2018 14th International Conference on Computational Intelligence and Security (CIS)

A Method of Detect Traffic Police in Complex Scenes

Ying Zheng
Beijing Key Laboratory of Information Service Engineering
Beijing Union University
Beijing, China
buu_zhengying@sina.com

Hong Bao*
Beijing Key Laboratory of Information Service Engineering
Beijing Union University
Beijing, China
xxtbaohong@buu.edu.cn

Xinkai Xu
Demonstration Center of Experimental Teaching in Comprehensive Engineering
Beijing Union University
Beijing, China
ldtxinkai@163.com

Nan Ma
College of Robotics
Beijing Union University
Beijing, China
xxtmanan@buu.edu.cn

JiaLei Zhao
College of Information Engineering
Hubei University for Nationalities
EnShi, China
745850441@qq.com

Dawei Luo
Beijing Key Laboratory of Information Service Engineering
Beijing Union University
Beijing, China
1361722864@qq.com

***Abstract*—Target detection has a wide range of applications in many areas of life, and it is also a research hotspot in the field of unmanned driving. Urban roads are complex and changeable, especially at intersections, which have always been a difficult and key part in the research of pilotless technology. Traffic policemen detection at intersections is a key link, but there are few existing algorithms, and the detection speed is generally slow. Aiming at this problem, this paper proposes a real-time detection method of traffic police based on YOLOv3 network.The YOLO network is robust and capable of quickly completing target detection tasks. According to the information investigated, there are currently few data sets on traffic police detection. In response to this problem, this paper adopts the transfer learning method, adopts the imageNet set to training model, learns the basic characteristics of people, and then selects 1000 pictures containing traffic police to conduct experiments. The average accuracy of traffic police detection is 77%, and the detection speed reaches 45FPS, which basically meets the requirements of real-time performance, indicating that the method is reasonable and feasible.**

***Keywords*-YOLO network; traffic police detection; transfer learning; machine vision**

I. Overview

In recent years, autonomous driving technology has received widespread attention. Self-Driving-Car and advanced driver assistance systems have developed rapidly [1]–[3]. As early as the end of the 1970s, developed countries such as the United States, Britain, and Germany began to study Self-Driving-Car technology, which promoted the Self-Driving-Car technology into a large-scale research phase and achieved breakthroughs results in feasibility and practicality. China's research on driverless technology is relatively late, and its scale is relatively small. Therefore, there is still a certain gap with foreign countries in terms of technology, but it has also achieved considerable results. The driverless cars mainly uses the sensing system to perceive the environment around the cars. The speed and direction of the driverless cars are changed according to the perceived road information, vehicle position, obstacle information, pedestrians, etc., so that the vehicle can travel safely and reliably on the road. Autonomous driving technology is developing rapidly, but it also faces many challenges. Target target detection is one of the key components of its technology, which has gradually attracted the attention of researchers and become one of the research hotspots of driverless technology.

Target detection is a kind of computer vision task that distinguishes the target in the image or video from other parts that are not interested, which determines whether there is a target, determines the target position, and identifies the target type. In the driverless technology, the environmental information around the car is obtained through vision, and the intersection scene is one of the most difficult problems faced by the driver. The scenes of urban road junctions are complex and changeable. There are many uncontrollable factors, especially during the morning and evening peak hours. Traffic is very congested. Traffic lights and traffic signs are difficult to separate the roads. Traffic police intervention is often required, and command pedestrians and vehicles with gestures. n this kind of scenario, if the driverless car is difficult to identify and understand the complicated and varied gestures of the traffic police, there is a great potential safety hazard. To identify and understand the traffic

978-1-7281-0169-9/18/$31.00 ©2018 IEEE
DOI 10.1109/CIS2018.2018.00026

police's gestures through computer vision, we must first detect the traffic police at the intersection. Therefore, it is very important to detect traffic police at intersections.

