15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

Road User Detection with Convolutional Neural Networks: An Application to the Autonomous Shuttle WEpod

Floris Gaisser f.gaisser@tudelft.nl
Pieter. P Jonker p.p.jonker@tudelft.nl
Intelligent Vehicles and Cognitive Robotics - Faculty of Machanical Engineering
Delft University of Technology, Mekelweg 2 2628CD Delft

Abstract

Over a million fatal accidents occur every year with road vehicles. Road user detection for Advanced Driver Assistance Systems and Autonomous Vehicles could significantly reduce the number of accidents. Despite the research focus on road user detection and such systems, there is a surprising lack of research in real-world applications. In this work, radar and camera data are combined on an autonomous shuttle called 'WEpod', driving on the public road in Wageningen, The Netherlands. With experiments we show that our method reduces the candidate region margin to $0.2m$ *and reduces the miss rate significantly. Furthermore, our specifically trained Convolutional Neural Network improves the performance by* 1.4% *over vision-based road user detection, and combined with radars we improve by* 7.6%*. Finally, with our approach we show a performance of* 95.1% *on the WEpod while driving on the public road.*

1 Introduction

Every year more than a million fatal traffic accidents occur [11]. About 25% of these could be attributed to lack of attention of the driver [18]. Therefore, road user detection has been a focus of research for Advanced Driver Assistance Systems (ADAS). Furthermore, Autonomous Vehicles also require reliable road user detection to safely navigate in traffic.

There are two common approaches to road user detection: vision-based and fusion-based. In vision-based systems, candidate regions are generated and subsequently classified. In fusion-based systems, data from multiple sensors is combined, either one sensor generates candidate regions for classification by the other sensor or both are simultaneously used to improve performance.

Pedestrians are the most vulnerable road users, more difficult to detect, and behave more erratic than other road users. Hence most vision-based research focuses on pedestrian detection [19, 2], although some literature can be found on vision-based car detection [6] and on-road vehicle detection [15].

Despite the extensive work on vision based systems, recent research shows that there is still significant room for improvement. Compared to a human benchmark, vision systems have about 100 times more false detections and miss 10 times more possible detections [19]. Fusion-based detection where one sensor generates candidate regions could reduce the number of false detections [14, 13, 8, 9].

Although automation in vehicles is growing, there is a surprising lack of research in on-line applications. Most research focusses on off-line processing of real world data and detection of a single type of road-user. However, for real-world applications, such as an autonomous vehicle, all road-users such as pedestrians, bicyclists, motors, cars, trucks and busses have to be detected in real-time on the vehicle.

The goal of this paper is to show road user detection in an autonomous shuttle that drives on the public road. Therefore, we introduce a fusion-based road user detection approach combining radar and camera sensors with our proposed dynamic candidate regions method. Furthermore we train a convolutional neural network (*ConvNet*) with contrastive loss for classification. This approach is evaluated on a standard dataset for benchmarking and we apply it in our autonomous shuttle.

This paper is organized as follows. First, a short description is given on related work of fusion-based detection and classification using Convolutional Neural Networks. Section 2 gives background on our approach, followed by the experiments in Section 3. The results are discussed and a conclusion is given in Section 4.

1.1 Related work

Detection can generally be split in two parts; first, detecting candidate regions of interest and second, classifying these as relevant or irrelevant. In general, two different sensors are used in fusion-based detection. Laser scanners / Lidars are often used for road user detection, however, they depend on light and are obstructed by fog and rain, making them unreliable in many real-world situations. [13, 12, 14, 10].

Radars detect objects with lower frequency electromagnetic wave reflections and are not much influenced by weather conditions. Literature has shown that smaller objects, such as pedestrians and bicyclists, can also be detected [17, 1] and hence using radars is a common choice in real-world applications, although they are seldom combined with visual data [9, 8].

Since all road users are visually distinguishable, a camera is generally well suited for classification. However, an abstraction from raw pixel data into classes is needed, which is generally described as a vector of probabilities for each of the classes.

ConvNets are the state-of-the-art method to classify multi-class visual problems [4]. Multiple convolutional, pooling and rectification layers are combined, so that the visual input is abstracted into lower dimensional data. This data describes the differences and unique visual components of each class. Multiple fully connected layers classify this data into probabilities for each of the classes [4].

