Pedestrian Detection Based on YOLO-D Network

Zhong Hong
School of Computer Science
Beijing University ofPosts and Telecommunications
Beijing. China
hertz0725@126.com

Lei Zhang and Pengfei Wang
School ofComputer Science
Beijing University of Posts and Telecommunications
Beijing, China
{zlei & wangpengfei}@bupt. edu.cn

Abstract- In view of the 1100....obustness of pedestrian detector based on artificial extractlon feature, a novel pedestrian detection method is proposed by referring to the research results of state-of-the-art object detection. Based on the YOLOv2 network and lnsplred by the DenseNet, we pass the low-level **feature** maps of YOLOv2 to the higher **layers** in **turn.** By combining the feature maps of different convolutionallayers, we propose a novel network architecture, namely YOLO-D network, to make detector performance better, In o..der to solve the occlusion JIroblem, we introduced a Head-Shoulders model and also reduced computer calculations. The experimental results show that compared with the original YOLOv2 network pedestrian detection method, this paper reduces the missed rate and false rate, Improves the localization JI.ecision, and the detection speed meets the real-time requirements,

Keywords-Pedestrian Detection; YOLOv2; YOLO-D Network; Head-Shoulders Model; Object Detection

I. INTRODUCTION

As a significant branch of the object detection, pedestrian detection has become the focus of research by experts and scholars in the last several years. Pedestrian detection technology as the basis for research on pedestrian tracking and behavior analysis, has broad application prospects in areas such as assisted driving, intelligent monitoring, and intelligent robot technology.

Although pedestrian detection has made great strides in the past decade [19, 20, 21], there are still many problems to be solved. Pedestrians have both rigid and soft characteristics, scale, attitude, angle of view, illumination, and partial occlusion are all factors influencing the detect results. Although there are many research findings of pedestrian detection, many problems still need to be further studied, especially the real-time detection and the robustness to the partial occlusion environment. Pedestrian detection in the monitoring scene is one of the main applications in the field. In the surveillance scenario, pedestrians are dense and partially occluded, and detection has real-time requirements.

This makes it challenging to train pedestrian detectors for crowded scenes in daily life. First, the difference between pedestrians and background is small compared to general objects. In other words, discrimination relies more on the semantic context [5]. The second problem is how to accurately locate each pedestrian.

With the flourishing of convolutional neural networks [22], object detection has made great progress in the past few years. At present, the deep learning methods in the object detection field are mainly divided into two categories: the object detection algorithm of two stage and the object detection algorithm of one stage. The former is to first generate a series of proposals as samples, and then use the convolutional neural network to classify the samples; the latter does not generate proposals, and directly converts the problem of object localization into regression problem processing. Due to the difference between the two methods, the performance is also different, the former is superior in detection accuracy and localization accuracy, and the latter is superior in detection speed.

YOLOv2 [7] has the same good detection performance as other deep networks, but it also has certain disadvantages, such as the inability to effectively use the low-level feature map, which cannot effectively detect the occluded pedestrians. Inspired by DenseNet, we will improve the YOLOv2 network architecture for better pedestrian detection.

The main contribution of this paper are as follows:

(1) We improved the YOLOv2 network and presented our own YOLO-D network, which has good detection performance.

(2) In order to solve the occlusion problem, we propose a Head-Shoulder model, reduced missed detection rate.

(3) Our experimental results are in line with expectations, YOLO-D network performs better than YOLOv2 on the INRIA dataset

II. RELATED WORK

A. Object detection

The task of object detection is to find the target of interest in the image or video, and at the same time realize the position and category of the output detection target. It is one of the core issues in the field of machine vision. The academic world has nearly 20 years of research history. With the rapid development of deep learning technology, the object detection algorithm has also shifted from the traditional algorithm based on manual features [12] to the detection technology based on deep neural network.

978-1-5386-6565-7/18/$31.00 ©2018 IEEE

From the original R-CNN [1], OverFeat in 2013, to the Fast/Faster R-CNN [2, 4], SSD [13], YOLO series [6, 7], and Mask R-CNN [8], RefineDet, and RFBNet. In less than five years, deep learning based target detection technology, in the network structure, from two tasks to one task, from single scale network to multi-scale network, from personal computer to the cellphone, many good algorithm techniques have emerged. These algorithms have outstanding detection performance and performance on the open object detection dataset.

