IOP Conf. Series: Materials Science and Engineering **261** (2017) 012005 doi:10.1088/1757-899X/261/1/012005

Deep Learning for Real-Time Capable Object Detection and Localization on Mobile Platforms

F. Particke[1,*], R. Kolbenschlag[1], M. Hiller[1], L. Patiño-Studencki[1] and J. Thielecke[1]

[1]Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Information Technologies, Erlangen, Gersmany

[*]florian.particke@fau.de

Abstract. Industry 4.0 is one of the most formative terms in current times. Subject of research are particularly smart and autonomous mobile platforms, which enormously lighten the workload and optimize production processes. In order to interact with humans, the platforms need an in-depth knowledge of the environment. Hence, it is required to detect a variety of static and non-static objects. Goal of this paper is to propose an accurate and real-time capable object detection and localization approach for the use on mobile platforms. A method is introduced to use the powerful detection capabilities of a neural network for the localization of objects. Therefore, detection information of a neural network is combined with depth information from a RGB-D camera, which is mounted on a mobile platform. As detection network, YOLO Version 2 (YOLOv2) is used on a mobile robot. In order to find the detected object in the depth image, the bounding boxes, predicted by YOLOv2, are mapped to the corresponding regions in the depth image. This provides a powerful and extremely fast approach for establishing a real-time-capable Object Locator. In the evaluation part, the localization approach turns out to be very accurate. Nevertheless, it is dependent on the detected object itself and some additional parameters, which are analysed in this paper.

1. Introduction

In future smart factories, every machine will communicate and fuse their information from the physical and virtual world, in order to accomplish fully autonomously their tasks with less effort. Especially smart mobile platforms like robots will enormously lighten the workload and will optimize production processes in Industry 4.0. In order to interact with the environment, the robot has to recognize and localize objects. In addition, especially in traffic control, short response times are important. The contribution of this paper is to establish an object detector and locator for a mobile platform, which is very fast and reliable and works even in unknown environments. With deep learning, currently being state of the art in computer vision, the problem is tackled in this paper by a CNN. Since 2012, every winning team in the ImageNet Large Scale Visual Recognition Challenge [1] has used a convolutional neural net [2], [3]. The recently released YOLO Version 2 (YOLOv2) detection network [4], [5] outperforms current state of the art networks in performance (e.g. VGG Net [6]) and precision (e.g. ZF Net [7]). Additionally, it is empowered to detect almost 9000 different object classes. Hence, it poses as an optimal candidate. As detection networks only locate objects within a camera image, an efficient and real-time capable approach is introduced to translate the powerful detection capabilities of a neural network into the 3D space. It combines detection

information of a neural network with depth information from a RGB-D camera mounted on the platform. In order to find the detected object in the depth image, it maps the bounding boxes, predicted by the detection network, to the corresponding regions in the depth image. It provides a powerful and extremely fast approach for establishing a real-time-capable Object Locator.

The paper is organized as follows. In Section 2 the object detection with YOLOv2 is shortly described, in Section 3 the spatial location approach using detection and depth information is proposed. This approach is evaluated in Section 4. Further research is pointed out in Section 5.

2. Object detection with YOLOv2 on the mobile platform

In this Section, the object detection with YOLOv2 (Subsection 2.1) and the integration into the mobile platform (Subsection 2.2) is described. Goal of this section is to summarize the key features in comparison to other convolutional neural networks. For further information the interested reader is referred to [4].

2.1 Object detection with YOLOv2

You Only Look Once (YOLO) is both name and motto of a fast and precise object detection network, which was released in 2016. The system was made to look only once at an image to detect objects. In Figure 1, the basic approach of YOLOv2 is depicted.

YOLOv2 belongs to the class of so called single-shot networks, e.g. SSD [8]. It models object detection "as a single regression problem straight from image pixels to bounding box coordinates" [9], called detection as regression. Thus, it is significantly faster than common networks and has a less complex network pipeline. Based on the entire image, it predicts category scores and box coordinates for a fixed set of bounding boxes, unifying separate components of object detection. YOLOv2 divides an image in a $S \times S$ grid (cf. Figure 1) and predicts B bounding boxes for each grid cell with four coordinates and a confidence score for those boxes. Additionally, each cell predicts C class probabilities. All in all, this leads to a predicted tensor, which is used for regression.