II. Related Work

Target detection. Nowadays, the target detection based on deep learning [4] has become the mainstream, and there are representative of R-CNN [5], Fast R-CNN [6], Faster R-CNN [7], SSD [8]etc.. The method of target detection based on region suggestion such as Faster R-CNN includes various candidate region generation parts and different feature layer processing procedures, so that the real-time performance of this algorithm is not guaranteed. Therefore, a target detection algorithm based directly on regression is generated without generating a target candidate region. The YOLO (You Only Look Once) [9] algorithm simply divides the picture into multiple parts, and then directly determines whether there is a target in each part through the deep neural network, and predicts the target category and the bounding box of the target. This method does not need to generate the target suggestion area, which saves the image processing time, so that the real-time detection is reliably guaranteed, and the fastest processing of 155 frames per second is possible. Although the YOLO network guarantees real-time performance, the accuracy needs to be improved. YOLO v3 improves the accuracy while ensuring the detection speed. Therefore, this article uses the YOLO v3 network to detect traffic police at intersections.

Data set. There is a common problem in the target detection methods based on deep learningdata sets. There are very few data sets about traffic police, and classification learning is very difficult, and it is difficult to train a reliable classification model. Traffic police and pedestrians have a common feature - people, so this paper uses the transfer learning method, through the imageNet data set as an auxiliary data set, learn the characteristics of people, and train an effective classifier. Then, through a small amount of traffic police data to learn the characteristics of the traffic police, to detect traffic police in the crowd.

YOLO network. YOLO treats target detection as a regression problem, and uses regression to detect the location of the target and identify the target category. The whole model framework of YOLO is very simple and fast. Under the Titan X GPU, it can achieve 45 frames per second, and the fast version can achieve 150 frames per second. It is able to directly detect and identify all targets of the entire map, which is equivalent to adding the surrounding environment information of the target. YOLO gradually entered the researcher's field of vision because it can guarantee the real-time detection. Based on the YOLO network structure, GAO Zong [10] and others combined the characteristics of pedestrians with small aspect ratio on the image, clustered the appropriate number of candidates and specifications, improved the YOLO network structure, and detected pedestrians in real time. Wang Lin [11] and others proposed to combine YOLO with the Pyramid Pooling Module in the Pyramid Scene Analysis Network (PSPnet) [12] to solve the problem that it is difficult to recognise pedestrians due to large dust and dark light. Jing [13] and others used the YOLO network to recognise human motion in video. In addition, in the field of Self-Driving-Car driving, YOLO network is also used to realize vehicle detection [14] , traffic sign detection [15] , license plate detection [16] and so on.

III. Problem statement

The complexity of urban roads, especially at intersections, has always been one of the difficulties in the study of driverless technology. The "communication" between the vehicle and the traffic police is based on the traffic police's gesture, and the traffic police is in the flow of people. To identify the traffic police's gesture, the traffic police must first be detected. Therefore, this paper proposes a traffic police detection method based on YOLOv3 network to lock traffic police in the crowd.

A. YOLO network structure introduction

The YOLO algorithm can be roughly divided into the following three steps. (1) Divide the entire picture into $S \times S$ grids. (2) Send the whole picture to the deep neural network, predict whether each grid has a target, a target's bounding box, and a target class. (3) Use the non-maximum suppression (NMS) of the predicted bounding box to filter out the best bounding box to get the best results. Here, the process diagram is shown in Fig. 1 using the schematic diagram in [9].

