Application Effect Analysis for Helmet Detection Algorithm based on YOLOV5

Churui Jia *

Henan University, Kaifeng, Henan, 475001, China

* Corresponding author email: hena@henu.edu.cn

Abstract. Safety helmets can provide better safety protection for the human head in industrial production, which is one of the important protective tools to protect the lives of workers. In the on-site environment, checking whether construction workers are wearing safety helmets mainly depends on manual inspection, which is time-consuming and labor-intensive. In order to solve the problem of hard hat detection and recognition in the complex environment of the construction site, in this paper, we propose a YOLOV5 hard hat detection method. Specifically, a pre-trained object model is used to locate the area of the worker and the face detection module is used to mark the face. Then, the helmet recognition module is used to extract the sub area of the worker's head. Finally, the extracted image is classified into two categories to judge the situation of the worker wearing the helmet. Focusing only on workers can reduce training time and improve the efficiency of the model, so the training label only uses the label of people. The model refines and extracts the head, inputs the separated head area into the subsequent network for subsequent analysis, and then uses the binary classification method to judge whether the helmet is worn. We compare the results of different models such as YOLOV5s, YOLOV5m and YOLOV5l, where YOLOV5l can achieve a best mAP of 91.9%. We also analyze the effect of light to the performance, where the detection effect will be affected by low visibility through the comparative detection of samples in the day and at night.

Keywords: Helmet Detection; YOLOV5; Deep Learning.

1. Introduction

With the rapid development of infrastructure in China, accidents on construction sites have also gradually increased. The construction site is a workplace with high safety risks and numerous potential safety hazards. There are 12 million injuries and even deaths caused by accidents on the construction site every year, which brings great harm to workers' families and social security. According to statistics, in the same accident, wearing a helmet can greatly reduce the rate of serious injury and mortality. Wearing safety helmets can effectively protect the head of construction personnel during engineering operation. However, there are still many workers working outside the site who do not wear safety helmets. Manual supervision by on-site security personnel alone is inefficient, and it is difficult to achieve full-area security coverage on site. Therefore, in order to improve the on-site safety factor and safety supervision efficiency, we urgently need a fast and accurate method for detecting the wearing of helmets.

With the development of technology, the detection of worker's safety helmet based on object detection has proved to be an effective solution to replace manual work. In early object detection, features are usually selected manually and classifiers are designed according to the detected objects. However, these methods are usually complex in design process and lack of versatility, and have great limitations in practical application. Due to the rapid development of deep learning, target detection based on convolutional neural network (CNN) is widely used. Girshick et al. [1-3] successively proposed R-CNN (Region CNN), Fast R-CNN and Fast R-CNN, which gradually improved the performance of PASCAL Visual Object Classes 2007 dataset. This algorithm has high accuracy, large computation and long running time [4]. Liu et al. used support vector machine (SVM) as the model, and realized helmet detection through skin color detection and location and Hu matrix as the feature vector. Wu [5] et al. proposed a mixed descriptor based on color, which is composed of local binary mode (LBP), Hu matrix invariant (HMI) and color histogram (CH) to extract the characteristics of

helmets with different colors. Shi Hui et al. used the pyramid structure of the image to obtain feature maps of different sizes to detect the location and category, and classify the parameters of the target frame. This method can identify the safety helmet worn by workers. Target detection algorithms can be divided into two categories. The first is two-step target detection algorithm. That is, candidate boxes are generated based on regions and then classified. The detection accuracy of this algorithm is relatively high, but it takes a lot of time. Typical algorithms mainly include R-CNN [3], FasterR-CNN, etc. The second type is single step target detection algorithm. The end-to-end method is used to predict the whole image, and the target position detection and classification are completed at the same time. Many researchers are committed to introducing the object detection into practical application scenarios. For example, Zhang et al. used RPN in Faster R-CNN to process small targets and negative samples, and used random forests to classify candidate areas, which greatly improved the poor detection effect in pedestrian detection. Song Huanhuan et al. [6] applied the improved RetinaNet to the field of vehicle recognition and realized rapid vehicle detection. Chen Lei et al. fused the feature output of the last layer of YOLOv3 network with the output of the previous layer to form a bidirectional pyramid feature layer, which improved the recognition effect of small and medium targets in remote sensing images. Deng Zhuanglai et al. detected and identified barn pests by compressing the SSD of the model, reducing grain loss. Zhang Haichuan et al. realized road crack detection by combining UNet++ with Conditional Generic Adversary Net (CGAN). It can be seen that deep learning has become a hot research field, and combining it with practical application scenarios has become a mainstream direction. Target detection is a task of image recognition in artificial intelligence, which needs to find the object of specific object target in the given image [7]. Common target detection models include Faster R-CNN [8], SSD [9], YOLO [10], etc.

