Real-time pedestrian detection based on improved YOLO model

Congcong Zhao
School of computer science and technology
Harbin Institute of Technology
Weihai, China
e-mail: zhaocongcong.edu@163.com

Bin Chen
School of computer science and technology
Harbin Institute of Technology
Weihai, China
e-mail: binchen@hit.edu.cn

***Abstract*—To investigate accurate and real-time pedestrian detection results is a mainstream trend in the field of intelligent security. Avoiding too many external disturbances causing the errors and omissions, in order to obtain a reliable and discriminative detection. This paper proposes a deep learning method based on improved YOLO model to efficiently detect pedestrians. It addresses two necessary above issues: (1) leverage real-time saliency region detection through surveillance camera; and (2) extract more detail discriminative feature with human parsing. The results show that our deep real-time saliency and detail discriminative feature with human parsing based on improved YOLO model, successfully learn both spatial and temporal cues, making pedestrian detection further ensures the accuracy and timeliness in practical application scenes.**

Keywords-intelligent security; YOLO model; real-time saliency region; detail discriminative feature; human parsing

I. INTRODUCTION

Pedestrian detection technology has always been a popular research task in the field of computer vision. In recent years, it has an extremely wide range of application fields: intelligent monitoring, self-driving urban vehicles, pedestrian semantic analysis, and intelligent robot system. As its name suggests, pedestrian detection utilizes computer vision to determine whether pedestrians are appearing in an image or video sequence. At the same time, it can compute an accurate position to the appearing position. In addition, pedestrian detection is also the basis of person re-identification (Re-ID). The Re-ID aims to match the same identity in non-overlapping camera views by given a person of interest. Due to its promising applications in video surveillance and smart retail, which recently has drawn more and more attention. Similarly, the combination of pedestrian detection and Re-ID will make it more valuable in artificial intelligence applications.

Pedestrian detection technology has been developed to the date, and more excellent detection methods have been proposed and applied. It can be mainly divided into two categories: classic detection methods and deep learning methods under the supervision depending on big data. In the process of development, the most traditional pedestrian detection method is the HOG + SVM pedestrian detection，which is proposed by Dalal [1] et al and has performed near perfection on the MIT pedestrian data set. Felzenszwalb et al propose an improved DPM algorithm [2], it has a strong robustness to the deformation of the target due to a component-based detection method. Dollar et al. proposes integral channel features [3] and aggregate channel features [4] to receive better pedestrian characteristics by integrating gradient amplitude features, LUV and gradient histograms. This process belongs to shallow learning, and the ability to characterize features is limited that often failing to achieve the desired effect.

With the support of hardware devices, the deep neural network has also been widely used in pedestrian detection. The method of deep learning can learn different characteristics from large-scale data, and compared to the traditional manual extraction feature method, which is more representative. Deep convolutional network model can automatically learn high-level semantic features from the data through a hierarchical structure, and exhibits superior performance in the classification task, gradually replacing the characteristics of manual design. This depth feature is also applicable to pedestrian detection. At present, most mainstream pedestrian detection methods are based on depth features. Due to the strong discriminative ability of depth features, the performance of detection methods has also been greatly improved. Girshick et al. [5] first proposed an R-CNN network based on the proposed region, the combination of convolutional neural networks and suggested areas making a huge breakthrough in the field of target detection. Anelia [6] et al. proposes a high accuracy precision model to achieve pedestrian detection with cascaded classifiers. In recent years, there are some video saliency models worthy of attention, involving video retiming [7] and video segmentation [8]. However, real-time efficiency has become a common bottleneck for current video saliency algorithms towards industrial applications. Fortunately, there are some excellent target detection models have been successively proposed by academia and industry, since the convolutional neural network is first applied to target detection. Separately R-CNN, Fast R-CNN, Faster R-CNN, SPP-Net, YOLO, SSD, DSOD.

YOLO model is suitable for practical pedestrian detection applications both in timeliness and performance through experimental comparison. YOLO has been iterated to the current YOLOv3 through a status update of the version. However, it is aimed at highlighting saliency pedestrian detection and focusing on human parsing [9] detail domain discrimination, YOLOv3 can lead to problems with false detections and missed inspections, although it has improved in detection speed, multi-scale detection, and model generalization ability.