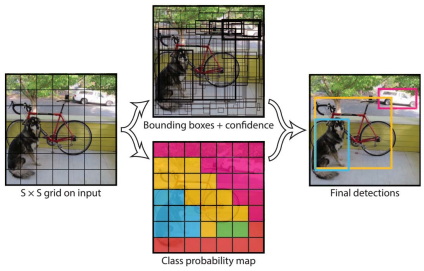


Figure 1: overall frame diagram

The YOLOv1 network consists of 24 convolutional layers and 2 full-link layers. Among them, the convolution layer is used to extract features, and the fully connected layer is used to predict the position and category of images. YOLOv1 achieves real-time detection targets, but its detection accuracy is lower than other state-of-the-art target detection algorithms. In order to improve the positioning accuracy and recall rate of objects, YOLOv1 network introduced the idea of anchor box in Faster R-CNN, and improved the design of network structure. The output uses the convolutional

layer instead of YOLOv1 to connect the fully connected layer, and joint use of coco object detection annotation data and imageNet object classification annotation data training object detection model, called YOLO9000 [**?**]. YOLO9000 has greatly improved in terms of identification, accuracy, speed and positioning accuracy. This year, the author of YOLO made some changes to the YOLO network. The model of YOLOv3 [**?**] is more complicated than the previous two versions. You can balance the speed and accuracy by changing the size of the model structure. YOLOv3 consists of continuous 3×3 and 1×1 convolutional layers, a total of 53 layers. YOLOv3 processed 608×608 images on Titan X to 20FPS, which is 3 times faster than SSD. In 51ms, AP_50 is 57.9 %. Compared with RetinaNet, the performance is similar, but it is 3.8 times faster.

B. the method of Traffic police detection

First, the input image containing the traffic police is divided into $S \times S$ grids, and each grid is responsible for detecting objects falling into the grid. If the coordinates of the center position of an object fall into a certain grid, then the grid is responsible for detecting the object. This way of dividing the image is very simple and can ensure that there may be targets in the grid. This is easier and faster than judging thousands of areas based on the regional suggestion method, which directly improves the efficiency of the whole process. Then, the feature is extracted through the convolutional layer, and the position coordinates of the traffic police are predicted. YOLOv3 only uses the convolutional layer, skips the connection layer and the upsampling layer, does not use any form of pooling, and uses the convolutional layer of stride 2 to downsample the feature map, which helps to prevent the pool usually Loss of low-level features caused by pooling. Finally, the non-maximum suppression is used to detect the traffic police in the figure.

YOLOv3 predicts the location of the traffic police, and the category is expressed by the tensor of $S \times S \times (B \times 5 + C)$. Where B represents the number of target bounding boxes to be predicted in the image. If B=2, it means that two bounding boxes are to be predicted at a time. 5 represents the target bounding box (x, y, w, h) and confidence score of this grid on the target is Conf, total of 5 parameters. C indicates the number of categories of targets in the data set used. YOLOv3 predicts 3 bounding boxes in each cell. The target confidence score Conf is used to indicate whether the bounding box of the grid prediction in the graph contains the target and whether the predicted bounding box is accurate. The formula is as follows:

$$\begin{aligned} Conf &= P(Class_i Object) \cdot P(Object) Iou_{pred}^{truth} \\ &= P(Class_i) Iou_{pred}^{truth} \end{aligned} \tag{1}$$

Where $P(Class_i Object)$ represents the probability that the target belongs to the i-th class if there is a target. $P(Object)$ indicates whether there is a target in the bounding box. A represents the ratio of the intersection and the union between the predicted target box and the real target box. This parameter will evaluate whether the selected border is valid during the test phase. The calculation formula is as follows,

$$IOU = \frac{area(B_t \cap B_p)}{area(B_t \cup B_p)} \tag{2}$$

Where B_t is the reference standard box for the training tag, B_p is the bounding box of the detection, and area() is the area. The predicted bounding box is obtained through the Anchor box. Suppose the predicted four coordinates are: t_x, t_y, t_w, t_h, the margin of the upper left corner of the target unit is c_x, c_y, and the width and height of the corresponding bounding box are $p_w p_h$, then the predicted value of the network is as follows :

$$b_x = \sigma(t_x) + c_x \tag{3}$$

$$b_x = \sigma(t_y) + c_y \tag{4}$$

$$b_w = p_w e^{t_w} \tag{5}$$

$$b_h = p_h e^{t_h} \tag{6}$$

The specific conversion process is shown in Figure 2.

