American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)
ISSN (Print) 2313-4410, ISSN (Online) 2313-4402

http://asrjetsjournal.org/

Detection of Pedestrians and Helmets in Large Construction Site

Chunming Tang[a]*, Fan Yang[b], Xiang Yu[c]

[a]*School of Artificial Intelligence Institute, TianGong University, Tianjin, 300387, CHINA*
[b]*School of Electronics & Information Engineering, TianGong University, Tianjin, 300387,CHINA*
[c]*Country Center for Engineering Practice and Training, TianGong University, Tianjin, 300387,CHINA*
[a]*Email: tangchunminga@hotmail.com*
[b]*Email: 307652409@qq.com*
[c]*Email: 1922489185@qq.com*

Abstract

It is necessary for workers to wear helmets when working in large construction sites. The traditional way to supervise the workers whether wearing helmets or not for safety is artificial, which have brought out many problems such as many blind spots, labor and time costing. Since a large number of surveillance cameras are currently installed in these construction sites, the surveillance video can be developed in taking the place of human supervision in an intelligent way. This paper designs a pedestrian and helmet detection network based on Faster R-CNN. In feature extraction, we have chosen the Residual Network (Resnet) combined with the Feature Pyramid Network (FPN) because the objects have small size, low resolution and less semantic information in whole scenes. We have also designed a parallel residual Block (PRB) combined with the Receptive Field Block (RFB) to strength feature extraction. The feature maps obtained from different convolution layers have been fused twice. And we have studied two fusion methods. Experiment results from our own dataset show that our proposed detection network improves the mAP by 8.74% and 2.3% respectively compared with Yolov3 and Faster R-CNN, at the cost of 0.3 FPS slower than Faster R-CNN.

Keywords: pedestrians and helmets' detection; large construction site; Faster R-CNN; Parallel Residual Block.

---

* Corresponding author.

1. Introduction

The helmet and pedestrian detection are classified into close shot and whole scene detection. Close shot detection is easy to implement, and many existing detection networks can solve it. While whole scene detection, like in harbor, the states of art of detection networks do not work. Therefore, our paper focuses on solving the problem. In the harbor's whole scene surveillance view, the workers are usually very small compared to the size of the whole image, the minimal value is 0.13% in occupied area proportion. And their helmets are even smaller. The difficulties in small objects detection are their low resolution and little semantic information. The detection networks are mainly two categories so far which is a two-stage region proposal object detection, such as R-CNN [1], Fast R-CNN [2], Faster R-CNN [3], R-FCN [4] and a one-stage single object detection, such as SSD [5], YOLO [6] and RetinaNet [7]. The characteristics of the former is higher accuracy and slower speed while the latter is on the contrary. It is reasonable to choose the latter in engineering application. As small objects effect on the detection accuracy immensely, we choose the former to develop while sacrificing the speed.

2. Related Works

2.1. Traditional Pedestrian and Helmet Detection algorithm

Traditional pedestrian detection algorithms include gradient histograms of interest [8], real-time pedestrian detection system based on Adaboost [9] and some others. Traditional helmet detection algorithms include skin color detection and Hu moment [10], color and shape-based helmet recognition [11] and some others. All these have been implemented on the close shot detection while failing in some whole scene construction sites.

2.2. Pedestrian and Helmet Detection based on Deep Learning

Compared with the traditional detection algorithms, deep learning can usually obtain higher detection accuracy, more accurate Bounding Box borders, and lower missed and false detection rate. Some networks are applicable to pedestrian detection [12] or helmet detection [13] in relatively close shot with fairly satisfactory results so far. But few networks can detect pedestrians and helmet at the same time. We improve the Faster R-CNN to detect pedestrians and helmets simultaneously with small size in low-resolution.

3. Approach

The structure of our network based on Faster R-CNN for the detection of pedestrians and helmets is shown in Figure 1, which the feature extraction network is improved and the ROI Pooling is replaced by ROI align [14], ROI align performs better in small object detection.

