ISSN 1213-6972 Journal of WSCG http://www.wscg.eu Vol.27, No.1, 2019

Curb Detection for a Pedestrian Assistance System Using End-to-End Learning

Hasham Shahid Qureshi
Technische Universität Berlin
Marchstr. 23
10587, Berlin
qureshi@tu-berlin.de

Rebecca Wizcorek
Technische Universität Berlin
Marchstr. 23
10587, Berlin
wiczorek@tu-berlin.de

Abstract

Our goal is to develop an assistance system for supporting road crossing among older pedestrians. In order to accomplish this, we propose detecting the curb stone from the pedestrians' point of view. Curb detection plays a significant role in road detection and obstacle avoidance, etc. However, it also presents significant challenges such as the small size of the target as well as, obstacles and different structures. To tackle these problems, we chose to fuse two sensors, a Camera and a Leddar, and use an algorithm that applies an end-to-end learning approach. The convolutional neural network was chosen to process the images acquired from the mono camera by filming the curb and its surroundings. The artificial neural network was selected to process the point cloud data of the Leddar acquired in the form of arrays from the 16 channels of the Leddar. A prototype was developed for data collection and testing purposes. It consists of a structure carrying both sensors mounted on a walker. The data from both sensors were collected with multiple factors taken into consideration, such as, weather, light conditions and, approaching angles. For the training of algorithms, an end-to-end learning approach was selected where we labelled the complete image or array rather than labelling the individual pixels or points in the data. The networks were trained and, the features from the parallel networks were concatenated and given as the input to the fully connected layers to train the complete network. The experimental results show an accuracy of more than 99% and robustness of the end-to-end learning approach. Both sensors are relatively inexpensive and are in fusion together, they are able to efficiently accomplish the task of detecting the curb stone from the pedestrians' point of view.

Keywords

Deep learning, curb detection, pedestrian assistance system, end-to-end learning, Leddar, monocamera, multi sensor data fusion, convolutional neural networks, artificial neural networks

1 INTRODUCTION

For people of all ages, out-of-home mobility is an indispensable part of leading an independent and self-determined life [Lim09a]. Mobility and participation in the society is a crucial part in keeping functionality and also to prevent the disability [mol04a]. This makes mobility of older pedestrians an important topic. Since the elderly demographic of the population (adults of 65 years and older) in the world is increasing, older pedestrians' mobility in the traffic environment requires special attention.

Older pedestrians were involved in 20% of all road-traffic accidents in Germany in 2013 [Bun13a]. They are involved in more accidents than middle-aged people when walking distance is taken into account [Ryt06a]. Moreover, in comparison to other age groups, older pedestrians require a longer recovery time after road-traffic accidents and their fatality rate is four times higher [Bun13a]. According to official police statistics in Berlin (Germany), most of these accidents happen at official crossings such as zebra crossings and traffic lights, and the main cause is the lack of attention older pedestrians pay to oncoming traffic [Bra15a].

These facts mean it is important to understand the underlying problems older pedestrians face in traffic in order to develop appropriate solutions to help prevent these accidents. In order to increase safety and support older pedestrians, in our project FANS (Fußgänger-Assistenzsystem für ältere Nutzerinnen und Nutzer im Straßenverkehr - Pedestrian Assistance System for Older Road Users) we are currently developing an assistance system. This work comprises a user-centred approach which includes the future target group in the design process of the prototype.

Our project initially involved investigating the reasons behind older pedestrians' lower attention to traffic. Two reasons were established. Firstly, older people tend to examine the terrain more frequently and attentively than younger people, which demands visual attention. This additional attention-demanding task impares their ability to detect hazards in the street environment [Wic16a]. Secondly, when approaching the road, most

people watch out for cars while walking, which can be viewed as multitasking behaviour. The act of walking requires cognitive resources and thus decreases the frequency at which visual targets such as passing cars are detected [Pro17a]. Based on the findings mentioned above and the analysis of official accident statistics [Sta13a], we can outline the requirements for the assistance system.