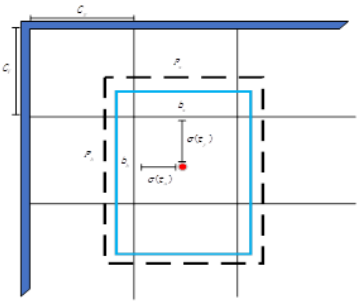


Figure 2: YOLOv3 network structure

The YOLOv3 network detects the traffic police and extracts the features of the entire picture, taking into account all the information around the traffic police. To optimize the

entire model, the loss function is as follows:

$$
\begin{aligned}
loss = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} (1_{ij}^{obj} \cdot [(x_i - \hat{x})^2 + (y_i - \hat{i})^2]) \\
+ \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} (1_{ij}^{obj} \cdot [(\sqrt{w_i} - \sqrt{\hat{w}_i})^2 \\
+ (\sqrt{h_i} - \sqrt{(h_i)})^2]) + \sum_{i=0}^{S^2} \sum_{j=0}^{B} (1_{ij}^{obj} \cdot (C_i + \hat{C}_i)^2) \\
+ \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} (1_{ij}^{noobj} \cdot (C_i + \hat{C}_i)^2) \\
+ \sum_{i=0}^{S^2} 1_i^{obj}) \cdot \sum_{c \in class} (p_i(c) - \hat{p}_i(c))^2
\end{aligned}
\tag{7}
$$

The parameter $_{coord}$ is used to enhance the importance of

Figure 3: Partial data setThe first line is the imageNet data set, and the second line is contains the traffic police data set.

the bounding box in the loss calculation, $_{coord} = 5$. The parameter $_{noobj}$ is used to reduce the influence of the non-target area on the confidence calculation of the target area, $_{noobj} = 0.5$. 1_i^{obj}indicates that the i-th image block in the image has a target. 1_{ij}^{obj} indicates that the jth prediction frame of the i-th image block has a target, otherwise 1_{ij}^{noobj}.

IV. EXPERIMENT

A. data set

Figure 4: The corresponding document of the traffic police image

Today, there are many public data sets in various fields, but there is no data set about traffic police. In the case of insufficient data sets, this paper uses imageNet data sets as auxiliary training data to train the model. The ImageNet dataset has more than 14 million images covering more than 20,000 categories; more than one million of them have clear category annotations and annotations of object locations in the image, taking 10,000 image training models. In this paper, through the crawler, 1000 images containing traffic police keywords are captured in the network as source data, which is manually labeled. As shown in Figure 3, the label of the traffic police picture is shown in Figure 4.

Figure 5: YOLOv3 detection results in the is _ policeman _ dataset data set

B. Experimental results and analysis

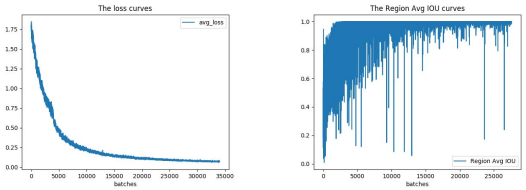
The entire training process uses the Stochastic gradient descent (SGD) and backpropagation algorithms to learn the network parameters. The batch size of the training is 8, the number of iterations is 20000, and the learning rate is 10^{-3}. The experiment is based on Ubantu 16.04, 64 operating system, and the GPU is GTX1080.

Table I: YOLOv3 compare

	mAP	person	police	fps
SSD-416	0.730153	0.589553	0.870754	46
YOLOv3-416	0.770559	0.617363	0.923755	50

In table 1, YOLOv3 performs well in efficiency and accuracy,means it can meet the needs of real-time detection of traffic police in the crowd.

It can be seen from the experimental results that the detection speed of each frame of picture is 0.026 seconds, that is 45FPS, real-time requirements are met.



