A Novel Safety Helmet Wear Detection Method based on Improved RFBnet

Liu Chen¹, a, Zhishuang Xue¹, b, Xiaofang Liu²*, Mingju Chen¹, c, Hui Huang¹, d

¹School of Automation and Information Engineering, Sichuan University of Science and Engineering, Zigong, China
²School of Computer Science and Engineering, Sichuan University of Science and Engineering, Zigong, China

a chen950916@foxmail.com
b 465223378@qq.com
* lxf1969@163.com
c 12347259@qq.com
d schuang_hui@163.com

Abstract—In order to solve the problems of error detection and missed detection in complex environments, a helmet wear detection method based on dense connection method and CIou loss is proposed. This method is based on RFBnet-512 neural network, and improves the backbone network of RFBnet-512 with dense connection to enhance the backbone network’s ability to extract features. At the same time, in order to avoid errors caused by inconsistent loss function and evaluation index, CIou loss is used to replace the original SmoothL1 loss of RFBnet-512 network, making the calculation of regression loss more close to the evaluation index. The experiment shows that compared with the original RFBnet-512 network, the detection accuracy is improved by 0.018, and the detection speed is significantly improved.

1. INTRODUCTION

Nowadays, production safety has drawn more and more attention from the society, and various enterprises have formulated strict production safety rules and regulations. Even so, in recent years, there have been a succession of safety accidents in the field of construction, mainly due to the weak safety awareness of workers who do not wear helmets and the difficulty of artificial safety supervision. In recent years, with the rapid development of computer vision, target detection methods for image and video emerge one after another. As a kind of target detection, safety helmet detection has important application values.

With the rapid development of neural network and the advantage of convolutional neural network in automatically learning image features, more and more researchers break through the difficulties faced by traditional image processing with the help of convolutional neural network. At present, the target detection algorithm based on neural network has achieved good detection effect and has been applied to some fields of society maturely. Target detection methods can be roughly divided into two categories: one is a two-stage detection network, such as Fast Region-Based Convolutional Network.
(Faster R-CNN) [1], Faster Region-Based Convolutional Network (Faster R-CNN) [2], Cascade Region-Based Convolutional Network (Cascade R-CNN) [3], et al. These detection networks include two stages of region extraction and region classification, with high detection accuracy, but the speeds are slow due to the relatively complex detection process. The other is the one-stage detection network, such as You Only Look Once version 3 (YOLO v3) [4] and Single Shot Multi-Box Detector (SSD) [5]. Besides, a large number of excellent detection methods based on SSD network appears, like Deconvolutional Single Shot Detector (DSSD) [6], Receptive Filed Block Net (RFBnet) [7], Multi-Level Feature Pyramid Network (M2Det) [8]. These detection networks have a high detection speed, but they are inferior to the two-stage detection networks in detecting small targets.

To sum up, in order to ensure the accuracy and efficiency of safety helmet wear detection, this paper proposes an improved RFBnet network, dense connection method is introduced to deepen RFBnet bankbone network, which strengthens the bankbone network feature extraction ability. At the same time, due to the small number of detection categories in this paper, the network detection speed can be accelerated by appropriately reducing the number of network channels. In order to further improve the detection accuracy of the proposed network, CIou loss [9] is used to replace the original SmoothL1 loss of RFBnet network, so as to avoid the error caused by inconsistent loss function and evaluation index.

2. THE PROPOSED DETECTION METHOD

In order to solve the problems of error detection and missed detection of existing safety helmet detection methods in complex environments, this paper uses an improved RFBnet-512 network based on one-stage detection. RFBnet-512 refers to the image with an input size of 512, and similarly, the RFBnet-300 network with an input size of 300. In order to make better use of the connection between different features, the dense connection of VoVnet network [10] is introduced into the bankbone network of RFBnet-512. In order to avoid errors caused by inconsistent loss function and evaluation index, CIou loss is used to replace the original SmoothL1 loss, making the calculation of regression loss more close to the evaluation index.