Therefore, this paper proposes a targeted model improvement on the above two aspects based on YOLOv3. In

978-1-7281-1859-8/19/$31.00 ©2019 IEEE
DOI 10.1109/IHMSC.2019.10101

terms of basic image feature extraction, YOLOv3 adopts a network structure called Darknet-53, which draws on the practice of residual network, and sets up a fast link between some layers. There are two contributions of this paper.

1) We adopt end-to-end training and salient prediction in dynamic real-time scenes, in order to enhance computational efficiency and detection capabilities, thereby reducing the probability of false positive.

2) We employ the emphasis on the human parsing to analyze the details and distinguish pedestrian in different ways, focusing on the upper and lower body semantic information guidance of various levels of pedestrians. The goal is to reduce the likelihood of omission detection from important human information markers.

II. Related Work

A. Multi-scale Feature for Object Detection

YOLOv3 [10] uses three different scale feature maps for object detection. Compared to the input image, the feature map used for detection has 32 times down-sampled. Due to the high down-sampled factor, the receptive field of the feature map is relatively large, so it is especially suitable for the detection image of relatively large object size. In order to achieve fine-grained detection, the feature map of the 79th layer starts to be up-sampled, and then merged with the 61st layer feature map, thus obtaining the 91st layer is thinner. The feature map of the granularity is also obtained through several convolutional layers to obtain a feature map which is down-sampled 16 times from the input image. Thus, it has a medium-scale receptive field and is suitable for detecting medium-scale objects.

B. 9 Scale a Priori Boxes

There three kinds of a priori boxes are set for each down-sampled scale, and a total of nine sizes of prior boxes are clustered. The nine a priori boxes in the COCO data set are: (10x13), (16x30), (33x23), (30x61), (62x45), (59x119), (116x90), (156x198), (373x326). Apply a larger a priori box (116x90), (156x198), (373x326) on the smallest 13*13 feature map with the largest receptive field on the distribution, which is fit to detect larger objects. In addition, the medium 26*26 feature map (medium receptive field) uses a medium a priori box (30x61), (62x45), (59x119), suitable for detecting medium-sized objects. It is suitable for detecting smaller objects to use the larger 52*52 feature map with a smaller a priori box (10x13), (16x30), (33x23). The design of a priori boxes for detecting the pedestrian size is presented in Figure 1 as follow. The model's border prediction formula is as follows：

$$b_x = \sigma(tx) + c_x \; ; \; b_y = \sigma(t_y) + c_y \tag{1}$$

$$b_w = p_w \mathrm{e}^{t_w} \; ; \; b_h = p_h \mathrm{e}^{t_h} \tag{2}$$

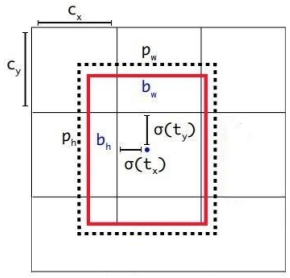


Figure 1. The design of a priori boxes for detecting the pedestrian size

Logistic is used instead of the output of softmax for prediction while predicting the object class. This way can support multi-label objects, such as a person with two labels: woman and person, separately. YOLO3 maps it to three scale output tensors for an input image, representing the probability of various objects in each position of the image. Regardless of the details of the neural network structure, taking into account the temporal and spatial characteristics of the scene, we improve the darknet model on the basis of YOLOv3.

III. Method

A. Pedestrian Real-time Saliency Regions Detection

On the one hand, prominent object detection is a key step in many image analysis tasks, as well as pedestrian detection of edge-level tasks, because it not only identifies relevant parts of the visual scene but also reduces computational complexity by filtering out uncorrelated segments of the scene. On the other hand, the visual impact of the human eye depends on the contrast between the target and the local area. The higher the contrast, the more perceptible human eye is to the target. It is a better way to use the contrast of each pixel and the surrounding area pixels to characterize the image. In this paper, we have improved on the YOLOv3 [11] model to reduce the false positive of pedestrian detection. In the meantime, we add a three-layer network at the end of darknet53, including a two-layer convolution and a fully connected network, in order to weaken the information of pedestrians under complex background conditions, to significantly enhance prospective pedestrian information.

