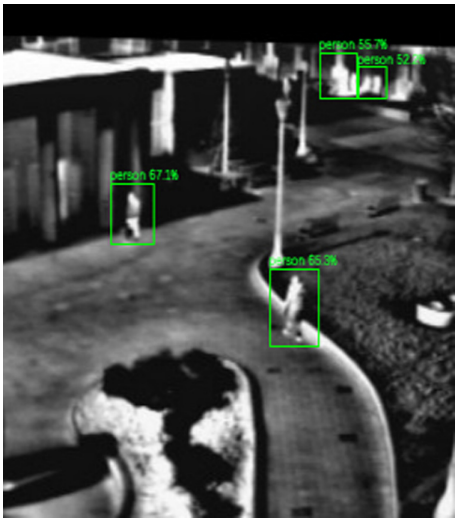

Fig. 7. SSD-VGG16

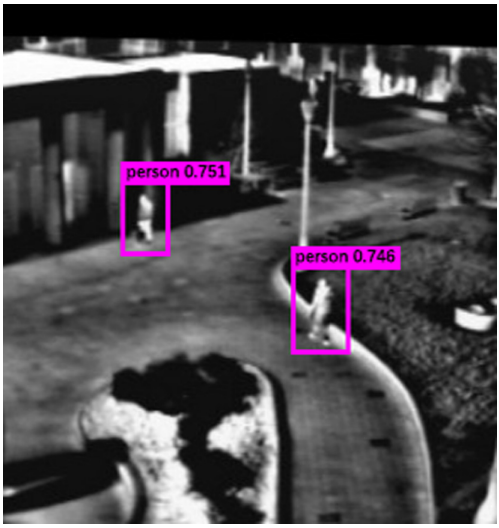

Fig. 8. YOLO-I

4 Conclusion

This paper proposes an improved infrared image pedestrian detection algorithm YOLO-I based on YOLOv3. The optimized YOLO-I has a significantly improved detection capability for grayscale images and small targets, which improves the practicality of infrared detection. This paper mainly aims at the detection environment of low pixels and small targets. Based on the actual detection situation, it is optimized on the basis of YOLOv3. The first is to increase the richness of the feature map size in the residual module. The deconvolution network and sliding step size are the convolution kernel of 2 performs upsampling and downsampling operations, and has two scale feature maps in the same residual module. Compared with the previous, the richness of feature information and the positioning ability of small targets were improved. It strengthened the utilization of shallow features and network-wide features, borrowed dense connections from DenseNet, enhanced the ability to transfer feature information, and effectively improved detection accuracy. After testing and optimization, YOLO-I is targeted at small infrared targets. The detection accuracy was significantly improved in the detection scene. The network in this paper is a reference value for pedestrians and vehicles driving at night. Our future work will explore a vehicle-mounted infrared camera equipped with the improved YOLO-I network, or equipped with the improved network in traffic. The camera is for pedestrians and drivers passing by at night, and hopes to improve safety.

Acknowledgements. This paper was supported by the Fundamental Research Funds for the Local Universities of Hei longjiang Province in 2018 (Grant No. 2018-KYYWF-1189) and Shanghai Aerospace Science and Technology Innovation Fund (Grand No. SAST2017-104).

References

1. Cui, M.: Application field and technical characteristics of infrared thermal imager. China Secur. Protect. **12**, 90–93 (2014)
2. Carlo, C., Salvetti, O.: Infrared: a key technology for security systems. Adv. Opt. Technol. **2012**, 838752 (2012)
3. ViolaI, P., Jones, J.M., Snow, D.: Detecting pedestrians using patterns of motion and appearance. Int. J. Comput. Vis. **63**(2), 153–161 (2005)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: International Conference on Computer Vision & Pattern Recognition (CVPR 2005), vol. 1, pp. 886–893. IEEE Computer Society (2005)
5. Felzenzwalb, P.F., Grishick, B.R., Mcallister, D., et al.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
6. Lecun, Y., Bottou, L., Bengio, Y., et al.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
7. Ning, S., Liang, C., Guang, H., et al.: Research on deep classification network and its application in intelligent video surveillance system. Electro Opt. Control **22**(9), 77–82 (2015)
8. Jensen, M.B., Nasrollahi, K.T., Moeslund, B.: Evaluating state-of-the-art object detector on challenging traffic light data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 9–15 (2017)
9. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
10. Ren, S., He, K., Girshick, R., et al.: Faster-R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2017). https://doi.org/10.1109/TPAMI.2016.2577031
11. Zhang, Z., Wang, H., Zhang, J., et al.: Aircraft detection algorithm based on faster-RCNN for remote sensing image. J. Nanjing Normal Univ. (Eng. Technol. Edn. **41**(4), 79 (2018). https://doi.org/10.3969/j.issn.1001-4616.2018.04.013
12. Yang, W., Wang, H., Zhang, J., Zhang, Z.: An improved algorithm for real-time vehicle detection based on faster-RCNN. J. Nanjing Univ. (Nat. Sci.) **55**(2), 231–237 (2019). https://doi.org/10.13232/j.cnki.jnju.2019.02.008
13. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
14. Redmon, J., Divvala, S., Girshick, R., et al.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
15. Ang, D., Jiang, Y.: Face recognition system based on BP neural network. Software **36**(12), 76–79 (2015)
16. Zhang, X., Yi, H.: Scene classification based on convolutional neural network and semantic information. Software **39**(01), 29–34 (2018)
17. Gao, W., Li, Y., Zhang, J., et al.: Research on forecast model of high frequency section of urban traffic. Software **39**(2), 81–87 (2018)