2.1 Densely Connected RFBnet-512

In the target detection neural network, different convolutional layers have different semantic information. For a common image, after repeated convolution and pooling in the neural network, the object feature information in the image will be lost to different degrees, resulting in that the deep feature map is not suitable for small targets detection. RFBnet-512 network which simulates the human visual system in the receptive field structure can improve the detection performance of one-stage network, and at the same time avoid producing more computational burden. It mainly uses VGG-16 network to extract image features, however, as the structure of VGG-16 network is obtained by stacking the convolutional layers and pooling layers, the feature loss between different convolutional layers is not taken into account, resulting in the very deep convolutional layer losing the semantic information from the shallow convolution. Aiming at the feature loss problem, Densenet [11] adopts the connection mode of dense blocks to realize the multi-layer features fusion, so as to make full use of the features of each layer and ensure the high efficiency of feature transferring. However, due to Densenet's connection is too dense, each convolutional layer will aggregate the features of the previous layers, resulting in feature redundancy in the deep network. In order to avoid feature redundancy, VoVnet optimizes the connection mode of Densenet by only aggregating all the feature layers before the last one-off aggregation, thus achieving efficient aggregation of different convolutional layers. The dense connection of VoVnet network is shown in Fig. 1.

Figure 1. The dense connection of VoVnet
In order to make the feature utilization of each layer more efficient, this paper uses the dense connection of VoVnet to improve the backbone structure VGG-16 of RFBnet-512 network, so that the feature extraction ability can be enhanced. For target detection, especially for small targets, high-resolution feature map of shallow network is needed to provide location information of small targets, and rich semantic information of deep network is needed to effectively classify small targets. The dense connection network can transfer the information useful for small target detection to the deep network. These feature extraction and transferring modes make the deep feature map of VGG-16 have the low semantic information from the shallow feature map, which make the improved RFBnet-512 network enhance the small target detection capability. Since this paper only detects the wearing of the helmet, compared with the COCO dataset, there are fewer detection categories. In order to reduce the weight of the model and improve the detection efficiency, the 1024 convolutional channels in the original RFBnet-512 network are reduced to 512 channels. The improved network in this paper is named DRFBnet.

2.2 CIou Loss is Introduced to DRFBnet

The loss function of target detection is consists of two parts, classification loss and regression loss. The regression loss in the original RFBnet-512 network adopts smooth L1loss, and the four points in the regression box are used to calculate the actual regression loss, but the detection and evaluation method uses Intersection over Union (IoU) to calculate the score, so as to predict the optimal detection box. We assume 4 points are independent of each other without considering their correlation. Therefore, in this paper, CIou loss is used to replace the original DRFBnet SmoothL1 loss, making the calculation of regression loss more close to the evaluation index.

Generally, the loss function based on IoU can be defined as:

\[ L = 1 - \text{IoU} + R(B, B') \]  

(1)

Where, \( R(B, B') \) is the penalty item of the prediction box \( B \) and the target box \( B' \). The overlapping area, center distance and length-width ratio of prediction box and target box are incorporated into the loss function to obtain CIou loss, and the penalty term of CIou loss is defined as follows:

\[ R_{\text{CIou}} = \frac{\rho^2 (b, b')}{c^2} + \alpha \nu \]  

(2)

Where \( \rho(\cdot) \) denotes the Euclidian distance between the prediction box and the target box, \( c \) denotes the diagonal distance of the minimum external rectangle between the prediction box and the target box, and \( \alpha = \frac{\nu}{(1 - \text{IoU}) + \nu}, \nu = \frac{4}{\pi^2} (\arctan \frac{\omega'}{h'} - \arctan \frac{\omega}{h}) \) are the parameters used to measure the consistency of length-width ratio. Finally, the CIou loss function is defined as follows:

\[ L_{\text{CIou}} = 1 - \text{IoU} + \frac{D^2 (b, b')}{c^2} + \alpha \nu \]  

(3)

3. EXPERIMENT AND ANALYSIS

3.1 Model Training

After a large amount of data training, we can verify the model. The training dataset is the key to train an excellent model. Since it is difficult to collect the image data in construction environments, the open source dataset is used to train and test the model in this paper. This dataset labels a person with a helmet as hat and a person without a helmet as person. There are 6,878 images under the dataset used in this paper, among which 6,328 are used for training, and 550 are used for testing.

In order to prevent the over-fitting of the model due to insufficient training data, the training sample images are randomly cropped, enlarged, distorted and mirror transformed to ensure sufficient training samples for model training. Initial training is done in Warmup style and six epochs are trained. The
optimization function is SGD with momentum=0.9, and the weight attenuation coefficient is 0.0005. The steps learning rate adjustment method (Gamma =0.1) is adopted, and the initial learning rate lr=0.004. In addition, in order to reflect the detection performance of DRFBNnet model in this paper, RFBnet-512, RFBnet-300 and YOLO v3 are trained with hat-det dataset in the same experimental environment, and corresponding training models are obtained.

