Ph.D Forum: Dynamic Obstacle Detection in Traffic Environments

Gopi Krishna Erabati
Electrical and Computer Engineering
University of Coimbra
Coimbra, Portugal
gopi.erabati@uc.pt

Helder Araujo
Electrical and Computer Engineering
University of Coimbra
Coimbra, Portugal
helder@isr.uc.pt

ABSTRACT

The research on autonomous vehicles has grown increasingly with the advent of neural networks. Dynamic obstacle detection is a fundamental step for self-driving vehicles in traffic environments. This paper presents a comparison of state-of-art object detection techniques like Faster R-CNN, YOLO and SSD with 2D image data. The algorithms for detection in driving, must be reliable, robust and should have a real time performance. The three methods are trained and tested on PASCAL VOC 2007 and 2012 datasets and both qualitative and quantitative results are presented. SSD model can be seen as a trade-off for speed and small object detection. A novel method for object detection using 3D data (RGB and depth) is proposed. The proposed model incorporates two stage architecture modality for RGB and depth processing and later fused hierarchically. The model will be trained and tested on RGB-D dataset in the future.

1 Introduction

In the past few decades, systems particularly vehicles are focused to be automated to trim down the cause of accidents by humans. Distraction, speeding, drunk driving, recklessness are some of the instances the human driver would create problems in traffic environments. The consequences of these actions by humans, may vary from damage to property to loss of life. With the motto of reducing accident severity and injury, pre-crash systems is becoming an active area of research among vehicle manufacturers, research institutions and universities. Advanced driver-assistance systems (ADAS) and self-driving (autonomous) vehicles are interesting solutions to such problems.

Obstacle detection is a fundamental step for autonomous vehicles. In computer vision literature, obstacle/object detection is a twofold process, to classify which category obstacle belongs to (obstacle classification) and to determine where obstacle is located in a given image (obstacle localization). This paper presents a comparison of vision based obstacle detection methods which uses 2D images as input and propose a new method for obstacle detection using 3D data. Obstacle detection using optical sensors is very challenging due to high within class variability in vehicle appearance. The appearance of vehicles may vary in shape, size, color; depends on pose and other objects; illumination changes; clutter background and occlusions.

ICDSC 2019, September 9–11, 2019, Trento, Italy

ACM ISBN 978-1-4503-7189-6/19/09.
https://doi.org/10.1145/3349801.3357134

This paper has the following structure: we present the state-of-art obstacle detection methods in Section 2. A comparison of three different methods of object detection using 2D data is presented in Section 3. Section 4 presents results and discussion. Section 5 presents a method of obstacle detection using 3D data and future work. Section 6 describes conclusion.

2 Related Work

The frameworks for object detection can be classified into three types : 1) Sliding window detectors with Neural networks, 2) Traditional object detection pipeline, generating regions proposals at first and then classifying each region into object category, and 3) Object detection as a classification problem with unified architecture to both classify and localize the objects. The region proposal methods include R-CNN, Fast R-CNN, Faster R-CNN [1], R-FCN, FPN, Mask R-CNN. The classification based one stage methods include YOLO, YOLOv2, YOLOv3 [2], SSD [3], DSSD.

3 Dynamic Obstacle Detection - Comparison

3.1 Two Stage Method : Faster R-CNN

Faster R-CNN [7] is the canonical model of deep learning based object detection. It comprises of two stages : 1) Propose Regions Of Interest (ROI), and 2) Classify and localize objects in ROIs. It is an improved version of its predecessors R-CNN and Fast R-CNN.

This model has similar design as Fast R-CNN except that it replaces the slow region proposal method of selective search by an internal deep network called region proposal network (RPN) to improve the speed, thus the name Faster R-CNN. RPN takes the output feature maps of CNN as input. It slides 3x3 filters over the feature maps to create class-agnostic region proposals using CNN. In Faster R-CNN anchor boxes of three scales and three aspect ratios are adopted. The inference time is about 0.2 seconds.

3.2 Single Stage Methods : YOLO, SSD

Single stage frameworks based on global classification/regression, can directly map from image pixels to bounding box coordinates and class probabilities, thus reduce time expense and work in real time applications. Single stage methods like YOLOv3 [2] and SSD [3] are compared in this work.

YOLOv3: This network only looks the image once to detect multiple objects. thus the name You Only Look Once (YOLO). In this method, the image is divided into S x S grid. If the center of object falls into a grid cell, that grid cell is responsible for detecting that object. Each grid cell predicts: 1) 'k' bounding boxes (which provides center of box relative to grid cell (x,y) and width, height of box relative to image (w, h)), 2) confidence scores P(Objects), and 3) conditional class probability $P(Class_i|object)$ - 'C' classes. Thus, the output size is S x S x k (C+5).