Through previous efforts have significantly improve the helmet detection performance, we argue few of them pay attention to analyzing the application results in different scenes with various settings, such as lighting, background complexity, etc. In this paper, we construct a helmet detection model based on YOLOV5, which takes whether to wear a helmet as the detection task to realize the helmet detection system. The detection effect under different algorithms is further analyzed to verify whether low visibility will affect the detection effect of the system.

2. Method

2.1 Revisiting YOLOV5

YOLOv5 is a relatively advanced target detection network algorithm in YOLO, which is constantly strengthened and innovated on the basis of previous generations of yolo. Secondly, YOLOV5 has a good effect in the detection of objects by COCO and PASCAL VOCD, so this paper uses YOLOV5 to detect whether workers wear safety helmets. YOLOV5 has given four algorithms. Among them, YOLOV5s algorithm is the network with the lowest depth and Ap precision in YOLOV5 series, while YOLOV5m, YOLOV51 and YOLOV5x algorithms are the enhancement and expansion of YOLOV5s. YOLOV5 network structure is divided into input end, backbone network, neck network and prediction part. YOLOV5 adds mosaic data enhancement to the input part; The backbone part adopts Focus slice structure and CSP (Cross Stage Partial Network) structure; Add FPN+PAN structure to Neck; In the prediction, the loss function of the boundary anchor frame is improved from completely overlapping joint loss (CIoU) to generalized overlapping joint loss (GIoU); In the post-processing of target detection, YOLOV5 uses weighted NMS (non maximum suppression) operation to filter multiple target anchor frames. Based on YOLOv4, YOLOV5 adds a Focus structure for slicing to the backbone network. In YOLOV5s algorithm, the input size is 3 × six hundred and eight × 608 ordinary image, after a slice operation through Focus, the conversion size is 12 × three hundred and four × 304, and then pass 32 convolution cores × three hundred and four × 303's normal convolution operation converts it to a size of 32. Unlike YOLOV4 algorithm, which only uses CSP structure in the backbone network, YOLOV5 algorithm has designed two new CSP structures. Taking YOLOv5s algorithm as an example, the backbone network uses CSP1_ 1 Structure and CSP1_ 3

Structure, CSP2 as neck_ 1 structure to strengthen feature fusion between networks. The network structure of YOLOV5s is shown in Figure 1.

