2019 IEEE 3rd Information Technology,Networking,Electronic and Automation Control Conference (ITNEC 2019)

Safety Helmet Wearing Detection Based On Deep Learning

Xitian Long, Wenpeng Cui, Zhe Zheng

State Grid Key Laboratory of Power Industrial Chip Design and Analysis Technology
Beijing Smart-Chip Microelectronics Technology Co., Ltd
Beijing, China
longxitian12@163.com

***Abstract*—In many scenarios, such as power station, the detection of whether wearing safety helmets or not for perambulatory workers is very essential for the safety issue. So far, research in safety helmets wearing detection mainly focused on hand-crafted features, such as color or shape. With rising success of deep learning, accurately detecting objects by training the deep convolutional neural network (DCNN) becomes a very effective way. This paper presents a deep learning approach for accurate safety helmets wearing detection in employing a single shot multi-box detector (SSD). Moreover, because of safety helmet usually relatively small and unfortunately SSD struggles in detecting very small objects, a novel and practical safety helmet wearing detecting system is proposed, Finally, extensive compelling experimental results in power substation illustrate the efficiency and effectiveness of our work.**

Keywords—Object Detection; Deep Learning; CNN; SSD

I. INTRODUCTION

Safety helmets wearing detection is a key problem for power station surveillance. The original methods of safety helmet wearing detection using color and shape for proposal generation, which means creating proposals by hand-selected features.

Recently, deep learning based methods achieve significant results for general object detection problem. The state-of-the-art methods for general object detection can be categorized into one-stage methods (e. g., YOLO [1], SSD [2], Retinanet [3]), RefineDet [4]), and two-stage methods (e.g., Fast/Faster R-CNN [5], FPN [6], Mask R-CNN [7]). Fig.1 shows the processing speed and accurate of several popular object detect deep learning solutions, the test results is training based on PASCAL VOC dataset and evaluated by mAP (Mean Average Precision) which will reflect both precision and recall, is a very important indicator of object detection results.

Generally speaking, two-stage methods usually have better detection performance while one-stage methods are more efficient. In this paper, we focus on the one-stage detector, due to requirement about the detection speed in our scenes. YOLO [1] and SSD [2] are two representative one-stage detectors. YOLO has a relative simple architecture thus very efficient, but cannot deal with dense objects or objects with large scale variants. As for SSD, it could detect objects with different size from multi-scale feature maps. Moreover, SSD uses anchor strategy to detect dense objects. Therefore, compared with YOLO, SSD achieves a better detection performance. Besides, SSD structure is very efficient and don't need extra time for region proposal, it can easily achieve the speed of more than 20 FPS on the graphics processing unit (GPU) and even some embedded devices, have potential to deploying in camera. Due to the above advantages, SSD becomes a very practical object detector in industry, which has been widely used for many tasks.

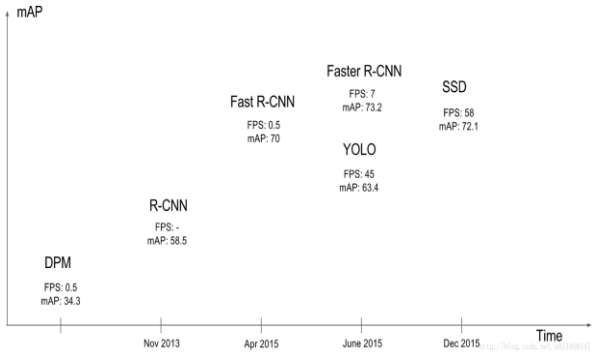

Fig. 1: Processing Speed and Accurate of several widely-used deep learning object detector , the FPS(frame per seconds) and mAP results is based on PASCAL VOC Dataset

Since SSD becomes first choice for safety helmet wearing detection, we should face its main drawback, the performance on small objects is not always so good. For example, on the test-dev of MSCOCO [8], the average precision (AP) of small objects of SSD is only 10.9%, and the average recall (AR) is only 16.5%. The major reason is that it uses shallow feature map to detect small objects, which doesn't contain rich high-level semantic information thus not discriminative enough for classification. However, the safety helmet is relatively small, Fig.2 illustrates the Pixel area scale of a safety helmet, which is much smaller than a person that wearing this safety helmet.