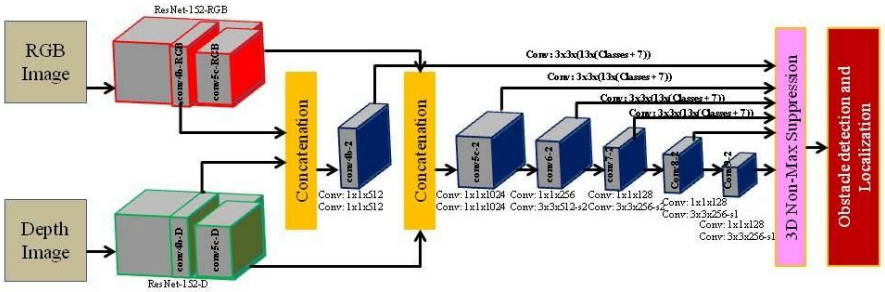

Figure 1: Network architecture for object detection using RGB-D data

SSD: This object detection technique also provides enormous speed gains over Faster R-CNN because Single Shot Detector (SSD) does the detection by passing through input image only once in single shot to predict object class and bounding box. Concretely, given an input image and a set of ground truth labels, SSD does: 1) Pass the image through a series of convolutional layers, yielding several set of feature maps at different scales. 2) For each location in each of feature maps, use a 3x3 convolutional filter to evaluate a small set of default bounding boxes. 3) For each box, simultaneously predict bounding box offset and class probabilities.

4 Results and Discussion

The three models are trained and tested with PASCAL VOC 2007 and 2012 datasets on Google Colab K-80 GPU. As shown in Figure 2, Faster R-CNN is able to detect the small objects in the background whereas YOLOv3 was not stable to detect small objects in background even though three scales were used to detect objects in YOLOv3. As shown in Table 1, Faster R-CNN suffers from speed as it could not provide a real time detection compared to YOLOv3 and SSD. SSD can be seen as a tradeoff between speed and small object detection, as it was able to detect small objects with real time inference time.

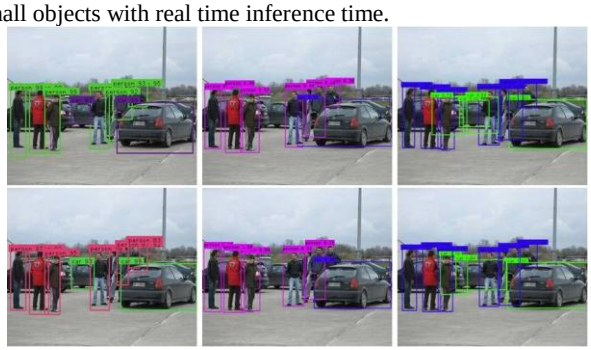

Figure 2: (left) Faster R-CNN(ResNet50, VGG16), (middle) YOLOv3(ResNet50, MBNetv1), (right) SSD(ResNet50, VGG16)

Table 1: Mean Average Precision (mAP) and speed

Model	Backbone	mAP	Speed(fps)
Faster R-CNN	ResNet50	72.7	3
	VGG16	69.8	4
YOLOv3	ResNet50	71.5	25
	MobileNetv1	69.6	29
SSD	ResNet50	74.1	24
	VGG16	73.2	26
	MobileNetv1	71.1	28

5 Object Detection with 3D data - proposal and future work

3D data (RGB + depth) is being used in many applications with the advent of low cost RGB-D cameras. RGB data provides information about appearance and texture and depth data provides additional information about object shape and it is invariant to lighting conditions. We propose a new method to detect objects from RGB-D data. The neural network takes RGB and depth image pair as input and produces object classification and localization as output, as shown in Figure 1. Our network consists of two sub-networks to process two modalities (RGB and Depth) and later features are fused hierarchically followed by multi-scale predictions to detect objects of different shapes and aspect ratios. ResNet-152 is used as a backbone for feature extractor. The features from conv4b and conv5c layers of both the sub-networks, which are used to learn appearance and geometric features from RGB and depth images respectively, are fused together. Later, SSD[3] type of multi-scale prediction architecture is used followed by non-max suppression to detect objects. To cope with the issue of scale ambiguity, 3D anchor boxes of different scales are attached to different layers of prediction. In the future, the proposed network will be trained and tested on RGB-D dataset and results will be published.

6 Conclusion

Two stage and single stage object detection methods with 2D data like Faster R-CNN, YOLO and SSD are compared in terms of speed and average precision on PASCAL VOC 2007 and 2012 datasets. SSD method seems to be tradeoff for speed and small object detection compared to YOLO and Faster R-CNN. A novel method for object detection with 3D data is proposed and will be trained and tested in future.

ACKNOWLEDGMENTS

This project is funded by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 765866.

REFERENCES

[1] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards realtime object detection with region proposal networks. In *NIPS, 2015, pp. 91–99.*

[2] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement. In *arXiv:1804.02767, 2018.*

[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *ECCV, 2016.*