B. Pedestrian Detection

Presently, the pedestrian detection method can be divided into a method based on background modeling and a method based on statistical learning. The method based on statistical learning can be divided into traditional pedestrian detection method and pedestrian detection method based on neural network. The traditional method is mainly based on the artificial design feature extractor. By extracting features such as HOG [12], Haar, LBP, the classifier performs pedestrian detection and has achieved remarkable results. Among them, representative is the HOG (histogram of oriented gradient) feature proposed by Dalal in 2005. It combines linear support vector machine as a classifier and achieves good results. Most of the subsequent algorithms are extended on this basis. However, the artificially designed pedestrian features are difficult to adapt to the large changes of pedestrians, and the high computational complexity limits the practical application.

With the aim to overcome the shortcomings ofthe traditional method of manual design feature generalization, relevant scholars apply the depth model to pedestrian detection [23]. For the past few years, deep learning has made major breakthroughs in the field of object detection. The current CNN-based general object detection algorithms are mainly divided into two categories: a two-stage method represented by Faster R-CNN and a single-stage method represented by SSD, in which Faster R-CNN has higher accuracy, while SSD is superior in speed. So as to obtain higher accuracy, the current mainstream pedestrian detection algorithm is based on the detection framework of Faster R-CNN.

Although the accuracy rate has made great progress, the attention to speed demand is slightly insufficient. A straightforward approach is to use a single-stage approach to the detection framework, however the current single-stage approach has not shown an accuracy advantage over the main pedestrian dataset.

III. METHODOLOGY

The network of this paper is based on the YOLOv2 network, it adopts the idea of RPN (Regions Proposal Network), removes the fully connected layer, and uses the convolution layer to predict the offset and confidence of the bounding boxes. These offsets and confidences are predicted for each location in the feature map to obtain the probability and location of the pedestrian.

YOLOv2 predicts the relative offset value of the center point of the bounding box relative to the position of the upper left comer of the corresponding cell. In order to constrain the center point of the bounding box to the current cell, the sigmoid function is used to process the offset value, so that the predicted offset value is in the range (0, 1) (see the scale of each cell as 1). In summary, according to the 4 offsets, t_x, t_y, t_w and t_h, predicted by the bounding box, the actual position and size ofthe bounding box can be calculated as follows:

$$\begin{aligned} b_x &= \delta(t_x) + c_x, \\ by &= o(ty) + cy, \\ b_w &= p_w e^{t_w}, \\ b_h &= P_h e^{th}. \end{aligned} \tag{1}$$

Where Pw and Ph are the width and length of the a priori box, (c-, c_y) is the coordinate of the upper left comer of the cell. The scale of each cell is 1 when calculating, so the coordinates of the upper left comer of the current cell are (1, 1).

Multiply the class information of each cell prediction with the confidence information predicted by the bounding box to obtain the class-specific confidence score for each bounding box:

$$Pr(Class_{;}) * IOU^{truth}_{pred} = Pr(Class_{;} \mid Object) * Pr(Object) * IOU^{truth}_{pred}, \tag{2}$$

where

$$IOU^{truth}_{pred} = \frac{area(box(tTuth) n box(pred))}{area(box(tTuth) U box(pred))} \tag{3}$$

In the Equation (2), the first item on the right side of the equation is the category information for each grid prediction, and the second and third items are the confidence of each bounding box prediction. This product includes both the predicted box belonging to a certain class and the information of the box precision. The Equation (3) indicates that ratio of the intersection area of the predicted box to the ground truth and the area of the union.

A. Improve ojYOLOv2

YOLOv2 uses a new feature extractor called DarkNet-19, which includes 19 convolution layers and 5 max pooling layers. The design principle of Darknet-19 is consistent with the VGG16 model. It mainly uses 3*3 convolution. After adopting 2*2 max pooling layer, the feature graph dimension is reduced by 2 times, while the feature map's channels are doubled. Similar to NlN (Network in Network), DarkNet-19 eventually uses global average pooling for prediction, and uses 1*1 convolution to compress feature maps between 3*3 convolutions to reduce model calculations and parameters. DarkNet-19 also uses the batch norm layer behind each convolutional layer to speed up the convergence and reduce the model overfitting.