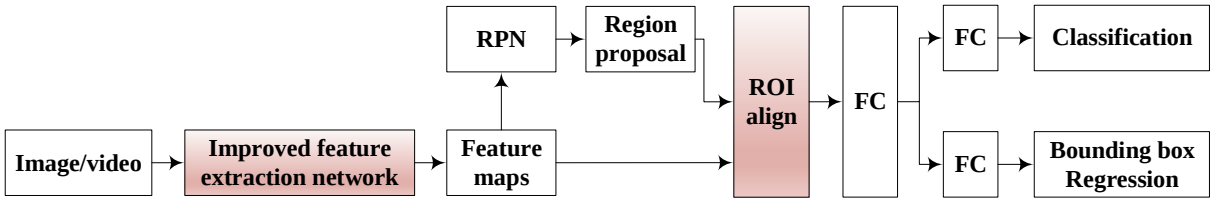

Figure 1: Structure of our improved network

3.1. Improved Feature Extraction Network

As shown in Figure 2, our improved feature extraction network is still based on Resnet50 [15] and FPN [16]. The Receptive Field Block (RFB[17]) and our proposed Parallel Residual Block (PRB) are used to increase the receptive field of the feature maps. And we also propose a multi-branch design for small object's detection. M2, M3 and M4 are the results of the first fusion. After PRB, the second fusion is performed to further enhance the semantic information by fusing high-level features after up-sampling with the low-level features, and obtain N2, N3 and N4. 3*3 convolutions have been used finally to reduce the dimension of the feature map, and obtain the features' outputs of P2, P3, P4 and P5. These four outputs also achieve the purpose of multi-layer prediction.

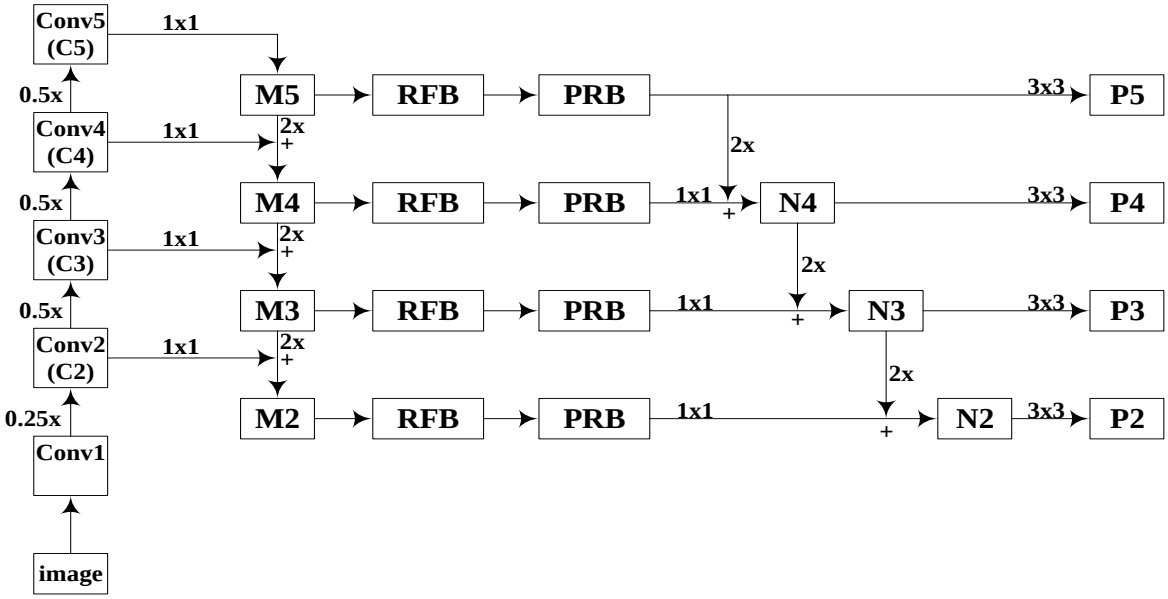

Figure 2: Improved feature extraction network

3.2. Parallel Residual Block (PRB)

The proposed PRB shown in Figure 3 is to deepen the depth and width of the feature extraction network to alleviate the degradation and gradient dissipation caused by the increase of the network depth which consists of three residual blocks. Multi-branch structure is shown in Fig. 3. There are a 3*3 and two 3*3 convolutions in the first and second branch, and a 3*3 convolution and a 3*3 dilated convolution [18] with rate set to 2 in the third branch apart from 1x1 convolutions. These three branches have different receptive fields, and finally being