Fig. 2: Pixel area scale of a safety helmet

The main purpose of this paper is to develop an innovative, effective and practical safety helmet wearing detection system based on deep learning for people working

in the power substation, the whole system is mainly used SSD to extracting features, positioning coordinates and classifying objects. However considerable optimizations were made to improve the ability to detecting the safety helmet which is always small in the image.

We make the following contributions: 1) collecting and labeling thousands of pictures that include people are wearing or not wearing safety helmet, some are wearing other kinds of hats, mostly under the power station scenes, thus build the safety helmet image dataset with PASCAL VOC format; 2) find a solution to solve the problem that safety helmets are too small to detect and combine this solution with SSD to propose a novel safety helmet wearing detection system; 3) training SSD with our image dataset which makes the system can recognize whether the hat that people are wearing is safety helmet or not; 4)furtherly improve the performance of whole system by analysis the context information of adjacent frames.

We organize our paper as follows: The next section describes the safety helmet wearing detection principle based on SSD. Novel safety helmet wearing detecting system is presented in Section III. Section IV presents extensive experiments under power station scene. The paper closes with a conclusion in Section V.

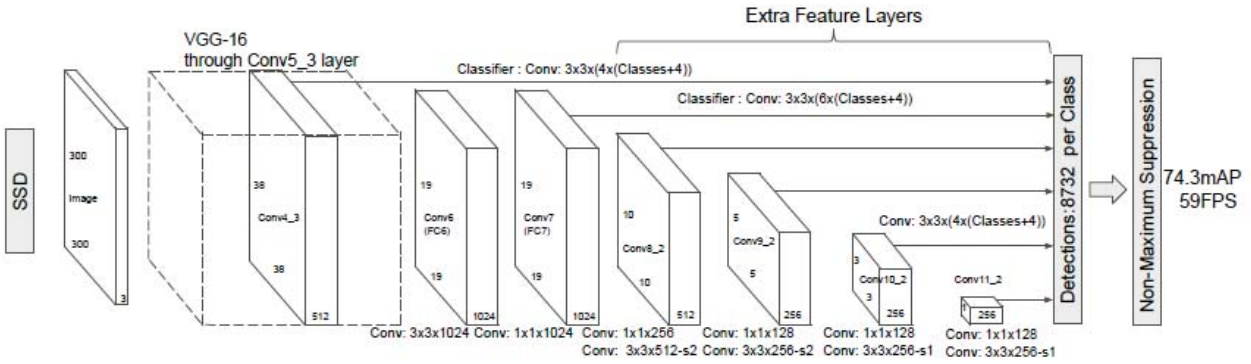

Fig. 3: Architecture of SSD with input-size of 300x300

II. SAFETY HELMET WEARING DETECTION BASED ON SSD

A. Architecture of SSD

SSD (Single Shot Multi-Box Detector) came out in 2015, boasting state of the art results at the time and real time speeds. As shown in fig.3, a backbone convolutional base (VGG16) is used. The SSD object detector comprises two main steps including feature maps extraction, and convolution filters application to detect objects. In addition, it uses anchors to define the number of default regions in an image, these anchors predict the class scores and the box coordinates offsets. For anchor matching, it begins by matching each ground truth box to the default box with the best jaccard overlap, then match default boxes to any ground truth with jaccard overlap higher than a threshold (0.5).

For each box, the SSD network computes two critical components including confidence loss which measures how confident the network is at the presence of an object in the computed bounding box using categorical cross-entropy and location loss which computes how far away the networks predicted bounding boxes are from the ground truth ones based on the training data [9]. The overall multitask loss function is defined as following:

$$L(x,c,l,g) = \frac{1}{N}(L_{conf}(x,c) + \alpha L_{loc}(x,l,g)) \quad (1)$$

Where N is the number of matched default boxes, the localization loss is similar to the Faster RCNN loss function- a smooth L1 loss between the predicted box (l) and the ground truth box (g) to predict the box offsets [5]. The confidence loss is the softmax loss over multiple classes confidences (c).

The major difference between the SSD from other architectures is that it was the first model to propose training on a feature pyramid. The network is trained on n number of feature maps, instead of just one. These feature maps, taken from each layer are similar to the FPN network but with one important difference. They do not use top down pathways to enrich the feature map with higher level information. A feature map is taken from each scale and a loss is computed and back propagated.