B. Detail Discriminative Feature with Human Parsing

In the practice of using the deep network model, it is difficult to capture different features using only a single branch network. If the two types of pictures are placed in one network, the training effect will be much worse than that of the two types of pictures. Therefore, it is necessary to introduce different sub-networks to learn and obtain more results for the feature differentiation of the retail domain. The detail field is further divided into two sub-branches: the upper body branch and the lower body branch. These two types of branches can jointly learn complementary feature representations through the guidance of different semantic information. The semantic information for calculating the upper and lower branches of human parsing, $E_ı$ indicates the upper body branch, $M_ı$ indicates the lower body branch, and a, β are branching coefficients. The formula is as follow:

$$H = \alpha * E\iota + \beta * M\iota . \quad (3)$$

Because the YOLOv3 model itself has a strong generalization ability, we balance the detection capability with the generalization ability to find the most important local region feature and use Global Max Pooling instead of global uniform pooling. The overall training flow chart of the experiment is shown in Figure 2.

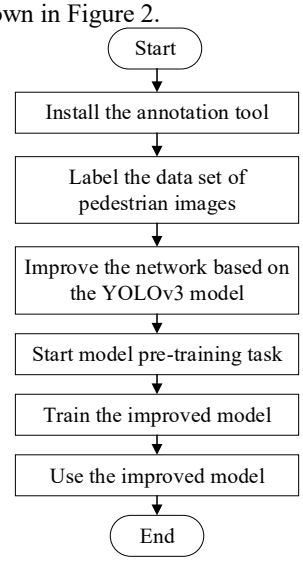


Figure 2. The overall training flow chart of the experiment

IV. EXPERIMENT

A. Dataset Details

At present, there are a lot of large-scale commonly used data sets, including PASCAL VOC, ImageNet, MS COCO, and other datasets. PASCAL VOC is a well-known data set in the fields of target detection, classification, and segmentation among them. The PASCAL VOC [12] contains approximately 10,000 images with bounding boxes for training and verification from 2005 to 2012, a total of 8 different challenges were held, and it is considered a benchmark data set for target detection problems. In addition, ImageNet released a target detection data set containing a bounding box. The training data set contains 500,000 images and belongs to 200 categories of objects. However, the amount of computation required for training is largely due to the data set is too large. Similarly, MS COCO dataset [13] was established by Microsoft Corporation, it is used in a variety of competitions: image header generation, target detection, key point detection, and object segmentation. Because it contains a wide range of computer vision research, this paper uses the MS COCO data set for pre-training operations for pedestrian detection. At the same time, we use the surveillance camera to shoot pedestrians from a more complex background for dataset labeling in the actual environment application[14]. The comparison of the common data set is shown in Table I.

TABLE I. THE COMPARISON OF THE COMMON DATA SET

Date set	*Category division*	*Content*
PASCAL VOC	20 kinds	Target detection, image classification
ImageNet	Full category	Target detection, segmentation, image semantics
MS COCO	80 goals	Target detection, segmentation, image semantics

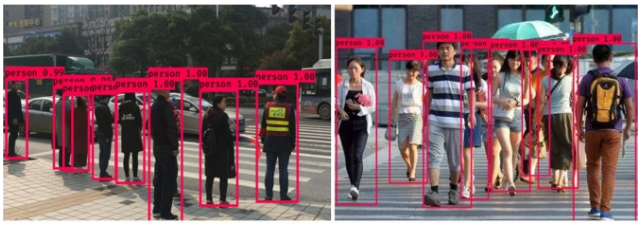
B. Model Training

We make an experiment with the YOLOv3 model based on the TensorFlow framework, the emergence of this framework has been widely recognized in academia and industry, especially for industrial application. We perform model improvement and related operations on the YOLOv3 model, such as edge determination, pedestrian overlap detection, etc., in order to achieve higher pedestrian detection accuracy in practical applications. In addition, the configuration information of our project environment including TensorFlow 1.5.1, Python 3.6.8, OpenCV 3.0.3. After pre-training weights and finetuning of the calibration data set, we increase the number of layers, modifying the corresponding configuration file based on the YOLOv3 model, therefore our pedestrian detection effect has been improved to a certain extent, especially in the real-time saliency detection and human parsing detail domain discrimination.