These ConvNets have to be trained; thus many images with known classifications are fed into the network

and a loss layer provides feedback of the performance to the network [4]. This approach puts an emphasis on learning a general visual description of the class. However, in road-user detection, the difference between a relevant and non-relevant detection also needs to be learned. A Siamese network with contrastive loss is an approach to learn this difference [5] and has shown better results than the traditional class-based training [16]. Therefore, we apply this approach in our system.

2 Method

For our fusion-based road user detection method we combine radar detections with classification of visual data. Other work [8, 9] reported similar approaches, however, we improved two aspects of their approach. Firstly, the dynamic candidate regions method fuses radar and image data more accurately. Secondly, the contrastive loss function used in training our ConvNet improves the precision and recall of the classification. In the next two sections we give a detailed description of these two aspects.

2.1 Dynamic Candidate Region

In our approach the radar detections are transformed into the camera image as regions of interests, which are then fed into a classifier. A dynamic projection of the detection location to the image plane combined with the detection distance and camera calibration allow us to generate candidate regions of interest at real-world scale in real-time. The method is detailed in the next paragraphs.

Detections in the radar plane are provided by the radar in the form of distance and angle (d_r, θ_r). Assuming that all objects are standing on the ground, they can be transformed into the vehicle coordinate system. This is detailed in equation 1, with the sensor location (x_r, y_r) and orientation (α_r) with respect to the vehicle (X_w, Y_w, Z_w).

$$\begin{bmatrix} X_w \\ Y_w \\ Z_w \end{bmatrix} = \begin{bmatrix} \cos\alpha_r & -\sin\alpha_r & x_r \\ \sin\alpha_r & \cos\alpha_r & y_r \\ 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} d_r * \cos\theta_r \\ d_r * \sin\theta_r \\ 1 \end{bmatrix} \quad (1)$$

In contrast to the work of Milch and Behrens [9] and Premebida and Nunes [13], we do not consider the road to be flat. Therefore, we incorporate the roll (β) and pitch (γ) of the vehicle, measured by the vehicle's inertia measurement unit. These values are obtained from the gravitational direction and the angles are defined to the horizontal coordinate system and hence they are not Euler angles, which can be seen from the rotation matrix in equation 2.

$$\begin{bmatrix} X_{rp} \\ Y_{rp} \\ Z_{rp} \end{bmatrix} = \begin{bmatrix} \cos\gamma & 0 & -\sin\gamma \\ 0 & \cos\beta & \sin\beta \\ \sin\gamma & -\sin\beta & \cos\beta * \cos\gamma \end{bmatrix} \cdot \begin{bmatrix} X_w \\ Y_w \\ Z_w \end{bmatrix} \quad (2)$$

As the detections are rotated with the vehicle's motion, they can be transformed to the camera coordinate system (X_c, Y_c, Z_c). This is described in equation 3, with the camera position (x_c, y_c, z_c) and orientation (α_c). These coordinates can be further projected to image coordinates (u, v), e.g. with the OpenCV [7] *projectPoints* function, also taking lens distortion into account.

$$\begin{bmatrix} Z_c \\ -X_c \\ -Y_c \end{bmatrix} = \begin{bmatrix} \cos\alpha_c & -\sin\alpha_c & 0 & -x_c \\ \sin\alpha_c & \cos\alpha_c & 0 & -y_c \\ 0 & 0 & 1 & -z_c \end{bmatrix} \cdot \begin{bmatrix} X_{rp} \\ Y_{rp} \\ Z_{rp} \\ 1 \end{bmatrix} \quad (3)$$

As the distance to the detection is available and ConvNets need a fixed sized input, every candidate region can be created in acordance with its real-world size. To allow pedestrians, cyclists and cars with a maximum height of 2 m to fit, crops of 2.4×2.4 m are created with a 0.2 m margin to compensate for variations. This margin is chosen based on the results of experiment 1 (Section 3.1). However, this is not wide enough for vehicles seen from the side. Fortunately, the radar also provides a width measure of the detection so additional crops to both sides can be created.