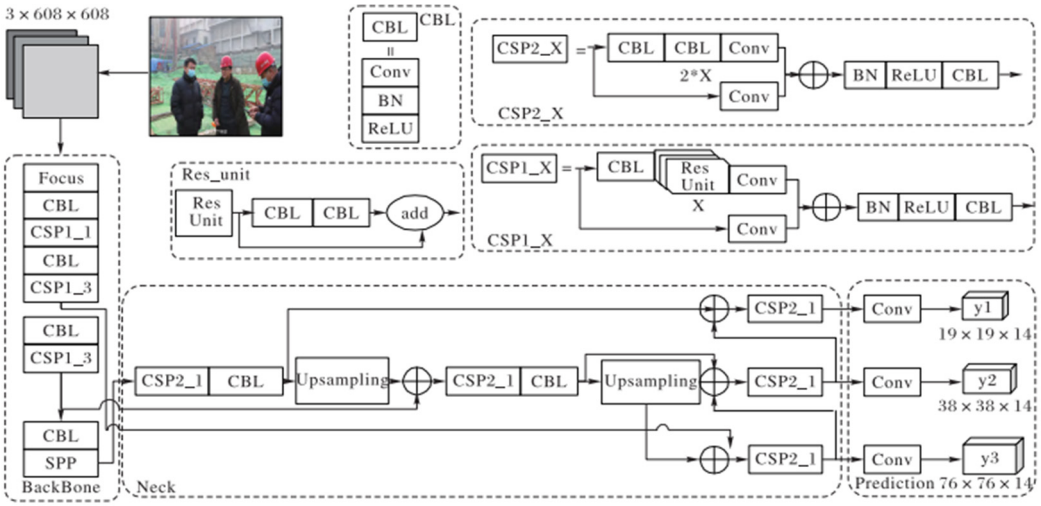


Fig 1. The framework of YOLOV5s network

2.2 Helmet Detection based on YOLOV5

Taking the head region of workers as the focus of detection, the neural network pays more attention to learning the head feature information of workers. As shown in Figure 2, the specific operation of this process is as follows: first, use the pre training object detection weight of YOLOV5 to locate the worker's area, use the face detection module to mark the face, then use the helmet recognition module to extract the worker's head sub area, and finally use the binary classification method for the extracted image to judge whether the worker wears a helmet.

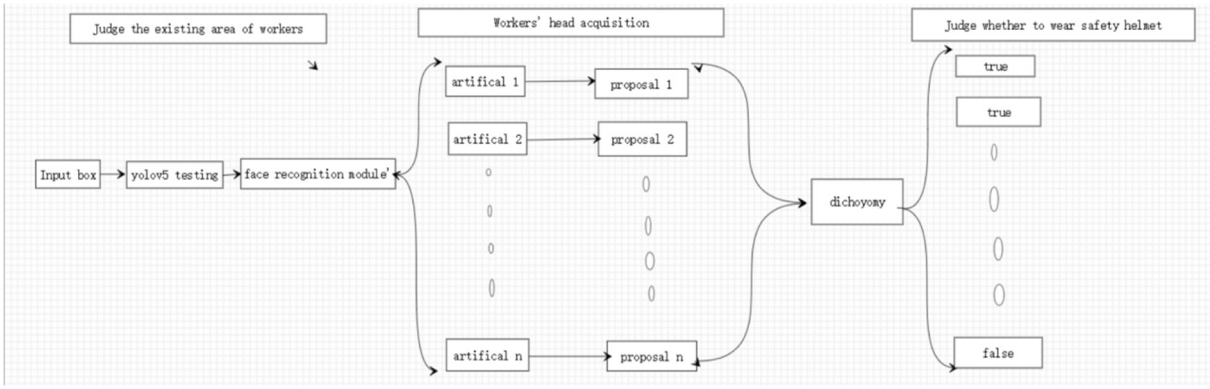


Fig 2. The framework of helmet detection based on YOLOV5s

Select YOLOV5 object detection framework to detect multiple types of target objects. The training tag uses the annotation tool (I use the wizard annotation assistant) to draw the ground truth, generate an xml file, and then complete the training process through code. In the helmet detection task in this paper, in order to make the model focus only on workers, it is required that only human tags are used for training tags, which can reduce training time and improve the efficiency of the model. YOLOV5 has also made changes in this structure compared with previous versions to prevent the gradient from disappearing after network deepening. Bottleneck adopts the idea of residual connection, whose structure is shown in Figure 3. SPP adopts the maximum pooling of 3×3, 5×5, and 7×7, and fuses different MaxPooling through Concat to achieve the best effect.

The detection of workers' heads is very important. Only a good head recognition can make a good helmet recognition. We completed the detection of human body through YOLOV5. The next step is to further refine the model to extract this area, input the separated head area into the subsequent

network for subsequent analysis, and finally determine whether the helmet is worn. However, when extracting the image of the area where workers exist, the size of the extracted area is often different. Therefore, this paper uniformly extracts in the upper middle part of each area, and fixes the extracted area as a square (the selected size is 60×60), the area shall not exceed the overall area of workers, but not less than the area of helmet. In this way, it can ensure that the extracted area contains the head area of workers, improve the fault tolerance rate, and do not increase the network computing. After extracting the head area of workers, the second classification method is used to judge whether they wear a helmet. This part of the function is realized by using the classical convolutional neural network.

