Real-time Pedestrian Detection Algorithm Based on Improved YOLOv3

Xianchang Xi, Zhikai Huang*, Lingyi Ning and Yang Zhang

1 College of Information and Engineering, Nanchang Institute of Technology No. 289 Tianxiang Avenue, Changdong Town, Nanchang County, Nanchang, China.
E-mail: huangzhik2001@163.com

Abstract. As a research hotspot in the field of current computer vision, pedestrian detection is widely applied to many fields, such as video surveillance and autonomous driving. However, the accuracy of pedestrian detection under video surveillance is poor, and the miss rate of small target pedestrians is high. In this paper, an improves the YOLOv3 algorithm and a YOLOv3-Multi pedestrian detection model had been proposed. First, referring to the residual structure of DarkNet, the shallow features and deep features had been up-sampled and connected to obtain a multi-scale detection layer. Then, according to different special detection categories, the spatial pyramid pool (SPP) is introduced to strengthen the detection of small targets. The experimental results show that our method improves the average accuracy by 2.54%, 6.43% and 8.99% compared with YOLOv3, SSD and YOLOv2 on the VOC dataset.

Keywords: Pedestrian detection; YOLOv3; Multi-scale detection; SPP.

1. Introduction
In the research of target detection, pedestrian detection has become a difficult and hot topic, because the image of pedestrian is affected by background occlusion, attitude and different shooting angle. For the problem of pedestrian detection, the current research methods of pedestrian detection are mainly divided into the traditional machine learning method and the deep learning methods. In terms of classic machine learning method. Gong, et al. proposed an algorithm based on mixed Gaussian background modeling combined with directional gradient histogram and SVM for classification [1]. Through the following three steps, such as, foreground segmentation, feature reduction and information updating, the finally false detection rate was reduced to 4%, at the same time, it showed good real-time performance and accuracy in complex scenes. Although researchers have made a lot of improvements in detection accuracy for target detection, there is still much room for improvement in detection speed and environmental impact by pedestrians with different postures.

With the development of artificial intelligence, the mainstream pedestrian detection method is based on deep learning algorithm. It is mainly divided into two categories: one is the two-stage target detection algorithm based on classification, that is represented by R-CNN, Faster R-CNN [2], Hypernet [3] and Mask R-CNN [4], and the other is the one-stage target detection algorithm that using regression algorithm by YOLO, SSD [5], G-CNN [6] and RON [7]. In recent years, breakthroughs have been made in the application of target detection algorithms based on deep learning in pedestrian detection. Han, et al. a video proposed a model based double stream network, which improves the detection accuracy and reduces the false detection rate in a small scale [8]. However, when dealing with complex background and occlusion for objects detection, the accuracy of detecting small objects still needs to be improved.
This paper focuses on the problem of the low accuracy of small object detection [9] in traffic camera and the high rate of missing detecting the small target pedestrians in real-time pedestrian detection. To meet the demand for higher requirement of real-time and detection speed, in here, an improve YOLOv3 algorithm which using DarkNet residual structure idea, combined the characteristics of shallow and deep on the characteristics of sampling to get multi-scale detection layer [10]. In this way, a fusion layer of different sizes of target location information and semantic information can be extracted. Finally, the prediction accuracy of the different scales of target can be improved by increasing the multi-scale fusion layer [11]. At the same time, the spatial pyramid pooling (SPP) module is used to achieve feature fusion at different scales, and improve the detection accuracy.

2. YOLOv3 Algorithm

The model of YOLOv3 is composed of two parts: the backbone network DarkNet-53 and the other is detection network. The backbone network DarkNet-53 draws on the residual idea of Resnet to improved the problem of gradient disappearance under training of convolution neural network and makes the model more easily convergent. There are 5 residuals in the residual structure. The residuals block consists of multiple residuals, mainly including the convolution layer, the batch normalization layer (BN), and the activation function (Leaky ReLU). Among them, the convolution layer is mainly employed for feature extraction, and the extracted features have been normalized, and the activation function is a nonlinear processing, which can effectively fit the nonlinear model. In the forward propagation of image convolution, the size transformation of the tensor is realized by changing the step size of the convolution kernel. There are 53 convolution layers and 23 jump connections. The detection network is composed of three YOLO layers, up-sampling, several Concat layers and convolutional layers, and the whole network level is summed to 107 layers. YOLOv3 outputs three feature graphs with the size of 13x13, 26x26 and 52x52, corresponding to the features of deep layer, middle layer and shallow layer respectively. Deep feature maps have small size and large receptive field, which is conducive to the detection of large scale objects. On the contrary, shallow feature maps are more convenient for the detection of small scale objects.