2.2 Classification

Convolutional Neural Networks (ConvNet) have been highly effective in image detection and classification and found their way to fusion-based pedestrian detection [14]. ConvNets learn a representation of the input images with different levels of abstraction. In contrast to the general approach to increase the network's size and complexity to improve the classification performance, we are bound by the available processing capacity. All candidate regions have to be classified within a 66 ms cycle time. In the following paragraphs, we describe our approach.

Neural Networks learn a lower dimensional representation of the input data through various convolutional and rectification layers. These are followed by fully connected layers that can learn the relation between the more abstract representation and the desired label output. Our approach is similar to this and is shown in Figure 1, where 64f stands for 64 filters, 6×6 px for the filter size and 2s for a step of 2 px.

The real-world size of the different road-users differ, a pedestrian is about 1×2 m, while a car is about 2×1.5 m seen from behind up to 6×1.5 m seen from the side. However, all these types have to be recognized from a $2.4 \times 2.4m$ crop, thus unrelated information is also present in each input. Therefore, the output is not a single prediction for each class, but rather a grid of 8×8 predictions. Each grid cell represents an area of 1×1 m and can be used to extract the smallest area in which a road-user might be present or not. Since fully connected layers cannot give such an output, this layer is replaced by a convolutional layer, which has a filter size of 1×1, creating a *convolutional fully connected* layer.

The general approach is to learn an abstraction toward the class label with *softmax*. However, two visually similar classes, such as a pedestrian and bicyclist,

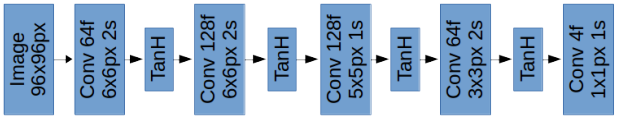

Figure 1. Our neural network architecture

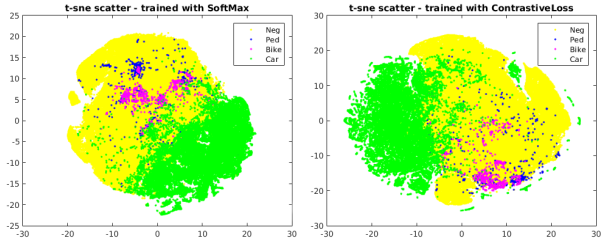

Figure 2. Class separation with t-sne

will often be classified wrongly. The underlying cause can be explained by visualizing the output of the last convolutional layer with t-sne. The outputs of pedestrians and bicyclists overlap as displayed in Figure 2. The contrastive loss method is used to increase the separation between these classes and reduce the difference between two similar inputs [5].

Our learning approach is to first train the network normally to obtain basic abstraction. Next, we obtain a set of pairs that have similar abstraction but belong to different classes as well as pairs that belong to the same class but have different abstractions. Except for the convolutional fully connected layer, the network is trained with contrastive loss.

3 Experiments

The goal of this paper is to show road user detection in a real-world application. For this we use the autonomous shuttle WEpod which is driving on the public road of the Wageningen University's campus in the Netherlands. In this paper we use the three front camera-radar pairs of the nine pairs around the vehicle.

In the sections below we report experiments to evaluate our method. However, we also use the KITTI dataset to have a comparable benchmark [3].

3.1 Experiment 1 - Dynamic Candidate Regions

The dynamic projection as described in Section 2.1, should make the candidate regions more accurate, since we do not assume the road to be flat. To evaluate this, three different types of road sections on the WEpod's route are chosen: a straight road, a curve and a speed bump. Three recordings of 20 seconds are taken for the flat and curved road types, while for the speed bumps only 2-3 seconds could be taken as bumps are short.