The purpose of the assistance system is to warn users when they are approaching the road. The warning reminds them to stop whatever else they are doing (i.e., walking, scanning the ground) and direct all their attention towards the traffic. This notification should be given to the users at a predefined distance in order to prevent them stepping onto the road without checking for traffic first.

The most important requirement for the system is for it to work as reliably as possible. That means reducing the number of events where the system fails to detect the curb stone (misses) to a minimum. However, the frequency of false alarms (occurring when the user is not close to a road) should also be kept fairly low. This is because the experience of false alarms can lead to the ignorance of warnings due to users' distrust of the system [Dix07a], [Mad06a], [Bli95a].

In order to generate appropriate warnings, the assistance system must be able to effectively identify when users are approaching a road. It was decided that this should done by detecting the curb stone using a suitable sensor solution. The sensors should not be too heavy to avoid imposing excessive weight upon the users, as well as not being too expensive for the older target group to afford.

2 RELATED WORK

Curb stone detection is an important research aspect in the field of mobile robotics and is especially important in the field of autonomous vehicles. It is a crucial component in ADAS (Advanced Driver Assistance Systems) such as parking assistance, vehicle positioning, etc. However, this research focuses on curb detection from the perspective of the driver (i.e., the car). Research into curb detection from the point of view of pedestrians is relatively rare. This is because curb detection from a pedestrians' perspective proposes a separate relevance and, in particular, a pedestrian's angle of view is entirely different. However, we can still extract some useful information regarding the sensors, as well as theories proposed to solve this problem, from recent research into mobile robotics and intelligent vehicle systems.

For curb detection, methods vary with regard to types of sensors and processing methods, which all have several advantages and drawbacks. It can be categorized based on the types of sensors used. For example, standard approaches using an inexpensive mono camera exploit the methods based on appearance information (i.e., image processing) [Pri16a]. Image-processing-based techniques can allow detection from long distances, but they are susceptible to decreased accuracy which can be caused by changes in the intensity of images such as shadows or changes in road surfaces, road markings, etc.

However, most methods rely on the 3D information extracted from LiDAR (Light Detection and Ranging) or imaging sensors. As opposed to monocular cameras, stereo vision cameras can exploit 3D geometry and are therefore more suited to detecting curbs [Kel15a], [Sod16a], [Fer14a], [Sei13a], [Enz13a], [Oni11a], [Hu12a], [Sie10a]. Stereo vision can provide high-resolution information which is not available in other 3D sensors. Since they provide a high resolution, appearance and geometry features are used actively to detect curbs using stereo vision. The geometry features, such as the height step [Kel15a], curvature [Sod16a], [Fer14a] and height variation [Kel14a], [Sei13a], [Hu12a] are commonly used with stereo-vision-based methods to detect curbs. These methods are relatively efficient, however the sensors used in these techniques are comparatively expensive and a 3-D sensor needs a 360° view which contradicts with our requirements.

Several mapping methods are used for curb detection. Digital Elevation Models (DEM) are the most widely used [Oni08a], [Kel14a], [Sie11a], [Enz13b]. All of these approaches can augment noisy sensor data through local or temporal filtering. However, they suffer the drawback that the cell sizes affect the accuracy of road-curb features. Therefore, small cell sizes are favoured, which tend to require higher computational efforts, such as higher memory consumption, making them difficult or even impossible to use in real time.

Considering the requirements of our assistance system, we decided to carry out the sensor fusion using the deep learning method. Therefore, we used a mono camera and a range sensor, assuming that the fusion of these two sensors could detect the curb more efficiently. In our case, the use of an expensive multi-layer LiDAR, which requires a 360-degree field of view, was not a feasible option. Hence, we decided to use a Leddar sensor. Leddar [Oli15a] is a propriety sensor from Leddartech which works based on the principles of LiDAR technology. It can detect, locate and measure objects in its field of view. These sensors are mounted on a walker (see section 3.2) to avoid older people having to carry the assistance system on their body.

