Research on Assisted Driving Technology based on Improved YOLOv3

Long Zhao[1,2,*], Xiaoye Liu [3], Qiang Wang[1] and Honglong Chen[1]

[1]Big Data Institute, Heilongjiang Oriental University, Harbin, Heilongjiang 150066, China
[2]College of Information and Computer Engineering, Northeast Forestry University, Harbin, Heilongjiang, 150040, China
[3] Harbin Far East Institute of Technology, Harbin, Heilongjiang, 150016, China
*Correspondence should be addressed to Long Zhao; zl_oriental@163.com

Abstract. In the era of automobile popularization, China and the developed countries such as Europe and the United States are facing the same problem of high car accidents. This paper uses deep learning technology to detect and identify lane lines, traffic lights, vehicles and pedestrians during driver driving. Improve the efficiency and accuracy of convolutional neural network training through migration learning and data enhancement techniques. Based on the current advanced YOLOv3 network, we have improved the network structure and loss function. The KITTI dataset has achieved the highest 2D target recognition accuracy, and the target recognition speed is higher than 36 frames/sec. The price of the assisted driving equipment we developed does not exceed RMB 5,000, which has a good market prospect.

1. Introduction

Traffic accidents caused by dangerous behaviors such as fatigue driving, distracted driving, drinking and driving, and speeding have become "the world's number one" [1]. According to data released by the World Health Organization, 1.2 million people worldwide die in car accidents every year, and one person is killed in traffic accidents every 25 seconds. In China, with the increase in car ownership, the number of people killed in car accidents has also risen sharply. At present, China has become one of the countries with the highest death rate from road traffic accidents.

In the existing domestic and international automatic driving research, target recognition is one of the core tasks, including road and road edge identification, lane line detection, vehicle identification, vehicle type identification, non-motor vehicle identification, pedestrian identification, traffic sign recognition. , obstacle identification and avoidance, etc. [2-3]. The target recognition system uses computer vision to observe the traffic environment, automatically recognizes the target from the real-time video signal, and provides a basis for real-time automatic driving, such as start, stop, steering, acceleration, and deceleration. Due to the extremely complicated actual road conditions, the performance of assisted driving technology based on traditional target detection is difficult to be greatly improved. The existing automatic driving technology generally relies on advanced radar systems to compensate, which significantly increases the cost of system implementation.

The video sequence captured by our proposed method in the road traffic scene contains various video targets, such as pedestrians, cars, roads, obstacles, various objects in the background, etc., and even identifies in the test image. The target object of the category of interest. At the same time, the program can also feed the test results to the driver in the form of voice announcements, provide the basis for decision-making of the vehicle control system, and remind the driver to drive safely. Since

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

we have improved the network structure and loss function of YOLOv3, our method achieves the best performance in the KITTI dataset.

2. Related Work

In recent years, due to the rise of deep neural networks, some significant advances have been made in the research of vehicle detection and tracking algorithms at home and abroad, mainly in vehicle detection algorithms based on deep convolutional neural networks [4-7] and correlation filtering. Vehicle tracking algorithm [8-10]. These target detection studies are distinguished from target frame selection methods and can be divided into 2D target detection and 3D target detection. From the application medium, they can be divided into LIDAR data for deep learning and image data for deep learning. Most of the initial research relied on expensive lidar systems such as Velodyne and hand-labeled environmental maps. In contrast, recent efforts have attempted to replace laser radar with cheap car cameras that are readily available on most modern cars. Although some algorithms have a significant improvement in accuracy, this comes at the cost of increased system complexity and computational cost. This is difficult to apply in the field of assisted driving with extremely high real-time requirements.

3. Approach

3.1. *Enhance Image Datasets Using "Polar Coordinate Transformation"*

The so-called polar coordinate transformation is that the pixel is represented by the original (x, y) by polar coordinate transformation (r, θ), and then represented as a two-dimensional image. Involved in mathematical Equations 1 and 2.

$$\theta_u = 2\pi \cdot u/U \qquad \forall \mathrm{ue}\{0,\cdots,U-1\} \tag{1}$$

$$\mathrm{x} = \|v \cdot \cos(\theta_u)\| \; and \; y = \|y = \sin(\theta_u)\| \tag{2}$$

Enhancing the image dataset can significantly improve the generalization ability of the model and reduce the possibility of overfitting. Use the diagram to indicate as shown in Figure 1.