$$SxSx(B \cdot 5 + C) \tag{1}$$

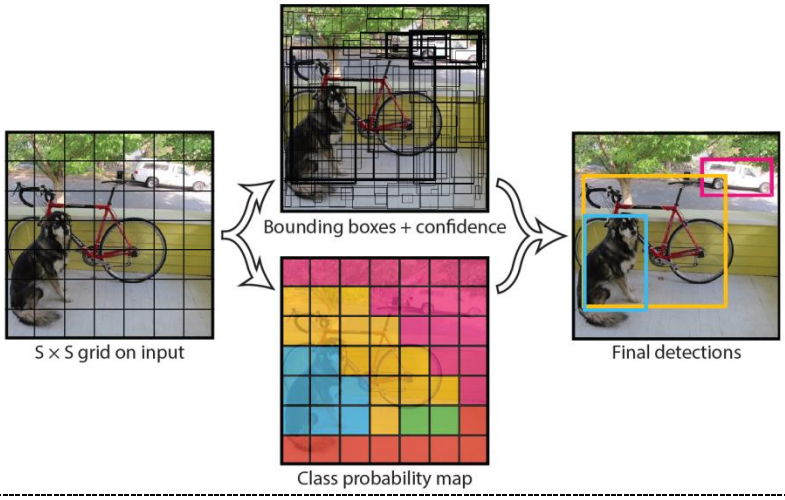

Figure 1. YOLOv2 divides an image into a grid, with each cell predicting a fixed amount of bounding boxes and a class score. Figure is adapted from [10].

Many detection frameworks rely on the powerful and accurate VGG-16 CNN [10] for feature extraction. But it has a major drawback, namely a complex pipeline. This makes it difficult to use in real-time systems, as it requires 30.69 billion floating point operations for a single image in its

convolutional layers [4, p.5]. To be real-time capable, YOLOv2 uses an own convolutional neural network called "Darknet". It is less complex but much faster than VGG-16, requiring only 5.58 billion operations while having only a slightly worse accuracy of 2% than VGG-16. In the ImageNet Challenge, it achieved a top-5 accuracy of 88%.

To sum up, the overall advantages for the integration on a mobile platform are

- YOLOv2 outperforms current state of the art networks in performance and precision
- YOLOv2 is able to detect over 9000 different object classes
- YOLOv2 is delivered as an enclosed program. Therefore, it can be used instantly without the necessity of being created and trained

2.2 Integration into a mobile environment

YOLOv2 is integrated into a mobile environment to fuse its detection capabilities with other sensors on a mobile platform. A robot laboratory is used, which provides space and hardware for testing and implementation of YOLOv2. The robot SUMMIT XL, developed by Robotnik [11], serves as a mobile platform. In order to communicate with sensors of the robot, YOLOv2 is combined with the Robot Operating System (ROS). ROS is a framework, providing a communication infrastructure between sensors, which enables them to pass messages to each other. These messages contain sensor data or status information [12]. The RGB-D Camera is used as a sensor for the detection and localization approach.

3. Spatial localization approach

The approach of fusing the classification and detection information of YOLOv2 with the Depth-Image is described in this section. A general overview is depicted in Figure 2. The RGB-Image is used as input of the YOLOv2 algorithm, which outputs the classification and the detection areas of the objects.

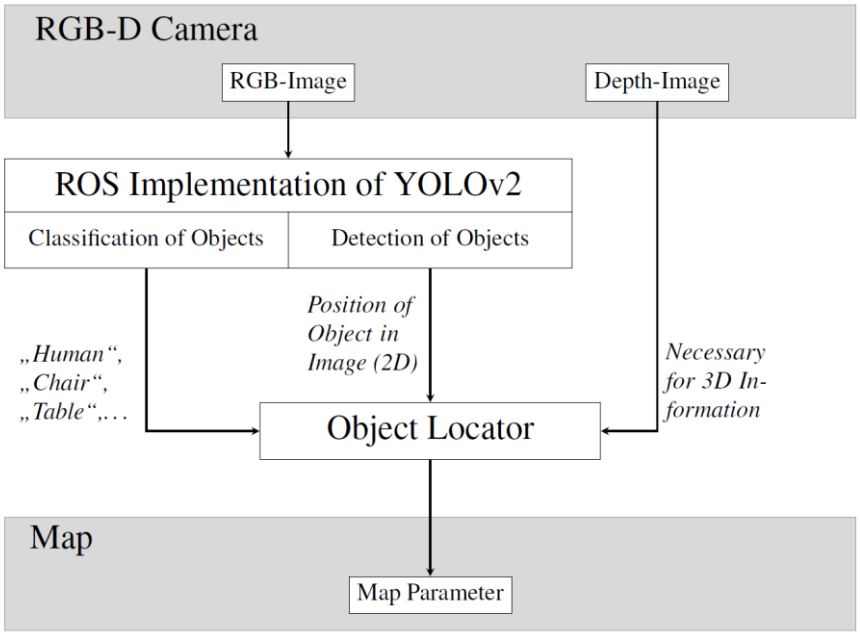

Figure 2. The road map shows the flow of information from the sensor to the created map, which is based on the fusion of deep learning (output of YOLOv2) and the depth-image.