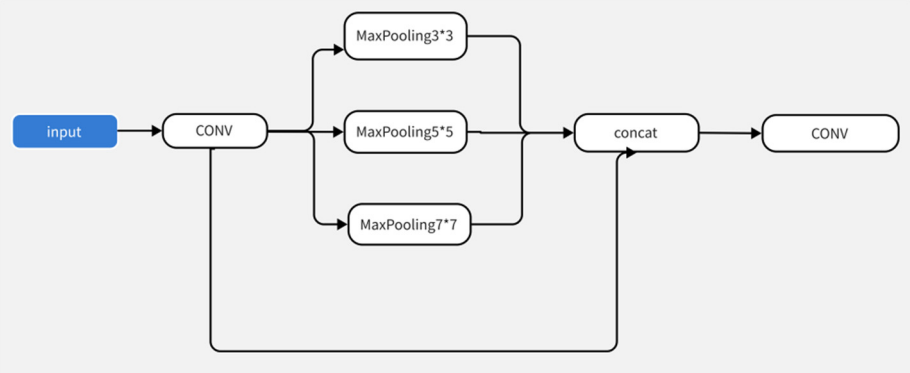


Fig 3. The structure of bottleneck in YOLOV5

3. Experiment and Result Analysis

3.1 Dataset

The quality of data sets is very important for the training of experimental models. In this paper, we select the data set from the GitHub, which is composed of the pictures of the site scene taken from the actual scene of the site. We use the wizard annotation assistant to annotate the image and save it in xml format. The data set introduction is as follows: This data set is used to detect the wearing of human head and helmet. It includes 7230 images, including 8596 images of workers wearing helmets and 109590 images of workers not wearing helmets. These images are from goolge or Baidu respectively. We use the wizard annotation assistant to manually mark them. Some negative objects come from SCUT-HEAD., And convert the data into txt format that can be used normally.

3.2 Experiment Settings

The environment required for the experimental system: Python 3.9, Python==1.5.7, PyQt5==5.15.3, PyQtChart==5.15.3, PyQt5 tools, GPUtil. The operating system and software environment are windows10+CUDA 10.2+cudnn 7.6.5+python 3.8.10+python 1.7.0. The hardware environment is GPU: NVIDIA GeForce GTX 1060, CPU: Intel (R) Core (TM) i7-6700, memory: 8.00GB, video memory: 6GB.

The model is trained with an SGD optimizer with an initial learning rate of 1E-2, a momentum of 0.937 and a weight decay of 5e-4. We report the results of YOLOV5 based on three backbones and 100 epochs in following subsection.

3.3 Evaluation Indicator

To evaluate the performance, we select the accuracy, recall and mAP as the evaluation metric. The accuracy and recall rate are calculated through the relationship between TP, FP and FN. The analysis of TP, FP and FN is as follows: TP: The detection frame whose IOU with the real frame is greater than the set threshold is regarded as the positive sample correctly identified by the model; FP: The detection frame whose IOU with the real frame is less than the set threshold is regarded as the positive

sample of model error recognition; FN: Target not identified as positive sample by the model (that is, the model does not detect). The calculation formula of precision P and recall R are as:

$$P = TP/(TP + FP) \tag{1}$$

$$R = TP/(TP + FN) \tag{2}$$

Then the mAP can be calculated based on the precision and recall, as formula (3):

$$mAP = \frac{\sum_{i=1}^{k} AP_i}{k} \tag{3}$$

3.4 Eexperimental Result

To evaluate the effectiveness of our methods, we compare the accuracy of three different backbones in Table 1. As shown in Table 1, three algorithms, YOLOV5s, YOLOV5m and YOLOV5l, are used for analysis and detection in the experiment.