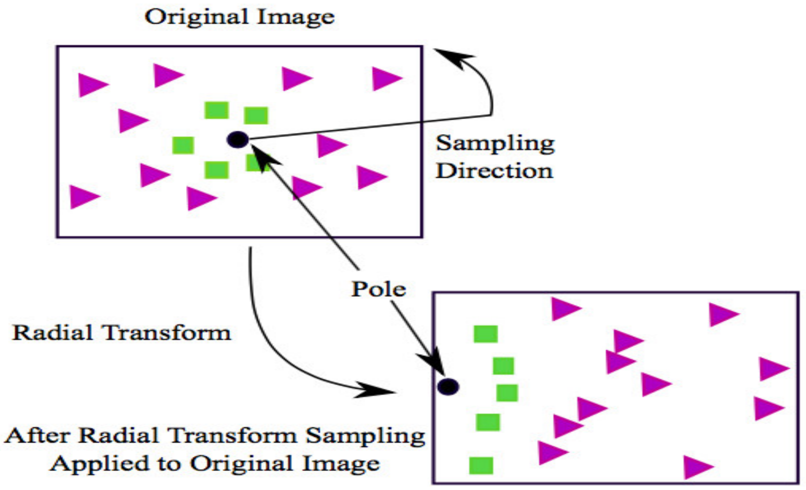


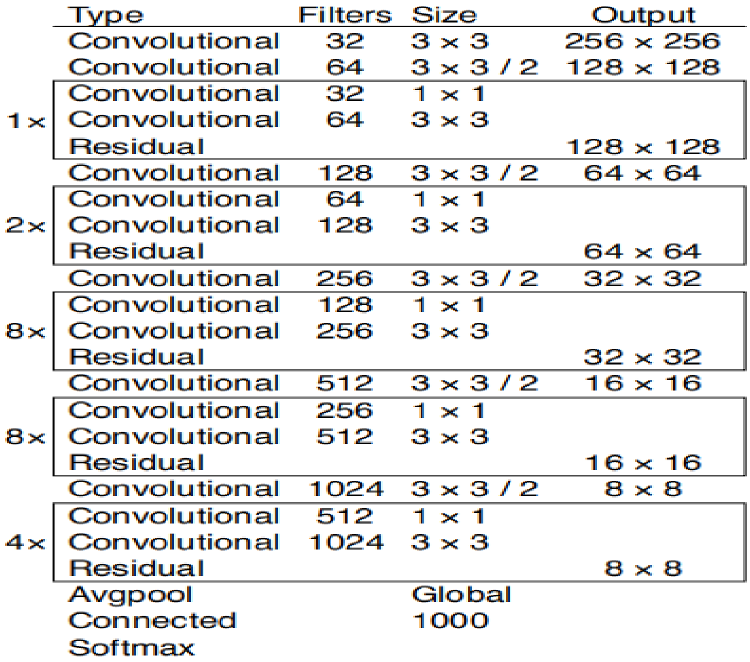
Figure 1. Polar coordinate transformation diagram

3.2. *Using Migration Learning to Speed up the Training of the Model*

In traditional classification learning, in order to ensure the accuracy and high reliability of the classification model obtained by training, there are two basic assumptions: (1) The training samples used for learning and the new test samples satisfy the independent and identical distribution. (2) There must be enough training samples available to learn a good classification model. However, in practical applications we find that these two conditions are often not met. First, over time, the previously available tagged sample data may become unavailable, with semantics and distribution gaps in the distribution of new test samples. In addition, tagged sample data is often scarce and difficult to obtain. This raises another important issue in machine learning. How to use a small number of tagged training samples or source domain data to build a reliable model to predict target areas with different data distributions. In recent years, migration learning has caused widespread concern and research. Migration learning is a new machine learning method that uses existing knowledge to solve different but related domain problems. It relaxes two basic assumptions in traditional machine learning, with the goal of migrating existing knowledge to solve learning problems in the target domain with only a small number of tagged sample data or even no. Migration learning is widely existed in human activities. The more factors shared by two different fields, the easier it is to migrate learning. Otherwise, the more difficult it is, even the "negative migration", which has side effects.

3.3. Target Recognition Network Design and Loss Function Design

YOLOv3 is the representative of the advanced one-stage target detection model [11]. YOLOv3 uses Darknet-53 as its backbone network. The existing CNN model learns the characteristics of objects by stacking multiple convolution and pooling layers, but the YOLOv3 network is a full-convolution network that uses a lot of residual hopping connections. The advantages of using residual structure: (1) A key point of the depth model is whether it can converge normally. The residual structure can ensure that the network structure can still converge under deep conditions, and the model can be trained. (2) The deeper the network, the better the characteristics of the expression, and the effect of classification and detection will increase. (3) 1*1 convolution in the residual, using the idea of network in network, greatly reducing the channel of each convolution, on the one hand reducing the amount of parameters (the larger the parameter, the larger the saved model), On the other hand, the amount of calculation is reduced to some extent. The structure of YOLOv3 is shown in Figure 2.