(a) avg loss (b) Region Avg IOU

Figure 6: loss and IOU

Figure 7: Traffic police test results in different scenarios

Figure 6(a) shows the model converge in 20000 iteration. As shown in Figure 6(b), the IOU value is occasionally low after the convergence of the model, which means that the detection is inaccurate sometimes. We analyze that there are occlusion and various scales of images in the data set, which is a great challenge to the model. It is shown that in the case of insufficient source data set, adding a large amount of auxiliary data for training can obtain a relatively reliable and effective classifier, that is, the training method used in this paper is effective and feasible.

The test results of the traffic police in different scenarios are shown in Figure 8.

V. Conclusion

For the target detection, the problem of extracting features takes a long time. This paper proposes a traffic police detection model based on YOLOv3 combined with practical problems. Through experiments, the average accuracy of the model is 77%. Although the accuracy needs to be improved, the detection speed reaches 45fps, which meets the requirements of real-time. The next step will be to improve the misdetection and missed detection, data set and labeling of traffic police detection to improve the accuracy of traffic police detection and optimize the research tasks of traffic police detection.

Acknowledgment

We really thank anonymous reviewers constructive suggestions. This part of study is partially founded by the national natural science foundation of China with the numbers 61871038 and 61672178Natural Science Foundation of Beijing with the numbers 4182022.

References

[1] Dai Yifan, He Chengkun. How to make a self-driving car safely on the road [N]. China Automotive News2016.8.15, 6

[2] LI Keqiang,DAI Yifan,LI Shengbo,BIAN Mingyuan. State-of-the-art and technical trends of intelligent and connected vehicles [J]. Journal of Automotive Safety and Energy.2017811-14

[3] Li Zhi, Han Guang, Zhang Xiuli. The future has come - the outbreak of driverless technology [J]. Modern Vocational Education, 2018(2)240

[4] Zhou Xiaoyan,Wang ke,Li Lingyan.Review of object detection based on deep learning[J]. Electronic Measurement Technology, 2017(11):89-93.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.

[6] Girshick R. Fast R-CNN[J]. Computer Science, 2015.

[7] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Trans Pattern Anal Mach Intell, 2017, 39(6):1137-1149.

[8] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[J]. 2015:21-37.

[9] Redmon, Joseph, Divvala, Santosh, Girshick, Ross, et al. You Only Look Once: Unified, Real-Time Object Detection[J]. 2015:779-788.

[10] GAO Zong, LI Shaobo, CHEN Jinan,et al. Pedestrian Detection Method Based on YOLO Network[J]. Computer Engineering,2018,44(5):215-219,226.

[11] WANG Lin, WEI Chen, LI Weishan, et al. Pedestrian detection based on YOLOv2 with pyramid pooling module in underground coal mine. Computer Engineering and Applications [J]. Computer Engineering and Applications,2018

[12] Zhao H, Shi J, Qi X, et al. Pyramid Scene Parsing Network[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2017:6230-6239.

[13] Jing L, Yang X, Tian Y. Video You Only Look Once: Overall Temporal Convolutions for Action Recognition[J]. Journal of Visual Communication & Image Representation, 2018.

[14] Wei Yang ,Ji Zhang ,Hongyuan Wang,Zhongbao Zhang, A vehicle real-time detection algorithm base on YOLOV2 framework,Proc. SPIE 10670,Real-Time Image and Video Processing 2018,106700N(21May 2018);doi:10.1117/12.2309844

[15] Zhang J, Huang M, Jin X, et al. A Real-Time Chinese Traffic Sign Detection Algorithm Based on Modified YOLOv2[J]. Algorithms, 2017, 10(4):127.

[16] Laroca R, Severo E, Zanlorensi L A, et al. A Robust Real-Time Automatic License Plate Recognition based on the YOLO Detector[C]// International Joint Conference on Neural Networks. 2018.

[17] Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2017:6517-6525.

[18] Redmon J, Farhadi A. YOLOv3: An Incremental Improvement[J]. 2018.