3. Improved Module

3.1. Spatial Pyramid Pooling

The improved YOLOv3-Multi algorithm uses SPPNet network structure like reference [12]. The idea is that SPP can generate output of fixed size at arbitrary input size, and can pool the features extracted from various scales. In addition, SPP uses multi-layer space box, while sliding window pooling only uses one window size. And multi-level pooling has strong robustness to the deformation of objects [13]. The SPP module is referenced in the original YOLOv3 network structure. The specific structure of the SPP is shown in figure 1.

![Figure 1. SPP module structure diagram.](image-url)
In the input convolution layer, there are four branches. One branch is direct output, and the other three branches are sampling under the maximum pooling of 5x5, 9x9, 13x13. Finally, the feature layer of the same scale can be obtained. The feature fusion of different scales can be realized to obtain richer feature expressions and improve detection speed and accuracy.

3.2. Improvement of Multi-Scale Prediction Layer

In view of the problem that small targets are easily missed due to occlusion and long distance in the process of pedestrian detection, the YOLOv3-Multi algorithm have been proposed that integrates shallow features and middle-level features with the help of multi-scale prediction idea of YOLOv3 algorithm. On the basis of the original YOLOv3 network structure, the deep feature map is enlarged to the same size as the shallow feature map through the up-sampling operation. Then the new scale target detection layer is constructed through the connection operation. On the basis of the original network, a 104x104 scale detection layer be added. Compared with other scale detection layers, the image is divided into finer units, which can detect smaller objects and hence improve the detection effect of small targets. The detail of YOLOv3-Multi network structure used in this paper has shown in figure 2.

![YOLOv3-Mult network](image)

**Figure 2.** YOLOv3-Mult network.

On the basis of adding a 104x104 scale detection layer, the candidate box sizes obtained by clustering analysis of the real box in the data set are applied to the 13x13, 26x26, 52x52, 104x104 scale detection layer for target detection. Because a large feature image has a small receptive field and strong ability to detect small targets, the candidate frame with small size is suitable for a large feature image. Because a small feature image has large receptive field and is relatively sensitive to large targets, the candidate frame with large size is suitable for a small image.

4. Experimental Results and Analysis

4.1. Experimental Environment and Training Parameter Setting

In the experimental platform built in this paper, the server CPU is Intel Xeon Gold 6240R, and the GPU are Nvidia Tesla M40*2. The operating system is Ubuntu20.04, and the memory of the graphics card is 12GB*2.
Pascal VOC dataset had been selected for our experiments. The experiments are mainly based on VOC 2012 training set and verification set to train the validation and VOC 2007 test set to train the test model. In all experiments, only pedestrians are trained and tested. The default input size is of image 416x416. The input channel is 3. We set the number of iterations is 40000. The batch size is 64, and the learning rate is 0.001.

4.2. Evaluation Index and Result Analysis
In here, pedestrian detection is carried out based on intersection cameras to judge whether the target is pedestrian or not. In order to evaluate the improved model more accurately and in real time, Average Percision (AP) is selected as the evaluation index [14]. AP is the result from the sum of all the precision rates of this class in the verification set divided by the number of images containing the target of this class. Its evaluation indexes comprehensively consider P(Precision) and R(Recall) to solve the single point value limitation of P and R. Then the definition criteria of AP evaluation indexes are shown in equation:

\[ AP = \frac{\sum Precision_c}{images_c} \]  

(1)

Where \( \sum Precision_c \) is the sum of all precision rates of the class in the verification set while \( images_c \) is the number of images containing the target of the cases.

The YOLOv3-Multi network proposed in this paper was compared with YOLOv3, SSD and YOLOv2 networks to analyze the variation trend of the average loss function of the improved model. And the P-R curve and AP value will be compared. The variation trend of the average loss function is shown in figure 3.

As can be seen from figure 3, the loss function value is relatively large at the beginning. But as the number of training iterations increases, the loss value decreases rapidly and gradually converges. When the training reaches 40,000 steps, the loss value is always stable at about 0.1, and the convergence degree of the model achieves the ideal effect and the model training is stable.