fused in channels. The features' fusion of different receptive fields can enhance their shallow information in small objects' detection.

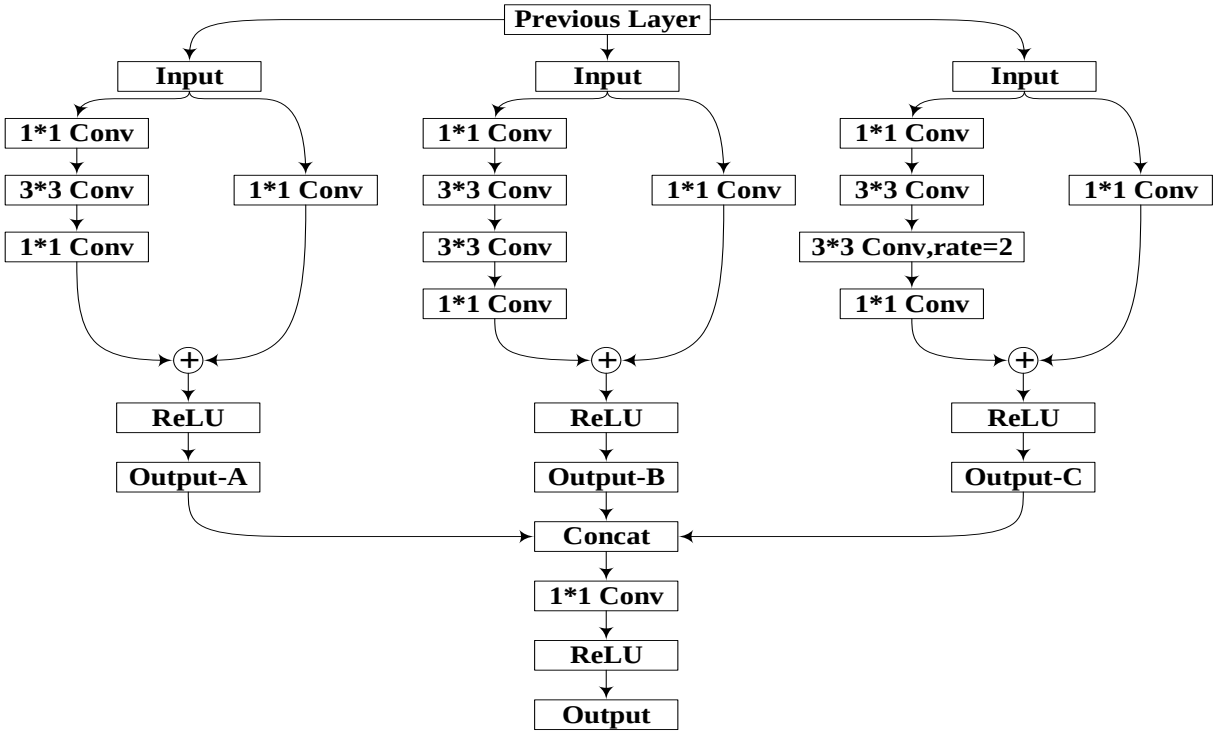

Figure 3: Structure of PRB

3.3. Generated feature maps

Figure 4 are an input image and its generated feature maps corresponding to each module in training. The first column is the C2, C3, C4 and C5 from bottom to top. The second to the fifth columns are the first fusion, RFB, PRB and the second fusion. The sixth column is P2, P3, P4, and P5 from bottom to top.