An example of the output of YOLOv2 is presented in Figure 3. Four objects are present in this scenario: chair, cup, sports ball and person. All objects are classified correctly and the bounding boxes are drawn around the objects. The resulting bounding boxes are used for the definition of the regions

of interest in the Depth-Image. An example of a Depth-Image with detected bounding boxes is presented in Figure 4.

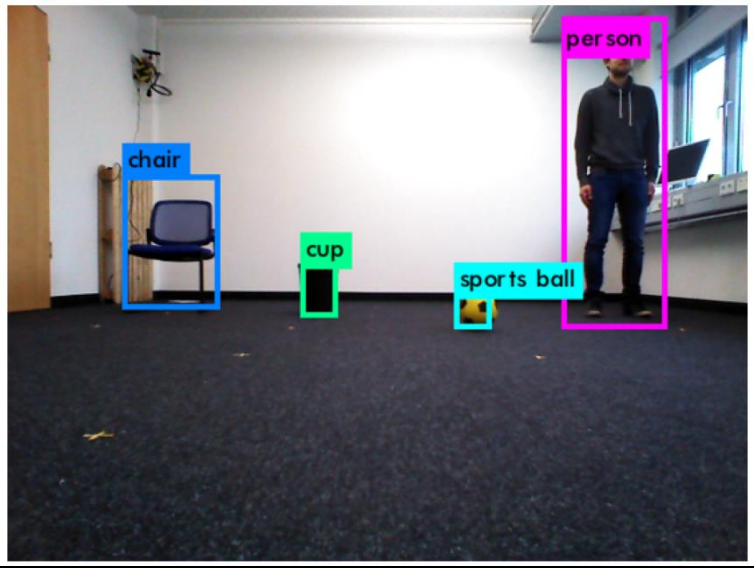

Figure 3. Output of block "ROS Implementation of YOLOv2" of Figure 2. The detection (bounding boxes) and the classification results are depicted. These results are used for the fusion in the Object Locator.

Figure 4. The depth image, which is enhanced by the detection results of YOLOv2. The bounding boxes select the interesting areas for the Object Locator.

The detected bounding boxes in the Depth-Image are used in the Object Locator. The Object Locator performs the following steps. Firstly, the selected bounding boxes are background filtered. The background is always the surface having the largest distance to the camera. So, all depth values within an interval between the maximum depth and a certain threshold are denoted as "background" and are set to zero. For the following experiments, the threshold was set to 0.05 m. Secondly, the depth values are clustered with K-means. Best results are given for K = 4. Finally, the mean average of

the biggest cluster is used for the distance estimate. The classified objects are mapped to an existing database, which represents the current map.

4. Evaluation

The proposed approach of Section 3 is evaluated in dependence of the distance for two different objects, a sports ball and a person. Figure 5 shows the difference between the true depth and the mean predicted depth values from 1m … 5m for the two objects. Due to the fact that the objects are placed in the center of the camera image (x = 0, y = 0), the depth corresponds to the distance between the object and the robot. It can be seen, that in nearer distances the depth values of the sports ball are more accurate than the values of the person. As the surface of the ball is plain, the dominating cluster contains more similar depth values than the surface of a person. As the person is very close to the camera, only the legs of the person are captured by the camera. Therefore, the mean depth value of this cluster slightly differs from the ground truth. The precision of the depth values corresponding to the person gets more accurate at a distance of 2 meters, as there are more parts of the body in the camera view. Hence, the person provides a new kind of surface, which is apparently better located.