	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Figure 2. YOLOv3 Structure

We borrowed from the literature [12] to use the spatial attention module on low-level features, and used the idea of channel attention module on advanced features to improve the network structure of YOLOv3 and improve the accuracy of target recognition. At the same time, the loss function of YOLOv3 is improved, and the target positioning accuracy of the image edge is improved. We recognize that certain levels of inaccurate information can cause performance degradation or even mispredictions. It's important to filter these and more valuable features. According to the characteristics of different levels of features, the channel is used to focus on high-level features, and spatial attention is used to select effective features for low-level features. In addition, spatial attention is not used for advanced features, as advanced features that contain high abstract semantics do not require filtering of spatial information. Because there are few semantic differences between the different channels of the low-level function, channel attention is not used for low-level functions.

In order to quickly deal with emergencies and minimize the probability of traffic accidents, the driver hopes that the field of view is large enough and the goal is clear enough. YOLOv3 is a typical 1-stage target detection algorithm with a huge speed advantage, but the positioning accuracy is not very high. We have improved the loss function of YOLOv3 as shown in Equation 3.

$$L_s = -\sum_{i=0}^{size(Y)}(a_s Y_i \log(P_i) + (1-a_s)(1-Y_i)\log(1-P_i)) \qquad (3)$$

Where Y is the ground truth and P is the salient map of the network output, indicating the balance parameters of the positive and negative samples, set a_s= 0.536, which is calculated from the groundtruth of the training set. However, the loss function simply provides a general guide to generating a saliency map. A simpler strategy is used to emphasize the generation of significant object boundary details. First, the Laplacian operator is used to obtain the boundary of the ground truth and saliency map of the network output, and then the cross entropy loss is used to supervise the generation of significant object boundaries. Figure 3 and Figure 4 below show the YOLOv3 test results before improvement and the improved YOLOv3 test results. By comparison, we found that the improved network small target detection ability is stronger.