C. Results and Analysis

The detection effect of pedestrian real-time detection under complex background has been significantly improved, through the improvement of the YOLOv3 model in practical engineering applications. The result of pedestrian detection during the real-time saliency region and detail domain discrimination of human parsing has been visualized in Fig.3. From Fig.3(a), it is revealed that the effect of pedestrian detection is significantly enhanced in terms of false positive compared with the current state-of-the-art YOLOv3 model. The real-time saliency regional features improving the model with the real environment under the constraints of dark lighting and obstructions. As shown in Fig.2(b), detail domain discrimination of human parsing based on YOLOv3, this way can reduce effectively the omission ratio of pedestrian detection. In the meantime, analyzing pedestrian characteristics through human body parsing and significantly enhances the accuracy of pedestrian detection.

To strengthen the experimental reliability, we compare our model with the YOLOv3 in terms of false positive and omission ratio. The comparison results of the mode have been given in Tab.2. The results in Tab.2 show that the improvement of our model based on YOLOv3, whose false positive is reduced by 0.5%, and the omission ratio is reduced by 0.6%, improving the model pedestrian detection effect fully.



(a) (b)

Figure 3. The result of pedestrian detection during the real-time saliency region and detail domain discrimination of human parsing ((a) False positive (b) Omission ratio)

TABLE II. THE COMPARISON RESULTS OF THE MODEL

Model	***False positive (%)***	***Omission ratio (%)***
YOLOv3	4.2%	7.8%
OURS	3.7%	7.2%
ΔModel	+0.5%	+0.6%

V. CONCLUSION

In this paper, we make a progress in both errors and omissions of pedestrian detection, improving the real-time saliency area and the human parsing detail field respectively based on the deep learning pedestrian detection YOLOv3 model. It further enhances the accuracy and model performance in pedestrian detection, making the timeliness of its model processing better.

REFERENCES

[1] Wang Y, Zhu X, Wu B. Automatic detection of individual oil palm trees from UAV images using HOG features and an SVM classifier[J]. International Journal of Remote Sensing, 2018: 1-15.

[2] Miao Q, Liu R, Zhao P, et al. A semi-supervised image classification model based on improved ensemble projection algorithm[J]. IEEE Access, 2018, 6: 1372-1379.

[3] Kieritz H, Becker S, Hübner W, et al. Online multi-person tracking using integral channel features[C]//2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2016: 122-130.

[4] Dollár P, Appel R, Belongie S, et al. Fast feature pyramids for object detection[J]. IEEE transactions on pattern analysis and machine intelligence, 2014, 36(8): 1532-1545.

[5] Jiang H, Learned-Miller E. Face detection with the faster R-CNN[C]//2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE, 2017: 650-657.

[6] Lan W, Dang J, Wang Y, et al. Pedestrian Detection Based on YOLO Network Model[C]//2018 IEEE International Conference on Mechatronics and Automation (ICMA). IEEE, 2018: 1547-1551.

[7] Bradley A V, Klivington J, van der Merwe R, et al. Seamless Forward-Reverse Video Loops: U.S. Patent Application 15/678,497[P]. 2018-3-29.

[8] Caelles S, Montes A, Maninis K K, et al. The 2018 davis challenge on video object segmentation[J]. arXiv preprint arXiv:1803.00557, 2018.

[9] Liang X, Xu C, Shen X, et al. Human parsing with contextualized convolutional neural network[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1386-1394.

[10] Liu W, Cheng D, Yin P, et al. Small Manhole Cover Detection in Remote Sensing Imagery with Deep Convolutional Neural Networks[J]. ISPRS International Journal of Geo-Information, 2019, 8(1): 49.

[11] Redmon J, Farhadi A. Yolsov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[12] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge[J]. International journal of computer vision, 2015, 115(3): 211-252.

[13] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//European conference on computer vision. Springer, Cham, 2014: 740-755.

[14] Luo Z, Branchaud-Charron F, Lemaire C, et al. MIO-TCD: A new benchmark dataset for vehicle classification and localization[J]. IEEE Transactions on Image Processing, 2018, 27(10): 5129-5141.