Realtime Road User Detection and Classification with Single Pass Deep Learning

Robin Manhaeve[1]
Promotors: Prof. dr. Luc De Raedt[1], Dr. ir. Kurt De Grave[2]
Supervisors: Dr. ir. Kurt De Grave[2], Dr. ir. Laura Antanas[1]

[1] Dept. of Computer Science, K.U.Leuven, Belgium
robin.manhaeve@cs.kuleuven.be
[2] Flanders Make, Belgium

In autonomous vehicles (AV), computer vision supplements sensors such as radar and lidar because it excels at classifying road users into separate classes. Correct classification improves the prediction of their future movement and the risk of collision. However, in the real-time road environment, response time is as important as accuracy. A lot of highly accurate methods have been developed, but many have a large delay (detection runtime) [2], making them less suitable for AV. These techniques are slow because they are based on a two-step method (detection proposals and a classifier). Often, the classifier has to be evaluated many times, creating a large and variable delay. Lately, a new class of less accurate but faster methods has appeared: the single-pass methods. In this thesis, we discuss different positions in the trade-off between speed and accuracy. We also test the use of LWIR images for AV. These might be vital for road user detection during non-optimal lighting conditions, but existing research is limited.
The method used in this thesis is essentially the YOLO-method. [4] It performs detection in a single evaluation of the neural network, resulting in a big speed-up compared to two-phase methods. The original paper used a network based on their *Extraction* network. Because Extraction is not available in the Caffe framework (used in this thesis as it is more prevalent in research than the YOLO-framework), GoogLeNet [5] is used, of which Extraction is a simplification with similar performance. The effect of a ResNet-50-based [3] network on the accuracy and delay is also tested. The method is evaluated on the following datasets: Caltech Pedestrian Detection Benchmark [1], KITTI Vision Benchmark Suite [2] and a new dataset developed by Flanders Make. This new dataset consists of several camera feeds, radar, lidar, and IR images. Only the center front visual light (VL) camera and the Xenics LWIR (shown in Figure 1) were used in the experiments. This dataset is split into two parts. The first part contains only VL images (2,589 frames). The second part contains both VL and IR images (681 frames). In contrast to other datasets, this part was recorded during the evening. This allows the performance of VL detection to be evaluated under non-optimal lighting conditions. To perform detection on the IR images, several setups are compared: only VL images, only IR images, evaluating separate networks for VL and IR, and performing detection with a single network on the

Fig. 1. LWIR image from the Flanders Make dataset

combined 4-channel VL+IR image. The IR network was trained by first taking a network trained on KITTI and the first part of the Flanders Make data, then merging the RGB channels into a single channel (summing the filters), and finally fine-tuning on the LWIR data.
YOLO was first evaluated on the Caltech Pedestrian Dataset. In comparison to the most accurate result (F-DNN), this method is noticeably less accurate (.64 as opposed to .89 mAP on the *reasonable* set). It does display a lower delay: 40 ms as opposed to 300 ms (the difference being underestimated as the same hardware was not available). The Caltech dataset proved challenging for this method because most objects are small, a known weakness [4].
The method's accuracy (.42, .41 and .30 mAP) on cars, pedestrians and cyclists of moderate difficulty from the KITTI Vision Benchmark, failed to compete with the start of the art (SAIT: .90, .73 and .76 mAP). However, since the processing time of the methods are available, this benchmark gives a better overview on the Pareto-optimal methods. Here, Pareto-optimality is defined in the trade-off between delay and detection accuracy. YOLO showed to be close to the Pareto-front by being quite fast. A network based on ResNet-50 was also trained on this dataset. Although it had a small positive impact on the accuracy (1-4% mAP), it slowed the method down by 50%, pushing it away from the Pareto-front.
On the second part of the Flanders Make dataset, detection using only VL performed very poorly (.06 mAP on cars). Only using IR images yields superior results (0.59 mAP on cars). Combining the detections from VL and IR images (evaluating two separate networks) further improves the result for cars (0.64 mAP), however, other classes performed worse due to errors being introduced by detections in the VL images. Detection on both images in a single network did not perform well (0.16 mAP on cars).

Single-pass methods are less accurate than two-pass methods, but they are a lot faster. Considering the trade-off between speed and accuracy, single-pass methods prove more applicable for AV. The use of very deep networks in AV is not desirable, as their impact on the delay is too big. Finally, incorporating IR images is vital for AV vision during non-optimal lighting conditions.

References

1. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. PAMI 34 (2012)
2. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. International Journal of Robotics Research (IJRR) (2013)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
4. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
5. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9 (2015)