Table 1. Comparison of traditional static and our proposed dynamic candidate regions method

Road	Straight	Curve	Bumps
Static			
Roll [deg]	1.64 ±0.34	1.00 ±0.43	1.68 ±0.67
Pitch [px]	4.2 ±1.3	2.0 ±1.3	9.6 ±10.6
Margin [m]	*0.30–0.75*	*0.23–0.57*	*0.65–1.62*
Dynamic			
Roll [deg]	0.27 ±0.18	0.17 ±0.18	0.12±0.12
Pitch [px]	1.1 ±0.7	0.9 ±0.7	1.3 ±1.4
Margin [m]	**0.09–0.22**	**0.07–0.17**	**0.08–0.21**

The horizon is manually annotated in each frame. For the static projection method, the roll and pitch values are set to zero. The pitch accuracy is calculated from the vertical pixel distance in the center, and the roll accuracy from the angle difference between the annotated and projected horizon. Table 1 details the accuracy and variation for both projection methods of the roll and pitch in each type of road section.

Furthermore, a margin is calculated from the roll and pitch variations, so that a detection of 2×2 m would fit in the candidate region. The first value represents a 2σ variation on a detection (d_r, θ_r) at 10 m distance and 28 deg angle, and the second at 25 m and 28 deg. From these results, a margin of 0.2 m is chosen for the candidate regions, resulting in a crop size of 2.4×2.4 m.

3.2 Experiment 2 - Learning

To compare the classification performance of our contrastive loss training with conventional learning, a training set of images of fully visible pedestrians, bicyclists and cars at a maximum distance of 25 m was created from the KITTI database. For evaluation, a sliding projected window approach creates image crops at different distances from the test images. Figure 3a details the recall and precision of the the different methods. Figure 2 shows the separation of the different classes from the fourth convolutional layer outputs.

3.3 Experiment 3 - Radar Fusion

Fusing image classification with radar candidate regions should improve the recall performance. The positive evaluation set is created as 2.4×2.4 m candidate regions from the 3D position of the ground truth and the camera calibration from the KITTI database. A total of 9 crops is created for each true detection, by adding random variation of max $0.3m$ in x and y direction to simulate the radar detection inaccuracy and the proposal accuracy. The negative evaluation set is created from random projected candidate regions. Detections are considered correct if the Intersection over Union (IoU) ≥ 0.5. The same networks as were used in experiment 2 are evaluated, and also per type of road user. The results in figure 3b show that fusion-based detection improves vision-based detection.

3.4 Experiment 4 - Real-world Application

While driving on the campus, recordings from the front three sensor pairs were obtained. In total, 423 pedestrians, 864 bicyclists and 1329 cars were manually annotated. The radar detections were used to generate dynamic candidate regions which were classified with our ConvNet. We distinguished between relevant and non-relevant classification (*CL-R/NR*) and classification as the correct type of road user (*CL-RU*). Furthermore, we combined the detection over three successive images and accepted the classification if two are the same (*CL-3*). The results are shown in Figure 3c.

4 Conclusion

This paper shows road user detection on an autonomous shuttle driving on the public road. To

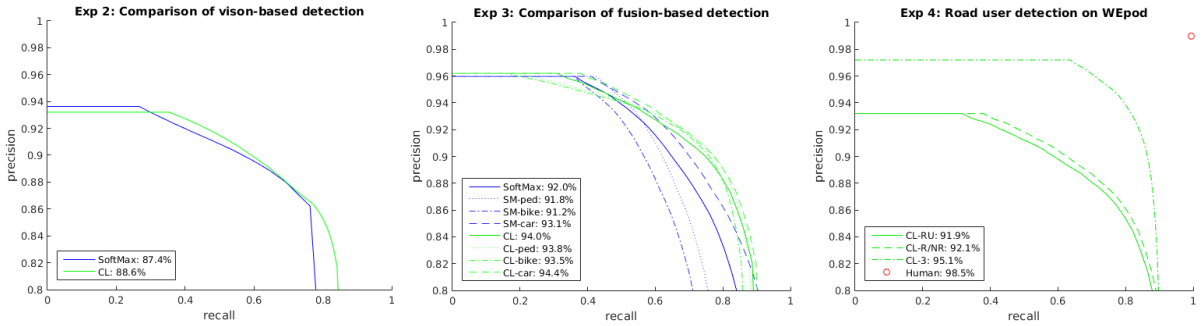

Figure 3. Results: a) Vision-based performance b) Fusion-based performance c) Real-world performance

achieve this we combined radar and camera sensors and introduced two improvements over existing fusion-based detection methods.