The SSD network computes the anchors for each scale in a unique way. The network uses a concept of aspect ratios and scales, each cell on the feature map generates 6 types of anchors, similar to the Faster RCNN [5]. These anchors vary in aspect ratio and the scale is captured by the multiple feature maps, in a similar fashion as the FPN. SSD uses this feature pyramid to achieve a high accuracy, while remaining the fastest detector on the market. Its variants are widely used in production systems today, where there is a need for fast low memory object detectors.

It is notable that the backbone-inside layer Conv4_3 is adopted for detecting objects of smallest size, the deeper layers are used to detect relative bigger objects. The range of the anchor size corresponding to each feature map is determined according to the object scale distributions on the training dataset. Although SSD can alleviate the problems arising from object scale variation, it still has limitation to detect small objects. The major reason is that it uses the feature of Conv4_3 to detect small objects, which is relatively shallow and does not contain rich high-level semantic information. Hence, we propose a safety helmet wearing detecting system to enhance the overall ability to detecting small safety helmet.

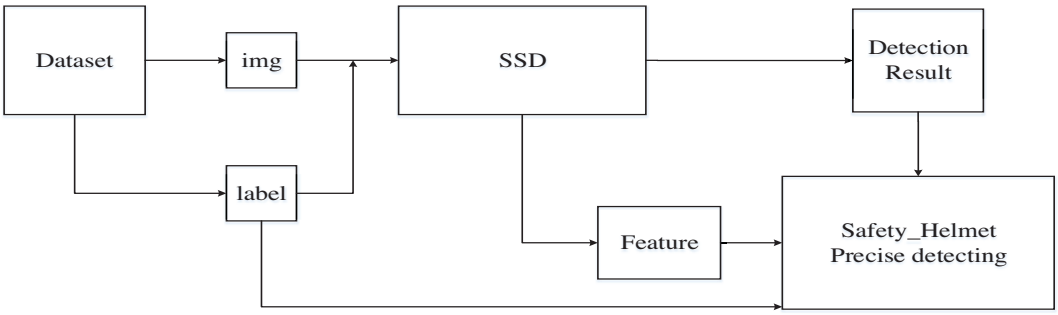

Fig. 4: Architecture of Safety Helmet Wearing Detecting System

B. Data Acquisition

We have prepared a relatively large dataset comprising more than five thousands image which were mainly captured from cameras placed in power station (include the station under construction) and some are from the internet or other dataset. The images contain indoor and outdoor environments under various conditions such as distance, lighting, angle, and camera type. Given the fact that each camera has its own color depth and temperature, field of view and resolution, all images passed through a preprocessing operation which ensures consistency across entire input data. Fig. 5 shows some examples of our dataset.

Fig. 5: Examples of our dataset

C. Labeling Strategy of SSD Training

With the small object issue of SSD, only labeling the safety helmet in the image dataset is absolutely not enough. Since person in our application scenario is much bigger than safety helmet, it is obviously much easier for SSD to detect person. Therefore, we use this characteristic in our dataset labeling strategy, person include whole head and body is labeled, no matter he is wearing safety helmet or not. Furthermore, person who wearing safety helmet is labeled with person_withhat, as for person not wearing safety helmet, we label it person_nohat. It is noticeable that for person who wearing hat but that hat isn't safety helmet, we also label it as person_nohat.

D. Safety Helmet Detection of SSD

In our scenes, Safety helmet detection is effective only when it's wearing by the worker, thus there is no need to identify safety helmet independently, the safety helmet detection is strongly connect with the person detection, SSD could compute the rough location and possibility that where may have safety helmet, the precise coordinates will provided by safety_helmet precise detecting module that will introduced in next section.

III. SAFETY HELMET WEARING DETECTING SYSTEM

A. Architecture of Our System

As illustrated in Fig.4, both the Detection Result and Feature that output from SSD are input to Safety Helmet Precise detecting module. This module is the key of whole system. The detection results of SSD in our system include the position coordinates of each person inside the image which in four direction of top-left, top-right, bottom-left, bottom-right. Besides, the extracted feature output from SSD will use the lower convolutional layer which has smaller receptive field, but include more detail information than Cov4_3.

During the training, the weights of SSD feature map are updated not only dependent on the detection result from SSD module but also safety helmet precise detecting module, hence, the label data is input to both this two modules.