Although YOLOv2 has excellent detection performance, it also has certain disadvantages. For example, it cannot effectively utilize low-level features to effectively detect small pedestrians. Pedestrian characteristics show the structure of the deep learning network, high-level features are highly abstract, express the overall characteristics of pedestrians, and the characteristics of the middle layer are relatively specific, expressing the local characteristics of pedestrians.

As we all know, in the last few years, the direction of convolutional neural networks to improve the effect is either

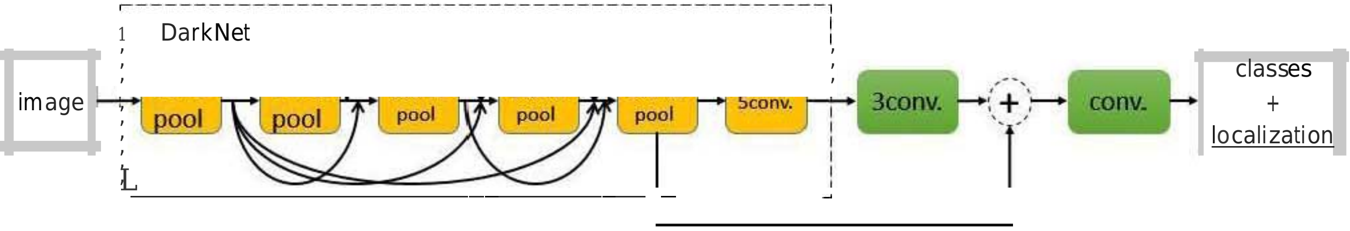


Figure 1 YOLO-D Network

deep (such as ResNet, which solves the problem of gradient disappearance when the network is deep) or wide (such as GoogleNet's Inception), while the author of the DenseNet is concerned about the feature, the ultimate use features achieve better results. DenseNet is a convolutional neural network with dense connections. In this network, there is a direct connection between any two layers. That is to say, the union of the outputs of all previous layers serves as the input of each layer of network, and the feature maps learnt by this layer are also passed directly to all the layers behind it for input.

It contains three dense convolutional blocks, each of which contains four convolutional layers. In each dense convolutional block, every convolutional layer can use the output of all previous convolutional layers as its input. The structure of each dense convolution block is shown in Figure 2.

Figure 2. Dense Convolutional Block

Inspired by DenseNet, we will combine the low-level feature maps and high-level feature maps of YOLOv2 to construct our YOLO-D network, as shown in Figure 1. In the YOLO-D network, we continuously transfer low-level feature maps to the high-level network. By combining different fine-grained features can increase the robustness of pedestrian detection, reduced the vanishing gradient, enhanced the feature transfer, more effectively utilized the feature, and to some extent reduced the number of parameters.

B. Loss Function

For the purpose of improving the detection of scale pedestrians, YOLO-D network pays attention to the application of low-level features. And to make our model generalization better, we redesigned the loss function. The optimized loss function is as follows:

$$LOSSYOLO\_D =$$

$$\sum_{1=0}^{W}\sum_{j=O}^{H}\sum_{k=0}^{A}\Big(I\ MaxIOU<Tllresll\ \lambda_{noobj} * \left(-b_{ijk}^{o}\right)^2$$

$$+ It<12800Aprior * \sum\nolimits_{rE(x,y,w,lI)} \left(prior_k^r - b_{ijk}^r\right)^2$$

$$+ 1_k^{truth}(\lambda_{coord} * \sum\nolimits_{rE(x,y,w,lI)} (truth^r - b_{ijk}^r)^2$$

$$+ \lambda_{obj} * \Big(IOU_{trutll} - b_{ijk}^{o}\Big)'$$

$$+ \lambda_{class} * (\textstyle\sum_{c=1}^{C} (truth^c - b_{ijk}^c)^2) \qquad (4)$$

As it shown in Equation (4), where W, H refer to the width and height of the feature map, A refers to the number of apriori boxes, and each λ value is the weight coefficient of each loss part.