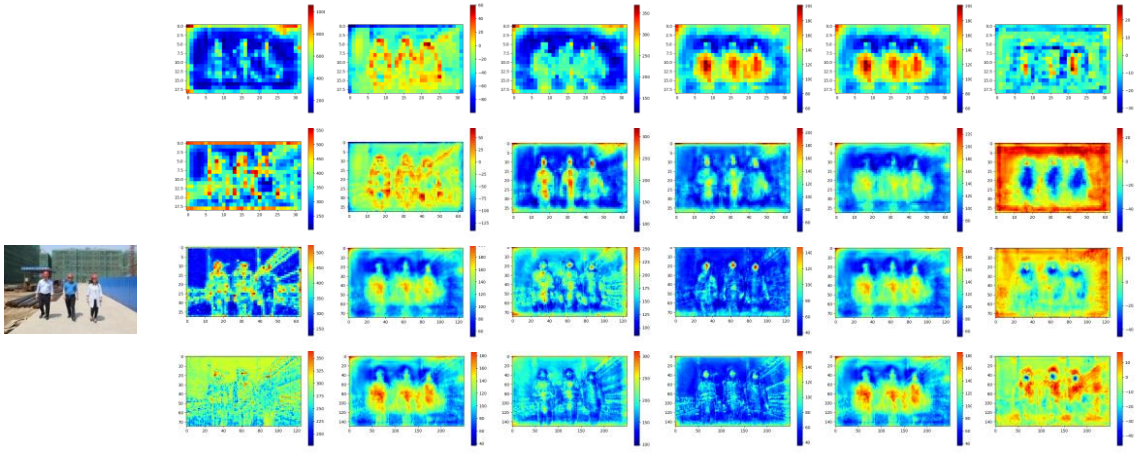

Figure 4: Input image and its generated feature maps

3.4. Research on two Fusion methods

In convolutional neural networks, the features generated by the multi-branch structure are often fused before the final output. Some features may be clear but have little semantic information, while some features may be vague but have a lot of semantic information, that is why the fusion is needed. There are two basic fusion methods. One is point-wise addition, and the other is vector concatenation. We have done some experiments to compare the effect of the two methods.

4. Experiments

4.1. Dataset and Experimental Environment

The dataset and the corresponding labels are all made by ourselves, which are characterized by large construction sites and small objects. There are three object categories in the dataset, namely people, head regions with helmet and without helmet. There are 3900 images in total, including 38559 objects. 3000 images are in training with 31664 objects, and 900 images in testing with 6895 objects. The average number of objects in per image is close to 10, and there are about 20% of relatively small objects, and the detection task is rather tough. This network is implemented by our workstation equipped with NVIDIA GeForce GTX 1080, which operating system is Ubuntu 16.04.6 LTS. CPU is an 8-core 16-thread Inter Xeon CPU E5-2650 v2@2.20GHz processor. The CUDA version is 8.0.

4.2. Comparisons of experimental results

We choose one-stage and two-stage feature extraction networks, Darknet53's YOLOv3 and Faster R- CNN's Resnet50 combined FPN's algorithm as comparisons. The experimental results are shown in Table 1.

Table 1: The comparisons of experimental results of three networks

Networks	Person	Head with helmet	Head without helmet	mAP	FPS
YOLOv3	0.7426	0.7157	0.6985	0.7189	**25**
Resnet50+FPN(Faster R-CNN)	0.8045	0.7574	0.7881	0.7833	5
Proposed (ours)	**0.8295**	**0.7707**	**0.7943**	**0.7982**	4.7

It can be seen from Table 1 that Faster R-CNN has improved 6.44% on mAP compared to YOLOv3. Our improved network has 1.49% higher in mAP and 0.3 slower in FPS than the original Faster R-CNN.