Fig. 6: Safety helmet roughly location extracting process of the system

As shown in fig.6, In our system, the effective scale is zoom out to a solo person after SSD detector, moreover the pixel ratio of safety helmet to person's body is usually within a certain range. Therefore, roughly location that may has safety helmet can be predicted as a default box. To some degree, the default box is similar to the proposals after region proposal network in Faster-RCNN [5], but only focus on safety helmet, which with an aspect ratio and scale based on the safety helmet size in training dataset.

The safety helmet precise detecting module will mapping the safety helmet's rough possible location to the feature map to extract the feature. During the training, the feature will input into the classifier and regressor to compared with the ground truth of safety helmet, and update the weights.

Finally, during the detection, the classifier of safety helmet precise detecting module will identify the possibility of safety helmet and the precise location of safety helmet will computed through regression.

B. Context Information Analysing in Video Surveillance

With regard to actual video surveillance of power station, the content change of frames is not so intense, thus, we can combine the detecting results of several adjacent frames to improve the recognition accuracy of whole system.

For example, if a safety helmet with similar position coordinates has been detected in four of five real time video frames, then the confidence of safety helmet detecting results will be largely increased. On the other side, when the safety helmet detecting results is very randomly in continuous frames, the system will report warning, which means current judgement is not so reliable and we should look insight into the details.

Fig. 5: Visualization of some detection results

IV. Results And Discussions

The experiments are conducted on our own dataset. The final dataset consists of total number of 5,229 images, divided into training set (2,895 images), testing set (1,076 images) and validation set (1,258 images).

For fair comparison, we training and fine-tuning the SSD detector and our safety helmet wearing detecting system (which includes a SSD detector) separately with the same image dataset. The dataset used for testing is also the same to evaluate their performance.

It's worth noting that, for further analysis and research, the testing experiments are implemented in two ways, one is using testing set and other one is using only new field photos captured by power station camera. In both experiments, the evaluation metric is Average Precision (AP).

TABLE I. Testing experiment results

Method	Safety Helmet AP (Testing set, IoU=0.5)	Safety Helmet AP (field photos only, IoU=0.5)
Our system	78.3	68.5
SSD(only)	70.8	55.7

Tab.1 illustrate that the system we proposed has a higher AP no matter in testing set or power station field photos, compared with SSD detector, it shows our safety helmet wearing detecting system does have better performance.

On the other hand, the processing speed of two systems is testing under the same hardware environment (GTX 1080), the experimental results show that they could both run faster than 20 FPS, even though SSD detector could run faster, but our system still can realize real time detecting, Therefore, overall, the system we proposed is outperform SSD detector for safety helmet wearing detecting.

TABLE II. Comparison on processing speed

Method	Input Size	FPS
Our system	512*512	21.6
SSD(only)	512*512	22.3

IV. Conclusion

In this paper, for power station surveillance, we have proposed an effective safety helmet wearing detecting system, which based on SSD and a novel safety helmet precision detecting module. And we build an image dataset especially for safety helmet wearing detecting under power station scenario.

The experimental testing results on our own dataset reveal that our system has significantly outperformed the original SSD detector, especially for detecting safety helmet. In addition, the system can achieve real-time speed, i.e., 21fps. These advantages demonstrate that our safety helmet wearing detecting system is more suitable for the application of power station surveillance.

Acknowledgment

This work was supported by the State Grid Corporation of China under Grant No. SGRIXTJSKF [2017] 351.

References

[1] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pages 779–788, 2016.

[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed,C. Fu, and A. C. Berg. SSD: single shot multibox detector. In Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I, pages 21–37, 2016.

[3] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Doll´ar. Focal loss for dense object detection. In IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, pages 2999–3007, 2017.

[4] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li. Singleshot refinement neural network for object detection. CoRR, abs/1711.06897, 2017.

[5] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real time object detection with region proposal networks. In Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, pages 91–99, 2015.

[6] T. Lin, P. Doll´ar, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 936–944, 2017.

[7] K. He, G. Gkioxari, P. Doll´ar, and R. B. Girshick. Mask R-CNN. In IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, pages 2980–2988, 2017.

[8] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick. Microsoft COCO: common objects in

context. In Computer Vision - ECCV 2014 -13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V, pages 740–755, 2014.

[9] Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., et al. (2017). Speed/accuracy trade-offs for modern convolutional object detectors. In IEEE CVPR, volume 4.