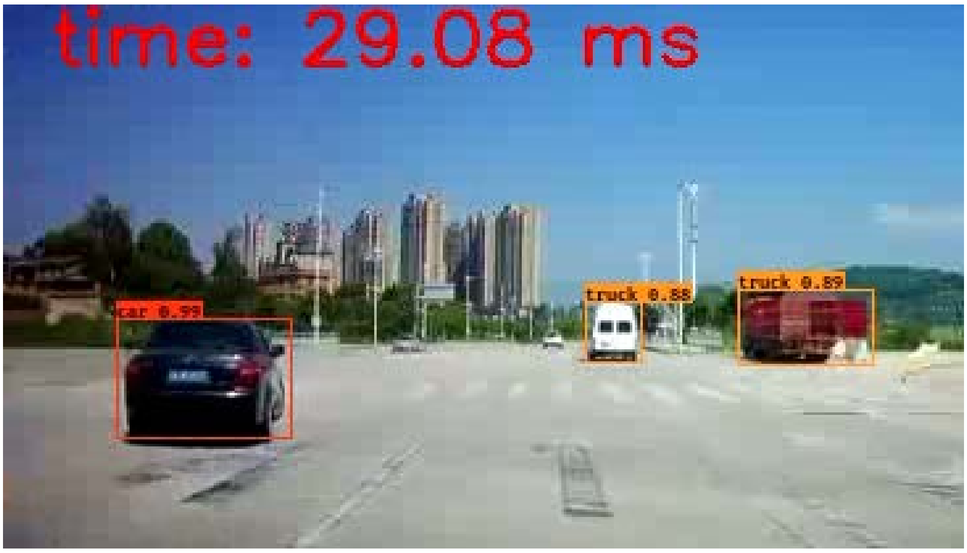


Figure 3. YOLOv3 target detection results

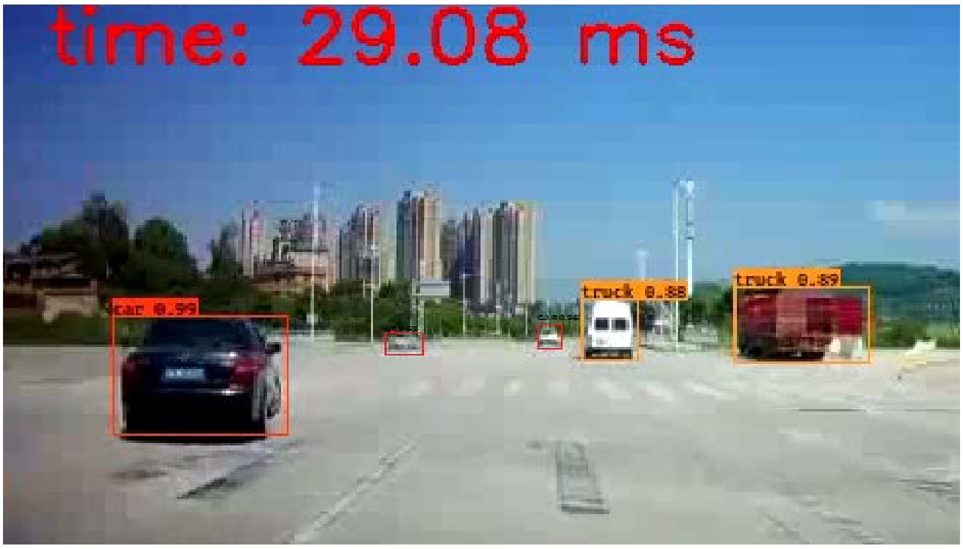


Figure 4. Improved target detection results for YOLOv3

3.4. Image Partition Recognition and Voice Prompts

In order to make it easier for car drivers to deal with various types of events in driving, we have identified and collected the images captured by the camera. The way we partition images is to perform horizontal dynamic partitioning according to specific needs. The height and width of the images we collect are H and W, respectively. The number of partitions to be partitioned is K (if K_Y is used for average partitions, and K_N is not used for average partitions), the area of each area after average partitioning. The calculation formula is as shown in Equation 4.

$$S_Y=H*W/K_Y \tag{4}$$

In order to give the driver the most direct voice prompt, we called Baidu's API.

4. Experiment

We evaluate the proposed network on the challenging KITTI dataset [13], which contains 7481 training images and7518 testing images with calibrated camera parameters. The KITTI data set is one of the most authoritative data sets for evaluating the performance of algorithms in the autopilot domain. We made detailed comparisons with other mainstream algorithms on the KITTI dataset. The comparison results are shown in Table 1. From Table 1, we can see that our target recognition accuracy is the highest, especially when it is difficult to achieve other goals.

Table 1. Performance comparison on KITTI dataset

	Cars			Pedestrians			Cyclists		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
LSVM-MD PM-sv	68.02	56.48	44.18	47.74	39.36	35.95	35.04	27.50	26.21
DPM-VOC +VP	74.95	64.71	48.76	59.48	44.86	40.37	42.43	31.08	28.23
Regionlets	84.75	76.45	59.70	73.14	61.15	55.21	70.41	58.72	51.83
Faster R-CNN	86.71	81.84	71.12	78.86	65.90	61.18	72.26	63.35	55.90
M3DOD	92.33	88.66	78.96	80.35	66.68	63.44	76.04	66.36	58.87
Ours	93.28	90.01	80.25	80.67	67.26	64.13	76.67	67.23	60.56

5. Conclusions

In this paper, we propose a new target recognition algorithm based on YOLOv3. We used the Spatial Attention Module on low-level features, the Channel Attention Module on advanced features, and we improved the loss function. Our method has been tested on the KITTI dataset to achieve the best results, especially when it is difficult to achieve other goals.

6. Acknowledgments

All the works were conducted at Forestry Intelligent Equipment Engineering Research Center.

The work is supported by Heilongjiang Oriental university key scientific research project in 2019(HDFKY190101) and Heilongjiang province education science "13th five-year plan" 2018 key project (GBB1318041)

7. References

[1] Mukhtar Amir, Xia Likun, Tang Tongboon. Vehicle detection techniques for collision avoidance systems: A review[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(5): 2318-2338.

[2] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun. Monocular 3d object detection for autonomous driving. In Conference on Computer Vision and Pattern Recognition(CVPR), pages 2147–2156, 2016.

[3] Song Wenjie, Yang Yi, Fu Mengyin, et al. Real-time obstacles detection and status classification for collision warning in a vehicle active safety system[J]. IEEE Transactions on Intelligent Transportation Systems, 2018, 19(3): 758-773.

[4] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia. Multi-view 3d object detection network for autonomous driving. In IEEECVPR, volume 1, page 3, 2017.

[5] X. Cheng, P.Wang, and R. Yang. Depth estimation via affinity learned with convolutional spatial propagation network. In European Conference on Computer Vision, pages 108–125, 2018.

[6] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. NIPS, 2015.

[7] T.-y. Lin, F. Ai, and P. Doll. Focal Loss for Dense Object Detection. ICCV, 2017.

[8] W. Luo, B. Yang, and R. Urtasun. Fast and Furious: Real Time End-to-End 3D Detection,Tracking and Motion Forecasting with a Single Convolutional Net. CVPR, 2018.

[9] S. Wang, D. Jia, and X. Weng. Deep Reinforcement Learning for Autonomous Driving. arXiv:1811.11329, 2018. URL http://arxiv.org/abs/1811.11329.

[10] S. Casas, W. Luo, and R. Urtasun. IntentNet: Learning to Predict Intention from Raw Sensor Data. CoRL, 2018.

[11] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[12] Ting Zhao,Xiangqian Wu.Pyramid Feature Attention Network for Saliency detection. CVPR, 2019.

[13] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomousdriving? the kitti vision benchmark suite. In ComputerVision and Pattern Recognition (CVPR), pages 3354–3361. IEEE, 2012.