The first loss is to calculate the confidence error of the background, but which prediction box to predict the background, we need to calculate the IOU value of each prediction box and all ground truth, and take the maximum value of IOU, if the value is less than a certain threshold, then this prediction box is marked as background and the confidence error of no object needs to be calculated.

The second is to calculate the coordinate error between the a priori box and the prediction width, but only between the first 12800 iterations. This is to be the shape of the prediction box that is quickly learned in the pre-training period.

The third largest term calculates the loss value of each part of the predicted box that matches a ground truth, including coordinate error, confidence error, and classification error.

C. *Head-Shoulders Model*

In the daily life monitoring scene, people are crowded, causing pedestrians to have a large amount of partial occlusion. In addition, the pedestrian's own occlusion, the mutual occlusion between the pedestrian and the object, make the video usually only shows the upper body of the person. Some pedestrians' trunks and legs are hindered by external environment, posture and clothes.

The shape of the head and shoulders of the human body is minimally changed and has better stability than other parts of the body. Therefore, in order to create a detector with high robustness to partial occlusion, we have introduced a head and shoulder model as shown in Figure 3.

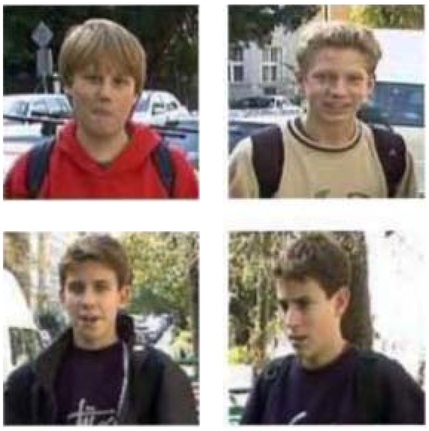


Figure J Head-Shoulders Model

In this paper, we detecting the head and shoulders of the pedestrians insteed of detecting their whole body, in this way, we reduced the amount of calculations and increased the speed of detection. At the same time, the recall rate is increased and the mss rate is reduced,

IV. EXPERIMENT

Our experiment is based on the open source framework DarkNet, where we train YOLOv2 and our own YOLO-D network. Then we will present our experimental results on dataset, training details and comparative experiments.

A. *Dataset*

The training samples in this paper are mainly from the INRlA. Ern pedestrian database and some pedestrian pictures in the PASCAL-VOC database. Combining multiple data sets will help our model training and make our model more robust. Finally. our training set has a total of I,500 images, containing about S,ooo positive samples of pedestrians. The INRIA test set serves as a test set for our experiments.

We manually label these imeges to mark the head-shoulders of each pedestrian. Trained the YOLOv2 network with the original data set and trained the YOLO-D network with our manually annotated data set.

B. *Evaluation index*

I) FPPT (False Positive per Image)

In this paper, we usc FPPI (False Positive pcr Image) to evaluate our detectors, it means that a given number N of sample sets, containing N images, with or without detection object in each image.

2) LAMR (Log-Average Miss Rate)

The pros and cons of the detector can be judged by the LAMR (Log-Average Miss Rate) indicator. LAMR reflects the overall missed detection of FPPI in the interval [10^{-2}, 10^{2}], the function as follows:

$$F_{LAMR} = e^{\frac{1}{9}(\sum_{i=1}^{9}\lg(miss-rate(10^{-2.25+0.25i})))} \qquad (3)$$