4.3. Experiments on Fusion Methods

Since the feature extraction network shown in Fig. 2 uses two feature fusion operations, which are the NMS and Soft-NMS. The NMS is to sort the detection boxes by scores firstly, and then the box with the highest score is kept, while other overlapping boxes more than a certain proportion are excluded. Soft-NMS is special for a

certain case in NMS with overlapping regions and high scores, which leads to one of the boxes being excluded. The Soft-NMS is to reduce the score of the boxes by a Gaussian function to solve partial occlusion. The eight experimental results are shown in Table 2.

Table 2: The eight experimental results

Exp	The 1st fusion		The 2nd fusion		NMS	Soft-NMS	person	head with helmet	head without helmet	mAP
	add	concat	add	concat						
Exp.1	√		√		√		0.8295	0.7707	0.7943	0.7982
Exp.2	√		√			√	0.8323	0.7731	0.7958	0.8004
Exp.3	√			√	√		0.8192	0.7727	0.8051	0.7990
Exp.4	√			√		√	0.8247	0.7788	0.8058	0.8031
Exp.5		√	√		√		0.8317	0.7623	0.7916	0.7952
Exp.6		√	√			√	0.8343	0.7609	0.7951	0.7968
Exp.7		√		√	√		0.8293	0.7796	0.8017	0.8035
Exp.8		√		√		√	0.8310	0.7834	0.8045	**0.8063**

Eight combinations are formed by the add or concatenation of NMS or Soft-NMS fusions. It is found that the mAP values using the Soft-NMS are higher than that using the NMS. The highest mAP value is using concat. Compared with the original Exp.1 also in Table 1, the highest mAP value is increased by 0.81%. Figure 5 shows the PR curve with a mAP value of 0.8063. Three actual detection scenarios are shown in Figure 6-8.

