Improved YOLOv4 Infrared Image Pedestrian Detection Algorithm

Jin Tao[1](✉), Jianting Shi[2], Yinan Chen[2], and Jiancai Wang[3]

[1] Graduate College, Heilongjiang University of Science and Technology, Harbin 150022, China
taojin@usth.edu.cn

[2] School of Computer and Information Engineering, Heilongjiang University of Science and Technology, Harbin 150022, China

[3] Academic Affairs Office, Heilongjiang University of Science and Technology, Harbin 150022, China

Abstract. Because pedestrians are always in the active state, each target is at a different distance from the camera, resulting in a certain difference in the size of similar targets in the figure. Therefore, an infrared pedestrian detection algorithm is proposed in the paper based on Yolov4 algorithm. Aiming at the problems of low recognition rate and high background influence in infrared image downlink human small target detection, the network structure of YOLOv4 is optimized. Compared with YOLOv4 and YOLOv3, the mean Average Precision is improved by 0.53% and 1.05%, which improves the detection accuracy in a certain extent.

Keywords: Infrared image · YOLOv4 · Pedestrian detection · Network structure · YOLOv3

1 Introduction

With the development of artificial intelligence technology and deep learning, computer vision is becoming more and more mature, and pedestrian detection technology has entered people's life. There is a broad prospect in the fields of security monitoring, vehicle unmanned driving and human behavior analysis etc. [1–12].Pedestrian detection technology is to study and judge the given image or verify whether there are pedestrians to be detected in each frame of video sequence, and the specific location of the target can be found accurately and quickly.

The traditional visible light technology can't be applied in the fields of night or unmanned driving. Infrared thermal image is based on the relative temperature information of the object, compared with the traditional situation, which is less affected by various additional factors and can be applied in many aspects. But there is no color in the image collected by infrared equipment, so the accuracy of pedestrian detection is low. Pedestrian detection algorithm can be divided into traditional algorithm and deep learning algorithm. The traditional algorithm mainly uses the artificial design to extract the image features, combined with machine learning to recognize and classify the image

X. Jiang (Ed.): MLICOM 2021, LNICST 438, pp. 226–233, 2022.
https://doi.org/10.1007/978-3-031-04409-0_21

features, so that to detect the target. However, the traditional algorithm is complex, sometimes it is difficult to design a reasonable method in the complex scene, and the weight parameter is difficult to get a more accurate value, so the generalization ability is poor.

In recent years, Convolutional Neural Network (CNN) [7] has made a significant breakthrough in pedestrian detection. Convolutional Neural Network (CNN) can learn the original features of the target through a large number of data automatically. Compared with the manually designed features, CNN has stronger abilities in discrimination and generalization [8]. In the meanwhile, the deep learning algorithm not only improves the detection efficiency, but also improves the detection speed, which is better than traditional method. Before the appearance of YOLO, the detection speed of deep learning was not very fast, and the real-time performance could not be guaranteed, especially in the future unmanned driving technology. Redmon et al. [5] proposed YOLO (you only look once, unified, real-time object detection) algorithm, which entered the field of one stage target detection. The idea of one stage solves the problem of speed in target detection, and improves the real-time performance with a certain accuracy greatly. Although the speed is improved, but the accuracy is not more precisely than other algorithms. After that, there are YOLOv2, YOLO9000 and YOLOv4. The network structure of YOLOv4 is simple and efficient, so it's easy to deploy and widely used. It is one of the preferred algorithms in many commercial fields. Combined with the practical application scenarios, it is applied to large-scale outdoor monitoring to detect the areas where pedestrians are forbidden. Moreover, YOLOv4 has great application prospects in infrared images of small object detection and pedestrian detection. The backbone network is better used in the network structure such as DarkNet or RESNET classifier, but also detected quickly. The most important thing is to build a simple environment, reduce the background detection error and make a strong generalization. Although YOLOv4 network has great advantages in multi-scale prediction and better classifier, there are several disadvantages that the accuracy of object recognition is poor and the recall rate is low, compared with other RCNN series detection algorithms.

To solve the above problems, the YOLOv4 algorithm is improved in the paper. The pedestrian detection accuracy (MAP) of infrared image is improved by 0.04%, compared with the original YOLOv4 algorithm.

2 Algorithm Structure of YOLOv4 Network

YOLO algorithm is the first work of one-stage detection. It is a target detection system based on single neural network proposed by Redmon and Ali Farhadi in 2015. In CVPR in 2017, Redmon and Ali Farhadi published YOLOv2 and YOLOv3, which further improved the accuracy and speed. After further improvement, YOLOv4 algorithm appeared. YOLOv4 is mainly introduced in three aspects: network input, structure and output, the network structure is shown in Fig. 1.