3.2 Training Results and Comparative Analysis

A small number of representative sample images are selected to test YOLO v3, RFBnet-300, RFBnet-512 and the proposed DRFBNnet detection method. Fig. 2(a) gives original test images. The test images have occlusion, background interference and targets with different sizes, which can fully test the detection performance of the model. (b), (c), (d) and (e) in Fig. 2 are the results obtained from YOLO v3, RFBnet-300, RFBnet-512 and DRFBNnet, respectively. For the first image, when there is occlusion target (the red arrow in the figure), other three comparison methods cannot detect the target indicated by the arrow. Among them, YOLO v3 has error detection. For the remote small targets in the second image (indicated by the red arrow), neither RFBnet-300 nor RFBnet-512 can detect them, while YOLO v3 and DRFBNnet can detect them correctly, which proves that the proposed DRFBNnet network enhances the detection ability of small targets. For the third image, due to the crowded targets, it can be seen visually that RFBnet-300 has obvious missed detection, while the other three detection models can achieve relatively good detection effect.
In order to prove the good performance of DRFBnet, DRFBnet is verified from two aspects: detection accuracy and speed. The hat-det test dataset is used to test YOLO v3, RFBnet-300, RFBnet-512 and DRFBnet, respectively. The test results are shown in Table 1. Compared with YOLO v3, RFBnet-300 and RFBnet-512, DRFBnet has the highest average accuracy (MAP), and the MAP value improves 0.011, 0.054 and 0.018 respectively, indicating that the proposed model improves the detection accuracy to a certain extent. In addition, although DRFBnet detection speed is slower than YOLO v3, compared to the FRBnet-512 network, the speed is 51% higher.

| Models      | Numbers of test images | Speed (frame/s) | MAP   |
|-------------|------------------------|-----------------|-------|
| YOLO v3     | 550                    | 39              | 0.814 |
| RFBnet-300  | 550                    | 24              | 0.771 |
| RFBnet-512  | 550                    | 18              | 0.807 |
| DRFBnet     | 550                    | 27              | 0.825 |

4. CONCLUSIONS
This paper proposes a safety helmet wear detection method based on improved RFBnet-512 network. Aiming at the problem of safety helmet wear detection in complex industrial environment, this paper enhances the feature extraction ability of the backbone network through dense connection, and reduces the number of model channels, thus effectively improving the detection accuracy and speeding up the detection speed. Experimental results show that compared with RFBnet-512, the improved DRFBnet safety helmet wear detection method has the advantages of better detection accuracy and higher detection speed.

ACKNOWLEDGMENT
This research was supported in part by Sichuan Science and Technology Program under Grant 2017GZ0303, in part by Fund Project of Sichuan Provincial Academician (Expert) Workstation under Grant 2016YSGZZ01, in part by Information Technology and Internet of Things Measurement and Control Technology Sichuan University Key Laboratory Open Fund under Grant 2019WZY04, and in part by Sichuan Education Department Project Fund under Grant 17ZB0302.

REFERENCES
[1] R. Girshick, "Fast r-cnn," Proceedings of the IEEE international conference on computer vision. IEEE, 2015, pp. 1440-1448.
[2] S. Ren, et al, "Faster r-cnn: Towards real-time object detection with region proposal networks," Advances in neural information processing systems. Curran Associates, 2015, pp. 91-99.
[3] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving Into High Quality Object Detection," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018, pp. 6154-6162.
[4] J. Redmon and A. Farhadi, "YOLOv3: an incremental improvement,"IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2018, pp. 89-95.
[5] W. Liu, D. Anguelov, D. Erhan, et al, "SSD: Single shot multibox detector,"European Conference
on Computer Vision. IEEE, 2016, pp. 21-37.

[6] C. Fu, W. Liu, A. Ranga, et al, "DSSD: Deconvolutional single shot detector,"Computer Vision and Pattern Recognition. IEEE, 2017, pp. 2881-2890.

[7] S. Liu, D. Huang and Y. Wang, "Receptive field block net for accurate and fast object detection,"Computer Vision-ECCV. Springer, 2018, pp. 404-419.

[8] Q. Zhao, T. Sheng, Y. Wang, et al, "M2det:A single-shot object detector based on multi-level feature pyramid net-work,"Proceedings of the AAAI Conference on Artificial Intelligence. 2019, pp. 9259-9266.

[9] Z. Zheng, P. Wang, W. Liu, et al, "Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression," arXiv, 2019.

[10] Y. Lee, J. Hwang, S. Lee, Y. Bae and J. Park, "An Energy and GPU-Computation Efficient Backbone Network for Real-Time Object Detection," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2019, pp. 752-760.

[11] G. Huang, Z. Liu, L. van der Maaten, et al, "Densely connected convolutional networks,"Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, 2017, pp. 4700-4708.