![Figure 3. Average loss curve.](image)

In order to evaluate the model more accurately, the YOLOv3-Mulit network proposed in this paper is compared with YOLOv3, SSD and YOLOv2 networks to calculate the recall rate and accuracy of various algorithms respectively. P-R curve is drawn in figure 4. The experiment demonstrates that the improved algorithm improves both recall rate and accuracy.
The plane area formed by P-R curve and coordinate axis is the calculation of AP. The larger the area is, the larger the AP value is. Table 1 below summarizes the AP value of the proposed YOLOv3-Multi network, YOLOv3, SSD and YOLOv2 network on VOC dataset.

| Network classes | AP      |
|-----------------|---------|
| YOLOv3          | 79.25   |
| SSD             | 75.36   |
| YOLOv2          | 72.80   |
| YOLOv3-Multi    | 81.79   |

As can be seen from table 1, the YOLOv3-Multi algorithm proposed that used our experiments could get much improvement in pedestrian detection. It had been compared with YOLOv3, SSD and YOLOv2, the AP value increases by 2.54%, 6.43% and 8.99%, respectively.

In this paper, a trained YOLOv3-Multi model and YOLOv3 model are used to collect images at road intersections for testing and comparison. As shown in figure 5, the YOLOv3-Multi model can detect pedestrians at intersections well and mark the position of detected objects, which is more accurate in the detection of small targets. Compared with the YOLOv3 network, it greatly reduces the rate of missed detection.
5. Conclusion
The YOLOv3-Multi pedestrian detection network model proposed in this paper is based on making changes in YOLOv3 network model. By add SPP module layer and increase the multi-scale prediction, this model can make accurate pedestrian detection for complex road conditions, and improve the efficiency of the pedestrian detection. The result show that YOLOv3-Multi network model in the VOC 2007 test set improves the average accuracy by 2.54%, 6.43% and 8.99% compared with YOLOv3, SSD and YOLOv2 network model. The test achieves good results. However, since the pedestrian detection accuracy of YOLOv3-Multi algorithm in this paper is insufficient, its detection performance still needs to be optimized. Therefore, the next step of research will focus on optimizing the loss function and optimizing the setting of Anchor class to improve the detection accuracy.

Acknowledgement
This project was supported by National Key Research and Development Plan of China (No.2019YFB1704502), the National Natural Science Foundation of China (Grant No. 61472173), the grants from the Educational Commission of Jiangxi province of China, No. GJJ151134, the Innovation Fund Designated for Graduate Students of Jiangxi Province, No. YC2020-S633.

References
[1] Gong L M, Xu M H, Liu D J, et al. 2018 Novel model of pedestrian detection based on Gaussian mixture model and HOG+SVM Journal of Shanghai University (Natural Science Edition) 24(3): 341–351.
[2] Ren S, He K, Girshick R and Sun J 2015 Faster R-CNN: Towards real-time object detection with region proposal networks Advances in Neural Information Processing Systems 28.
[3] Kong T, Yao A, Chen Y and Sun F 2016 HyperNet: Towards accurate region proposal generation and joint object detection Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: 845-853.
[4] He K, Gkioxari G, Dollar P, et al. 2017 Mask R-CNN Proceedings of the IEEE Interactional Conference on Computer Vision: 2961-2969.
[5] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S E, Fu C Y, Berg A C 2016 SSD: single shot multibox detector European Conference on Computer Vision: 21–37.
[6] Najibi M, Rastegari M, Davis L S 2016 G-CNN: An iterative grid based object detector Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) DOI: 10.1109/CVPR.2016.260.
[7] Kong T, Sun F, Yao A, et al. 2017 RON: Reverse connection with objectness prior networks for object detection IEEE Computer Society I: 2.
[8] Xie H and Shin H 2021 Two-stream small-scale pedestrian detection network with feature aggregation for drone-view videos Multidimensional Systems and Signal Processing: 1-17.
[9] Li J N, Liang X D, Wei Y C, et al. 2017 Perceptual generative adversarial networks for small object detection Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[10] Redmon J and Farhadi A 2019 Yolov3: An incremental improvement arXiv:1804.02767. 2018.
[11] Li Y, Chen Y, Wang N, et al. 2019 Scale-aware trident networks for object detection IEEE.
[12] He K, Zhang X, Ren S, Sun J 2015 Spatial pyramid pooling in deep convolutional networks for visual recognition IEEE Trans. on Pattern Anal. Mach. Intell. 37(9): 1904–1916.
[13] Lazebnik S, Schmid C, Ponce J 2006 Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories Proceedings of CVPR.
[14] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, Silvio Savarese 2019 Generalized intersection over union: a metric and a loss for bounding box regression Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition CVPR: 658-666.