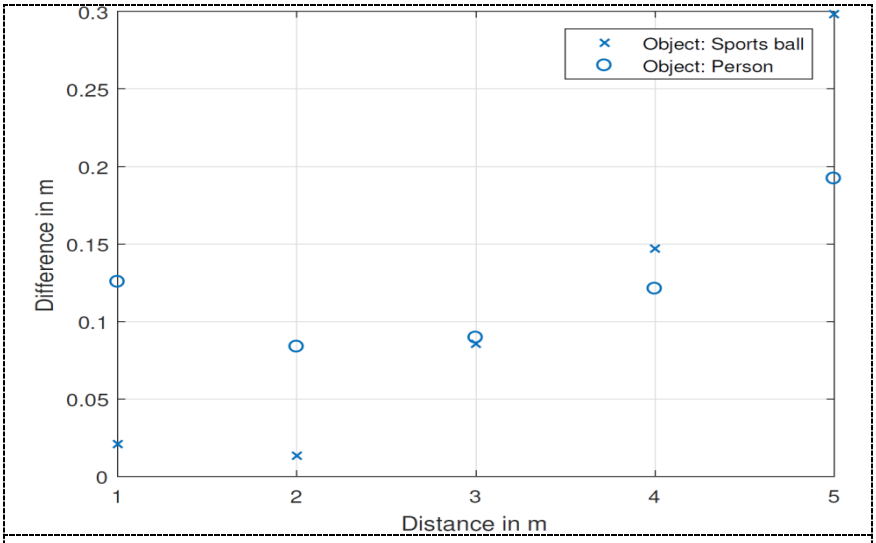

Figure 5. Difference between ground truth and mean predicted depth values in relation to the distance of the object.

With an increasing distance, the depth values of the person are more accurate than the values of the ball, because there are more depth values to evaluate in the chosen cluster and the cluster mean value is less influenced by outliers.

The prediction of the depth values and therefore the localization depends not only on the distance, but also on the surface of the evaluated object. Objects with a plain surface are better localized. In addition, it is shown that the position of near objects is estimated very accurately. Nevertheless, the localization only works well if the correct cluster was chosen. In the case of the ball and person, the number of four clusters was suitable, because no other object was put in the same cluster and every chosen cluster had enough values for a representative centroid. All in all, the approach seems really promising for the fusion of localization with RGB-D and classification with YOLOv2.

5. Conclusion and further work

This paper addressed the object detection within the RGB image of the camera and the fusion of the detection results with depth information. YOLOv2 was used for detection and was integrated into a robotic framework, which made a very fast and powerful object detector available for mobile platforms. It turned out that the detection confidence is very dependent on the number of image points

and the object type. Empowered with an internal object detector, new possibilities were given to find the correct depth regions. The difference between the ground truth and the mean predicted depth values in relation to the distance of the object were analysed. It was shown that there is a strong dependence between the distance of the object to the depth sensor and the accuracy of the approach. Nevertheless, the proposed approach showed very promising results. In posterior works, the localization could be further improved by additionally considering spatial information in the clustering process. Furthermore, instead of the K-Means algorithm the X-Means algorithm could improve the proposed Object Locator. It introduces an approach to find the optimal number of clusters in the cluster space by optimizing the Bayesian Information Criterion. Last but not least, it is necessary to extend the Object Locator to provide a statistically proven confidence score.

References

[1] Stanford Vision Lab. Imagenet large scale visual recognition competition (ilsvrc). http://www.image-net.org/challenges/LSVRC/. [Online; accessed 21-June-2017].
[2] Chiyuan Zhang. Mocha.jl: Deep learning for julia. https://devblogs.nvidia.com/parallelforall/author/czhang/, 2015. [Online; accessed 21-June-2017].
[3] Krizhevsky A, Sutskever I, and Hinton G.. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 25, pages 1097–1105. Curran Associates, Inc, 2012.
[4] Redmon J and Farhadi A. Yolo9000: Better, faster, stronger. arXiv preprint arXiv:1612.08242, 2017.
[5] ThundeRatz robotics. ros_yolo2 - yolov2 integration with ros. https://github.com/ThundeRatz/ros_yolo2, 2017. [Online; accessed 21-June-2017]
[6] Liu, S., & Deng, W. (2015, November). Very deep convolutional neural network based image classification using small training sample size. In *Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on* (pp. 730-734). IEEE.
[7] Zeiler, M. D., & Fergus, R. (2014, September). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818-833). Springer, Cham
[8] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C, and Berg A. Ssd: Single shot multibox detector. In European Conference on Computer Vision, pages 21–37. Springer, 2016.
[9] Redmon J, Divvala S, Girshick R, and Farhadi A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 779–788, 2016.
[10] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
[11] Robotnik. Mobile robot summit xl | robotnik. http://www.robotnik.eu/mobile-robots/summit-xl/. [Online; accessed 21-June-2016].
[12] Open Source Robotics Foundation. Robot operating system. http://www.ros.org/about-ros/. [Online; accessed 21-June-2017].