C. *Experiment resuhs*

I) FPPT-MR ill TNRTA Dataset

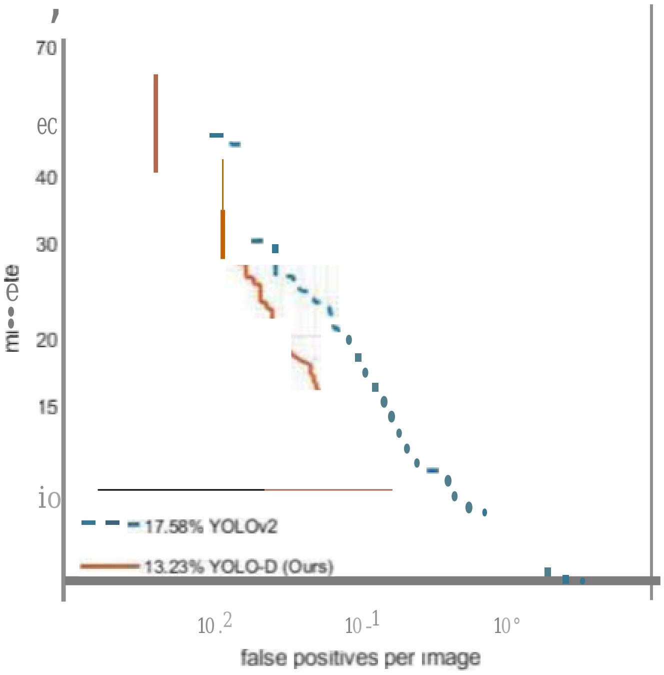


Figure 4 fPPI·MR in INRlA Dataset

As shewn in Figure 4. on the INRIA pedestrian dataset, the Log-Average Miss Rste of YOLOv2 is 17.58%.. while our YOLO-D network is approximately 4.3% lower, which is 13.23%.. This difference shows thal the YOLO-D network architecture is more reasonable and can better solve the crowd occlusion problem.

2) Mainstream algori/hm's LAMR

TABLEt MAINSTREAM ALGORITHM LAMR

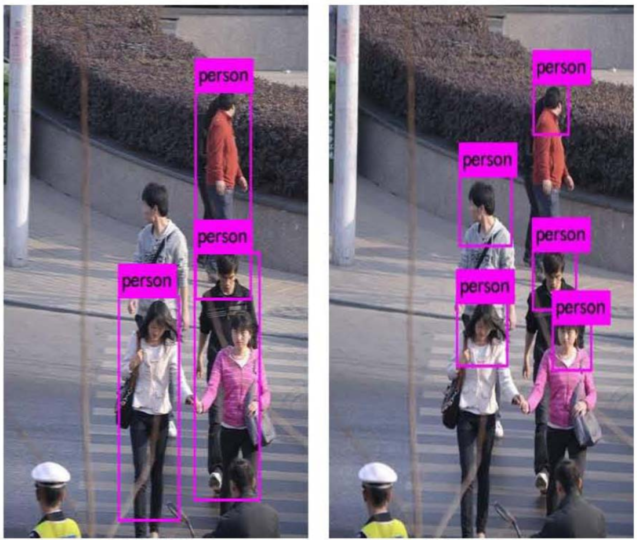
Algorithm	LAMR
V,	72.48%
Hoo	45.98%
HogLbp	39 10%
LatSvm-V2	19.96%
ConvNet	19 89%
YOLOv2	17.58%
YOLO-D (Ours)	13.23%

As shown in Table 1, in order to comprehensively evaluate the effectiveness of the proposed algorithm, we compare it with mainstream HOO, LatSVM, ConvNet and other pedestrian detection methods on the INRIA pedestrian test set, and we have achieved the best results.

Mainly because the traditional pedestrian detection method is based on manual design features and thus lack of robustness. and we designed a deep network that effectively learns pedestrian characteristics and trained a robust classifier.

Therefore, the YOLO-D network has advantages over other pedestrian detection algorithms, and the missed detection rate is relatively low.

3) Detect in real life scene



(a) YOLOv2 (b) YOLO-D

Figure 5. Detect in real life scene

As depicted in Figure 5, we tested the YOLOv2 and YOLO-D networks separately using pictures from real-life scenarios, the results showed that the YOLO-D miss detection rate was significantly lower than YOLOv2. YOLO-D detected 5 pedestrians, and YOLOv2 detected only 3 pedestrians.

V. CONCLUSION

In this paper, we present a novel network YOLO-D to improve the pedestrian detection accuracy in real life crowded scenes. The experimental results show that the proposed algoritlun has good generalization ability, and the accuracy of this method is significantly higher than that of the main popular pedestrian detection method under the public pedestrian dataset.