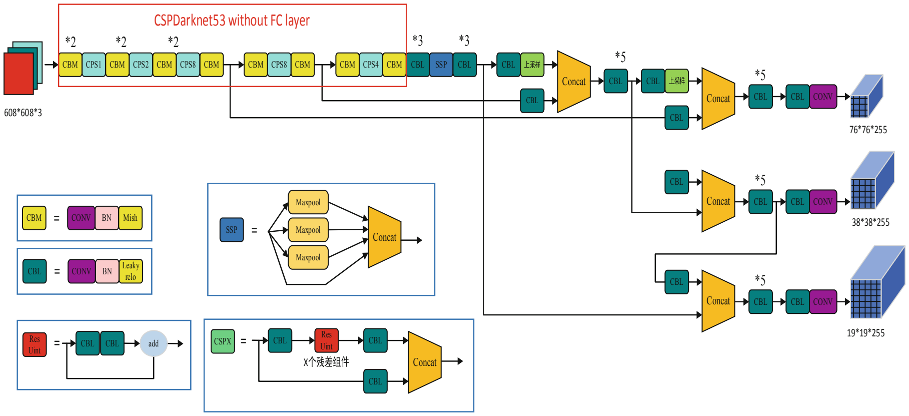

Fig. 1. YOLOv4 network structure.

3 Improved Infrared Pedestrian Detection Algorithm Based on YOLOv4

In the night vision infrared pedestrian detection and location task, first, the sampling machine is in the high spot to make the volume of the pedestrian target smaller; Second, because of the active pedestrians, each target is at a different distance from the camera, resulting in the differences of similar targets in the images. These two factors lead to a certain deviation between the final detection results and the actual situation. So the structure of the feature extraction network is optimized to enhance the ability of the network to capture the target position.

In the feature extraction network CSPDarknet adopted by YOLOv4, standard convolution of 3 × 3 size is mainly used for feature extraction. Because the shape and size of the receptive field of standard convolution are fixed, it will also extract the features of non-target areas when detecting small targets, which leads to more interference factors in the features extracted by final convolution and more interference effects on the prediction of the detector. So in the actual detection situation, deformation convolution is used as the core component, the deformation feature extraction module is constructed to improve the effectiveness of target feature extraction based on the standard convolution of YOLOv4. Compared with standard convolution, deformable convolution has the following advantages: firstly, the efficiency of receptive field is improved, that is, the feature map is more accurate in mapping target information; Secondly, the effectiveness of feature extraction in convolution kernel is improved. Convolution kernel can adapt to the position of the target for sampling, and the extracted feature information matches the target better; In addition, the deformable convolution can extract features more specifically for the region of the target, the stability of the feature graph (that is, the weight parameter will not change) is better than the standard convolution. When the feature graph is transferred in the network, the deformation process during model training can be expressed by the following formula and the deformation convolution used in the paper.

$$y(p) = \sum_{k=1}^{K} w_k \cdot x(p + p_k + \Delta p_k) \cdot \Delta m_k \tag{1}$$

p—Convolution kernel coordinates.
k—The number of convolution kernels (for example, 3 × 3 convolution kernels with 9 kernels).
w—Weight.
p_k—The offset of the Kth kernel.
Δp_k—The offset of the model needs to learn.
Δm_k—Offset control parameters that the model needed to learn.

The feature extraction network module is optimized based on deformation convolution. The composition of the optimized deformation feature extraction module is shown in Fig. 2.

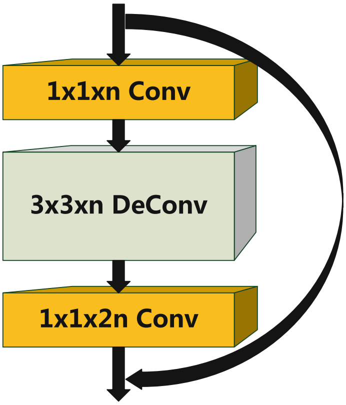

Fig. 2. Deformation feature extraction module.