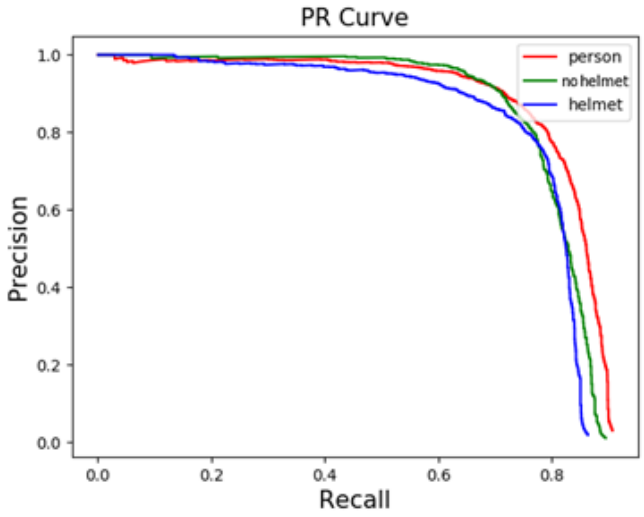

Figure 5: PR curve with mAP value of 0.8063

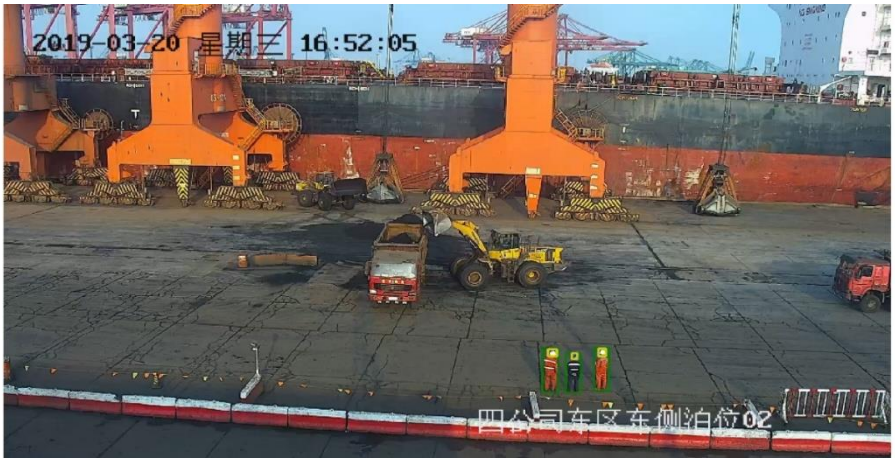

Figure 6: Detection result in harbor

Figure 7: Detection result in construction site

Figure 8: Detection result in the case of multiple objects and occlusion

5. Conclusion

We focus on solving the problems of small object detection in large construction sites. The feature extraction network is our mainly developed part which the PRB and the RFB blocks are added, and two fusions are performed. The improved feature extraction network increased the mAP value by 1.49%. We have also conducted experiments on the two fusion methods. And find that concatenation performs better than add. Finally, our proposed network improves mAP by 2.3% compared to the original network in actual pedestrian and helmet detection scenarios. But this network has certain limitations. When the objects are very small or occluded, it is hard to be detected.

References

[1]. V Zeiler, Matthew D, and R. Fergus. "Visualizing and Understanding Convolutional Networks."European Conference on Computer Vision Springer (2014).

[2]. Girshick, Ross. "Fast R-CNN." Computer ence (2015).

[3]. Ren, Shaoqing, et al. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." IEEE Transactions on Pattern Analysis & Machine Intelligence 39.6(2017):1137-1149.

[4]. Zhang, Hanwang, et al. "PPR-FCN: Weakly Supervised Visual Relation Detection via Parallel Pairwise R-FCN." 2017 IEEE International Conference on Computer Vision (ICCV) (2017).

[5]. Liu, Wei, et al. "SSD: Single Shot MultiBox Detector." European Conference on Computer Vision Springer International Publishing (2016).

[6]. Redmon, Joseph, et al. "You Only Look Once: Unified, Real-Time Object Detection." Computer Vision & Pattern Recognition IEEE (2016).

[7]. Lin, Tsung Yi, et al. "Focal Loss for Dense Object Detection." IEEE Transactions on Pattern Analysis & Machine Intelligence PP.99(2017):2999-3007.

[8]. Gan, Guolong, and J. Cheng. "Pedestrian Detection Based on HOG-LBP Feature." Seventh International Conference on Computational Intelligence & Security IEEE (2012).

[9]. Guo, Lie, et al. "Pedestrian detection for intelligent transportation systems combining AdaBoost algorithm and support vector machine." Expert Systems with Applications 39.4(2012):4274-4286.

[10]. Xiao-Hui, Liu, and Y. E. Xi-Ning. "Skin Color Detection and Hu Moments in Helmet Recognition Research." Journal of East China University of ence and Technology (2014).

[11]. Zhang, Geng, et al. "The Method for Recognizing Recognition Helmet Based On Color and Shape." International Conference on Machinery (2017).

[12]. Ouyang, Wanli, and X. Wang. "Joint Deep Learning for Pedestrian Detection." IEEE International Conference on Computer Vision IEEE (2014).

[13]. Rohith, C A, et al. "An Efficient Helmet Detection for MVD using Deep learning." 2019 3rd International Conference on Trends in Electronics and Informatics (2019).

[14]. Kaiming, He , et al. "Mask R-CNN." IEEE Transactions on Pattern Analysis & Machine Intelligence PP(2017):1-1.

[15]. He, Kaiming, et al. "Deep Residual Learning for Image Recognition." IEEE Conference on Computer Vision & Pattern Recognition IEEE Computer Society (2016).

[16]. Lin, Tsung Yi, et al. "Feature Pyramid Networks for Object Detection." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) IEEE Computer Society (2017).

[17]. Liu Songtao, Huang Di, Wang Yunhong. "Receptive Field Block Net for Accurate and Fast Object Detection." 2017 IEEE Conference on Computer Vision and Pattern Recognition (2017).

[18]. Zhang, Xiaohu, Y. Zou, and W. Shi. "Dilated convolution neural network with LeakyReLU for environmental sound classification." 2017 22nd International Conference on Digital Signal Processing (2017).