Our work is only a preliminary exploration of pedestrian detection from the network architecture level, although it has achieved good results, how to further reduce the pedestrian miss rate still has a long way to go.

REFERENCES

[I] Girshick R , Donahue J , Darrell T , et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[J]. 2013.

[2] Girshick R Fast R-CNN[ej// IEEE International Conference on Computer Vision. IEEE, 2015.

[3] Zhang L, Lin L, Liang X , et al. Is Faster R-eNN Doing Well for Pedestrian Detection?[ej// European Conference on Computer Vision. Springer, Cham, 2016.

[4] Ren S, He K, Girshick R , et al. Faster R-CNN: towards real-time object detection with region proposal networks[ej// International Conference on Neural Information Processing Systems. MIT Press, 2015.

[5] Mao J, Xiao T, Jiang Y , et al, What Can Help Pedestrian Detection?[ej// Computer Vision & Pattern Recognition. IEEE, 2017.

[6] Redmon J , Divvala S, Girshick R , et al. You Only Look Once: Unified, Real-Time 0 bject Detection[ej// Computer Vision & Pattern Recognition. IEEE, 2016.

[7] Redmon J , Farhadi A . YOL09000: Better, Faster, StrongerjCj/' IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2017.

[8] He K , Gkioxari G , Dollar, Piotr, et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, PP(99):I-I.

[9] Zhang S , Wen L , Bian X , et al. Occlusion-aware R-CNN: Detecting Pedestrians in a Crowd[J]. 2018.

[10] Wang X , Xiao T , Jiang Y , et al, Repulsion Loss: Detecting Pedestrians ina Crowd[J]. 2017.

[II] Huang G , Liu Z, Laurens V D M , et al. Densely Connected Convolutional Networks[ej// IEEE Conference on Computer Vision & Pattern Recognition. IEEE Computer Society, 2017.

[12] Dalal N , Triggs B . Histograms of oriented gradients for human detection[ej// IEEE Computer Society Conference on Computer Vision & Pattern Recognition. IEEE, 2005.

[13] Liu W , Anguelov D, Erhan D , et al. SSD: Single Shot MultiBox DetcctorjC]// European Conference on Computer Vision. Springer, Cham, 2016.

[14] Westlake N , Cai H, Hall P . Detecting People in Artwork with CNNs[CW European Conference on Computer Vision. Springer International Publishing, 2016.

[15] Tomè, Denis, Monti F , Baroffio L , et al. Deep Convolutional Neural Networks for pedestrian detection.[J]. Signal Processing Image Communication, 2016, 47(C):482-489.

[16] Liu J , Gao X, Bao N , et al, Deep convolutional neural networks for pedestrian detection with skip poolingl Cj/' International Joint Conference on Neural Networks. IEEE, 2017.

[17] Tian Y , Luo P , Wang X , et al. Pedestrian detection aided by deep learning semantic tasks[J]. 2015.

[18] Yang B , Yan J, Lei Z, et al. Convolutional Channel Features[Cl/IIEEE International Conference on Computer Vision. IEEE Computer Society, 2015.

[19] Dollar, P, Wojek C, Schiele B , et al. Pedestrian detection: A benchmark[J]. Proc.confon Computer Vision & Pattern Recognition, 2009:304-311.

[20] Benenson, Mathias, Timofle, et al. Pedestrian detection at 100 frames per second[ej// IEEE Conference on Computer Vision & Pattern Recognition. IEEE Computer Society, 2012.

[21] Geronimo, David, LOpez, Antonio M, Sappa A D, et al. Survey of pedestrian detection for advanced driver assistance systems[J]. IEEE Trans Pattern Anal Mach Intell, 2010, 32(7):1239-1258.

[22] Krizhevsky A, Sutskever I, Hinton G E. lmageNet classification with deep convolutional neural networks[ej// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105.

[23] Benenson R, Omran M, Hosang J, etal. Ten Years of Pedestrian Detection, What Have We Learned?[C]// European Conference on Computer Vision. Springer, Cham, 2014:613-627.