Compared with the module before optimization, the optimized module mainly uses 3 × 3 deformation convolution to replace the standard convolution, and uses 1 × 1 convolution layer to construct feature channel for dimension reduction and dimension elevation. In the beginning, a 1 × 1 standard convolution layer is used to reduce the dimension of the input feature map, at the same time, the redundant features are deleted; Then, the feature map is extracted by 3 × 3 deformation convolution; Finally, 1 × 1 standard convolution is used to enhance the dimension of the extracted feature to increase the amount of information. In order to enhance the reuse of target location information, coordinate attention mechanism module is added based on the attention mechanism of YOLOv4 to enhance the coordinate information. Coordinate attention mechanism module optimizes based on SE channel attention mechanism, extracts the horizontal and vertical feature weight information of feature graph, and achieves accurate target position coordinate saliency mark through aggregation.

In order to optimize the location accuracy of the anchor, the "Guided Anchoring" mechanism is added in the detection layer of YOLOv4 to improve the quality of the anchor and the candidate region through the adaptive generation of the network anchor. The core principle of guided anchoring is to decompose the location attribute, the location attribute of a target is usually by four parameters (x, y, w, h) to represent location and size. That is to say, if the position information of a target can be expressed as $p(x, y, w, h|F)$ in the feature graph F, it can be decomposed into $p(x, y|F)p(w, h|x, y, F)$. This decomposition method shows that the definition of the position information of a target is to determine the existing region firstly, the shapes and sizes are closely related to the regional coordinates. Guided Anchors includes two branches, one is responsible for the prediction of anchor center coordinates, and the other is responsible for the prediction of anchor shape. The structure of anchor generator is shown in Fig. 3.

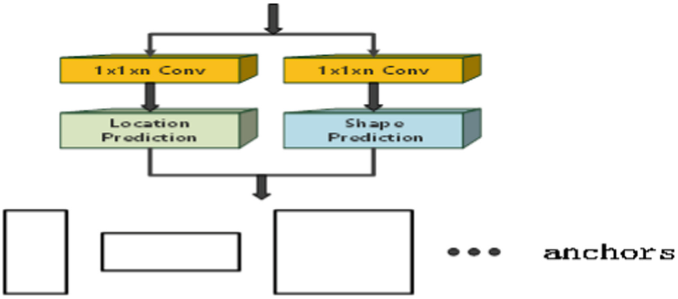

Fig. 3. Anchor network structure.

The main function of the center coordinate prediction branch is to determine whether regions in the feature map may have the center points of anchors, which is the binary classification problem. The specific implementation method is that the input feature map is converted into score map through 1×1 convolution layer, and then the final probability map is obtained by activating the sigmoid function of the elements on the map in the way of element-wise. At the same time, the threshold value ε_L is set to select the area where there may be anchors center point. Taking the point on the feature graph F as the anchor center point (i, j), the probability value is $p(i, j|F)$, and mapping it back to the coordinates in the input image is $((i + 1/2)s, (j + 1/2)s)$, S is the down sampling step of the feature graph related with the original image, by which the point on the feature graph can be mapped to the size of the original image for detection and output.

4 Experimental Results and Analysis

4.1 Evaluating Indicator

The related performance indicators of the infrared pedestrian detection algorithm, such as intersection and union ratio, IOU, precision and recall, are used to evaluate the advantages and disadvantages of infrared pedestrian detection. According to the size of the corresponding value to judge the quality of the model.

$$\mathrm{IOU} = 2 * \mathrm{area}(\mathrm{S}_1 \cap \mathrm{S}_2)/(\mathrm{area}(\mathrm{S}_1) + \mathrm{area}(\mathrm{S}_2)) \tag{2}$$

$$\mathrm{Recall} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \tag{3}$$

$$\mathrm{P} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}) \tag{4}$$

S1—Pedestrian area predicted by infrared image.
S2—Pedestrian area marked by people;
TP—The correct prediction of infrared image downlink;
FN—The situation of wrong prediction;
FP—It's not a pedestrian area, but it's predicted to be pedestrian.

4.2 Experimental Steps and Innovation Analysis

The data sets used in the experiment is from the OSU Thermal Pedestrian Database. Before used, the data set is cleaned, and 1500 ordinary samples, 400 difficult samples, and 200 negative samples are selected to form a 2100 training set; 300 ordinary samples and 200 difficult samples are used as 500 test sets.

According to the two training models, the improved YOLOv4, YOLOv4and YOLOv3 are compared and tested. The test results are shown in Table 1.