With Experiment 1 we showed that the road cannot assumed to be flat. On a flat road, the candidate region would have an offset up to 0.75 m, missing half of most road users. Moreover, in the case of speed bumps, the candidate region would miss a whole car or most of pedestrians and bicyclists, thus increasing the miss rate. With our dynamic candidate regions method, the offset is reduced to 0.2 m and added as a margin to the candidate region. We reduced processing time by having smaller and more effective candidate regions.

Our ConvNet with contrastive loss improved the performance with 1.4% over the conventional approach, as shown in experiment 2. Experiment 3 combined the dynamic candidate regions with our ConvNet which increased the performance with 7.6%.

The road user detection was benchmarked on an existing dataset, but in future we will present our own dataset based on the WEpod recordings also containing more road user types. Moreover, it will contain more data and different variations compared to other datasets. With more data we expect the performance gap to be closed even further.

Experiment 4 showed that we obtained a performance of 91.9% on road user detection for our WEpod vehicle driving on the public road. This performance is still below the human benchmark of 99% precision and 99.5% recall for a single image. However, by combining the classification of three successive images the performance is increased to 95.1%. Furthermore, we are much closer to the human benchmark and hence the WEpod can drive safely on the public road.

References

[1] D. Belgiovane and C.-C. Chen. Bicycles and human riders backscattering at 77 ghz for automotive radar. In *2016 10th EuCAP*, pages 1–5. IEEE, 2016.

[2] R. Benenson, M. Omran, J. Hosang, and B. Schiele. Ten years of pedestrian detection, what have we learned? In *ECCV*, pages 613–627. Springer, 2014.

[3] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012 IEEE Conference on, pages 3354–3361. IEEE, 2012.

[4] I. Goodfellow, Y. Bengio, and A. Courville. Deep learning. Book in preparation for MIT Press, 2016.

[5] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR, 2006 IEEE conference on*, volume 2, pages 1735–1742. IEEE, 2006.

[6] D. Held, J. Levinson, and S. Thrun. A probabilistic framework for car detection in images using context and scale. In *Robotics and Automation (ICRA), 2012 IEEE Conference on*, pages 1628–1634. IEEE, 2012.

[7] Itseez. Open source computer vision library, 2015.

[8] T. Kato, Y. Ninomiya, and I. Masaki. An obstacle detection method by fusion of radar and motion stereo. *IEEE Transactions on Intelligent Transportation Systems*, 3(3):182–188, 2002.

[9] S. Milch and M. Behrens. Pedestrian detection with radar and computer vision. 2001.

[10] L. Oliveira and U. Nunes. Pedestrian detection based on lidar-driven sliding window and relational parts-based detection. In *Intelligent Vehicles Symposium, 2013 IEEE*, pages 328–333. IEEE, 2013.

[11] W. H. Organization. *Global status report on road safety 2015*. World Health Organization, 2015.

[12] C. Premebida, J. Carreira, J. Batista, and U. Nunes. Pedestrian detection combining rgb and dense lidar data. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4112–4117. IEEE, 2014.

[13] C. Premebida and U. J. C. Nunes. Fusing lidar, camera and semantic information: A context-based approach for pedestrian detection. *The International Journal of Robotics Research*, page 0278364912470012, 2013.

[14] J. Schlosser, C. K. Chow, and Z. Kira. Fusing lidar and images for pedestrian detection using convolutional neural networks. In *2016 IEEE International Conference on Robotics and Automation*, pages 2198–2205. IEEE, 2016.

[15] S. Sivaraman and M. M. Trivedi. Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis. *IEEE Transactions on Intelligent Transportation Systems*, 14(4):1773–1795, 2013.

[16] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2012 IEEE Conference on, pages 1701–1708, 2014.

[17] N. Yamada, Y. Tanaka, and K. Nishikawa. Radar cross section for pedestrian in 76ghz band. In *2005 EMC*, volume 2, pages 4–pp. IEEE, 2005.

[18] K. Young, M. Regan, and M. Hammer. Driver distraction: A review of the literature. *Distracted driving*, pages 379–405, 2007.

[19] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele. How far are we from solving pedestrian detection? *arXiv preprint arXiv:1602.01237*, 2016.