Table 1. Comparison of different algorithms

	YOLOV5s			YOLOV5m			YOLOV5l		
	P	R	mAP0.5	P	R	mAP0.5	P	R	mAP0.5
Human	0.846	0.893	0.877	0.844	0.906	0.887	0.856	0.914	0.897
Head	0.889	0.883	0.871	0.900	0.911	0.900	0.893	0.913	0.901
Hemlet	0.917	0.921	0.917	0.913	0.929	0.916	0.927	0.929	0.919
Average	0.884	0.899	0.888	0.886	0.915	0.901	0.892	0.919	0.906

First of all, on the whole, the highest map 0.5 of YOLOV5l is 0.906, and the lowest map 0.5 of YOLOV5s is 0.888. It is the only algorithm without 0.9. The center of YOLOV5m is 0.901. The values of P and R have the same effect as that of map 0.5. They are both the lowest in YOLOV5s and the highest in YOLOV5l, which are 0.884 and 0.899 in YOLOV5s, 0.886 and 0.915 in YOLOV5m and 0.892 and 0.919 in YOLOV5l, respectively. The difference between P's YOLOV5s and YOLOV5m is only 0.002, the maximum difference between YOLOV5s and YOLOV5l is 0.8, the minimum difference between R's YOLOV5m and YOLOV5l is 0.004, and the maximum difference between YOLOV5s and YOLOV5l is 0.020. In general, the minimum value of p appears between YOLOV5s and YOLOV5m, while the minimum value of R appears between YOLOV5m and YOLOV5l, indicating that the accuracy rate of YOLOV5s and YOLOV5m is the closest, while the recall rate of YOLOV5m and YOLOV5l is the closest. Secondly, the data of human body were analyzed, and the minimum mAP0.5 of YOLOV5s was 0.877. The maximum mAP0.5 of YOLOV5l is 0.897, and the middle mAP0.5 of YOLOV5m is 0.887. It has the same effect as the whole. According to pr analysis, YOLOV5s0.846 0.893, YOLOV5m0.844 0.906, YOLOV5l0.856 0.914. The difference between P YOLOV5s and YOLOV5m is 0.002 and the minimum value is YOLOV5m, while the difference between R YOLOV5m and YOLOV5l is 0.008. The three groups of head mAP0.5 were 0.871 0.9 0.901, maintaining the same rule as above. PR data are 0.89 0.83, 0.9 0.911, 0.893 0.913 respectively. The accuracy rates of YOLOV5s and YOLOV5m are similar, and the recall rates of YOLOV5m and YOLOV5l are similar. Finally, look at the detection effect of the safety helmet. The detection values of the three groups are almost the same, and the detection effect is close to and above 0.9. The detection effect is good. Pr was 0.917 0.921, 0.913 0.929, 0.927 0.929, respectively. The accuracy rate of YOLOV5s and YOLOV5m is the closest to the recall rate of YOLOV5m and YOLOV5l. In the four classifications, the accuracy rates of yolv5s and YOLOV5m are similar, and the recall rates of YOLOV5m and YOLOV5l are similar. From the perspective of classification, the mAP0.5 from human body to helmet is getting higher and higher, indicating that the detection effect is gradually strengthened.

Then look at the P R values under the respective algorithms and categories. R is always greater than P, and the recall rate of the system is always greater than the accuracy rate. The probability of FN (manually marked but not detected) is relatively low. Finally, between categories, whether P, R or mAP0.5, the helmet is larger than the head and larger than the human body, and all data of the

helmet are above 0.9, which has excellent detection effect. We also visualize the detection results. As shown in Figure 1, the model can detect the human body and helmet.



(a) Original image (b) Detection result

Fig 4. Visualization of detection results based on YOLOV5

According to statistics, in many object detection applications, the detection effect will be greatly reduced or even undetected at night or in the case of low visibility, so will the safety helmet detection system based on YOLOV5 also be affected. To quantitatively analyze the model detection accuracy under different lighting scenarios, we also conduct a set of experiments, who's average IoU and mAP can be seen in Table 2.