Table 1. Model checking performance comparison

Indexes	Precision	Recall	F2-1 score	IOU	mAP
YOLOv3	0.90	0.77	0.83	64.75%	82.04%
YOLOv4	0.92	0.78	0.86	63.92%	82.56%
Improved YOLOv4	0.89	0.85	0.87	64.86%	83.09%

Among them: the meaning of each index in Table 1 is as follows: precision represents the proportion of the part that the classifier considers to be a positive class and is indeed a positive class in all classifiers. Recall represents the proportion of the part that the classifier thinks is a positive class and is a positive class in all the positive classes. F1score calculation formula: 2 * precision * recall/(precision + recall). IOU (intersection and union ratio) represents the overlap ratio of the candidate bound and the ground truth bound, that is the ratio of their intersection and union. The ideal situation is complete overlap, that is, the ratio is 1. The mean accuracy (mAP) represents the average value of each category of AP. From the analysis in Table 1, the column F2-1score shows that the overall robustness and recall rate of the improved YOLOv4 algorithm are better than YOLOv3, which comprehensively reflects that the optimization of backbone network and detection network is of great help in improving network performance. The improved YOLOv4 algorithm is used for pedestrian detection, and the test results are shown in Fig. 4.

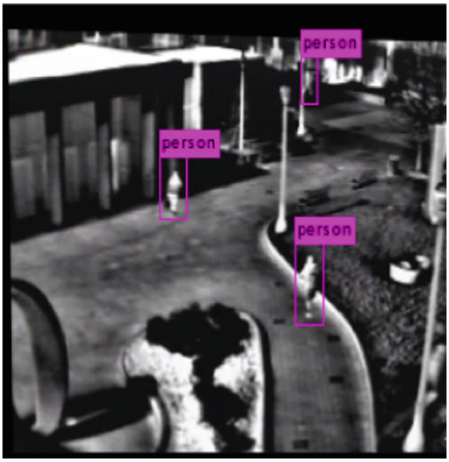

Fig. 4. Test results with improved YOLOv4 algorithm.

In order to verify the comprehensive performances, three algorithms are compared with the same training sets and the same testing sets to obtain the ROC curve. The comparison of improved YOLOv4(Im-YOLOv4), YOLOv4 and YOLOv3 are shown in Fig. 5.

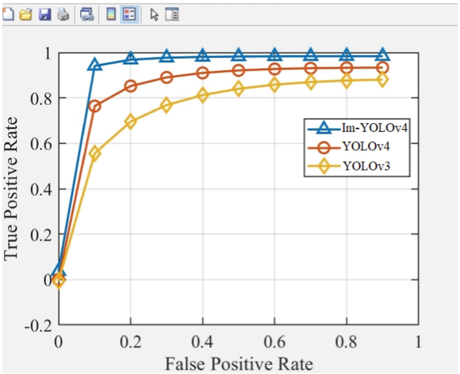

Fig. 5. The comparison of comprehensive performances.

5 Conclusion

In the paper, an improved infrared pedestrian detection algorithm based on YOLOv4 is proposed. The optimized algorithm improves the detection ability of gray image, small target, and the practicability of infrared detection. The deformation convolution is used as the core component, and the deformation feature extraction module is constructed to enhance the effectiveness of target feature extraction, to strengthen the ability of feature information transmission, and effectively improve the detection accuracy.

Acknowledgements. This work has been partially supported by "The National Key Research and Development Program of China (2018 YFC0810500)" and "2020 scientific research project of basic scientific research expenses of provincial colleges and universities in Heilongjiang Province. (2020-KYYWF-0680)".

References

1. Jensen, M.B., Nasrollahi, K., Moeslund, T.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. 9–11 (2017)
2. Girshick, R: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
3. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN, towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 92–199 (2015)
4. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Computer Vision – ECCV 2016. ECCV 2016. (LNCS), vol. 9905. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
5. 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)
6. Mukai, L., Tao, Z., Wennan, C.: Research on infrared pedestrian small target detection technology based on YOLOv4. In: Infrared Technology, pp. 1002–18891 (2020)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: International Conference on Computer Vision & Pattern Recognition (CVPR 2005), IEEE Computer Society, vol. 1, pp. 886–893 (2005)
8. Yao, Y., Wang, N.: Fault diagnosis model of adaptive miniature circuit breaker based on fractal theory and probabilistic neural network. Mech. Syst. Signal Process. **142**, 106772 (2020)
9. Sandler, M., Howard, A., Zhu, M., et al.: MobileNetV2: inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE (2018)
10. Lecun, Y., Bottou, L., Bengio, Y., et al.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
11. Felzenzwalb, P.F., Grishick, R.B., Mcallister, D., et al.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
12. Shi, J., Zhang, G.: Improved infrared pedestrian detection algorithm based on YOLOv3. J. Heilongjiang Univ. Sci. Technol. 21–37 (2020)