Table 2. Detection performance under different lighting scenarios

	P	R	mAP	IoU
Day	0.875	0.882	0.871	0.594
Night	0.823	0.865	0.843	0.505
Total	0.848	0.875	0.864	0.549

Most of the daytime iou values can reach 0.64, with more samples of grade 0.7 or above, such as 0.717 0.786 0.704, and fewer unrecognized phenomena (None). In the nighttime iou values, most of the data are above 0.53, almost no more than 0.7, and there are multiple groups of samples of grade 0.5 or below, such as 0.47 0.451 0.486. There is a certain gap between the daytime iou values and the nighttime iou values, and there are relatively many unrecognized cases. In the map table, the map value in the daytime is 0.871, which is higher than the map value at night, 0.843. It shows that night and low visibility will indeed affect the detection effect of the system. Looking at the analysis of PR, it is 0.875 in the daytime, 0.882 at night, and 0.823 0.865 at night. It can be seen from the data that both the accuracy rate and the recall rate are greater in the daytime than in the night. In night images, image enhancement is generally required due to insufficient clarity, such as contrast enhancement and histogram equalization. Then, the histogram equalization image is binarized. For the form of binarization, we carry out operations such as expansion and corrosion, so as to achieve the diagnosis and analysis of connected areas and obtain potential goals. But the map in the daytime is 0.871, and the map at night is 0.843. It can be seen that there is no big difference in the impact. The reason is that although these pictures of site workers on the Internet are at night, they are more or less illuminated, not blurred, so there is no big difference.

4. Conclusion

This paper proposes how to implement the helmet detection system through YOLOV5, compare the AP accuracy of different algorithms and verify that low visibility will affect the detection effect of the project. Pre train the object detection weight, locate the area of the worker, mark the face with the face detection module, extract the head sub region of the worker, and finally use the binary classification method to judge the situation of the worker wearing the helmet. The training tag uses the wizard annotation assistant to draw the ground truth, generate an xml file, and then complete the training process through YOLOV5. The experimental results show that YOLOV5s network is the

smallest, with the lowest speed and the lowest AP accuracy. The accuracy rates of yolv5s and YOLOV5m are similar, and the recall rates of YOLOV5m and YOLOV5l are similar. From human body to helmet. The detection effect is gradually strengthened. In the case of low visibility, the detection effect will be reduced. In night images, image enhancement is generally required due to insufficient clarity, such as contrast enhancement and histogram equalization. Then, through the binary processing of the histogram equalization image, we can expand and corrode the binary image to achieve the diagnosis and analysis of the connected domain and get the potential target.

References

[1] Girshick R, Donahue J, Darrell T, et al. Rich featurehierarchies for accurate object detection and semantic segmentation[C]// Proceedings of the 2014 IEEE Conference on Computer Visionand Pattern Recognition. Piscataway: IEEE, 2014: 580-587.

[2] Girshick R. Fast R-CNN[C]// Proceedings of the 2015 IEEEInternational Conference on Computer Vision. Piscataway: IEEE, 2015: 1440-1448.

[3] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[4] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[J]. arXiv preprint arXiv: 1311. 2524, 2013.

[5] Hao Wu, Jinsong Zhao. An intelligent vision-based approach for helmet identification for work safety[J]. Computers in Industry, 2018, 100: 267-277.

[6] Zhang L L, Lin L, Liang X D, et al. Is faster R-CNN doingwell for pedestrian detection? [C]// Proceedings of the 2016European Conference on Computer Vision, LNCS, 9906. Cham: Springer, 2016: 443-457.

[7] JIao L C, Zhang F, Liu F, et al. A survey of deep learning-based object detection[J]. IEEE Access, 2019, 07: 128837-128868.

[8] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39 (06): 1137-1149.

[9] Liu W, Anguelov D, Erhan D, et al. SSD: sin-gle shot multibox detector: European conference on computer vision[C]. Amsterdam: Springer Press, 2016: 21-37.

[10] Redmon J, Divvala S, Girshick R, et al. Youonly look once: unified, real-time object detection: IEEE conference on computer vision & pattern recogni-tion[C]. Piscataway NJ: IEEE Press, 2016: 779-788.