3 MULTI-SENSOR DATA FUSION USING END-TO-END LEARNING

We started our work by implementing the detection system with only one sensor. An end-to-end learning approach using Convolutional Neural Network (CNN) was chosen to detect the road and its surroundings from the pedestrian's point of view via a mono camera. This work was inspired by the work of Bojarski et. al. [Boj16a]. The authors implemented an end-to-end learning approach using a CNN in the context of autonomous driving. If a network is used in the context of end-to-end learning, it learns the whole processing pipeline without the need to label explicit parts of the data. For example, in the case of the image dataset, it is sufficient to label the whole image rather than labelling the individual pixels in the image, which saves considerable time during the annotation of the data (for further details, see [Qur18a]). The camera which has been used has the focal length of $4.0mm$ with the optical resolution of 1280×960 and has the maximum frame rate of 30fps @ 640×480.

In order to further improve detection accuracy, we integrated the Leddar as the second sensor in the system. The Leddar sensor is based on the optical time-of-flight technology which sends short light pulses about 10,000 times per second to actively illuminate the desired area. The sensors then capture the light that is scattered back from objects and processes the signals to accurately determine their location and other attributes, such as shape and design. In our project, we are using the Leddar M16, which is a 16-segment solid-state LiDAR sensor module. The Leddar M16 sensor module uses 16 independent detection channels to deliver continuous and precise detection combined with exceptional lateral discrimination. It has a detection range of 146m and a data acquisition rate of up to 100 Hz [Oli15a].

The Leddar has been mounted on a walker so that the channels are facing the curb stone vertically with an approximate angle of 45° in the predefined distance of $2m \pm 1m$. The schematics can be seen in Figure. 1.

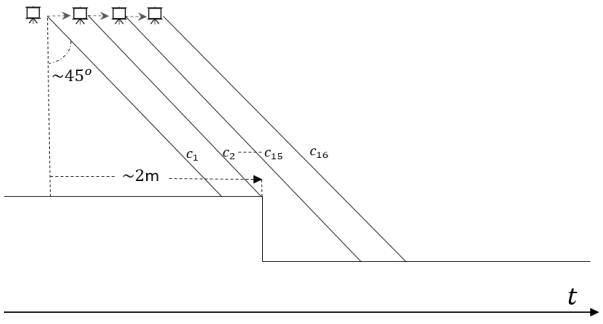


Figure 1: Schematics of Leddar sensor

In Figure 1, "c" represents the channel of the Leddar. If the Leddar is moved towards the curb stone and it approaches the predefined window, as there is a difference in height between the pavement and the road, there will be a change in value for one channel (channel 16) as the lateral distance between the Leddar and the deflecting surface increases. This means that, for channel number 16, at t = 0, the deflecting surface is the pavement and, at t = 1, the deflecting surface is the road. Therefore, a profile can be made where this channel indicates the difference, which provides an array of data in the shape $(16,1)$. Similarly, a profile can be made for each channel.

3.1 Proposed Methodology

For the camera images, we used the input plane of $(227,227,3)$, that means RGB images were used to train the CNN architecture in the first step. However, it is impossible to use the same CNN network to fuse both sensors because the dimensionality of the input data differs. Therefore, we decided to use two different algorithms to process the data of each sensor separately and merge them later at the classification stage. For the images, CNN was used as it provides the best results for learning the hidden patterns in images. In order to cater the effect of light in the RGB images, we decided to use the grey-scale images to train the CNN. In this way the algorithm neglects the effects of light (e.g. brightness, shadowing, etc.), especially in sunny conditions and this led to a better prediction. We also reduced the pixel size of the images to decrease the training time of the algorithm. This led to an input size of resolution $(225,225)$. However, Leddar information is comprised of point cloud data acquired from the 16 channels, with each channel representing one data point. Thus, the Artificial Neural Networks (ANN) were chosen because ANNs do not reduce the dimensionality of the input data, meaning no relevant information is lost. Moreover, ANNs are easy to fuse with CNNs and doesn't hinder the speed of the network.

To fuse the two sensors which have heterogeneous input streams, we trained two networks in parallel, before concatenating the features from both networks. These concatenated features were used as an input for the fully connected layers to establish the symmetry between both sensors. After this, the complete network was trained to tune the hyper parameters. These fully connected layers also serve the purpose of classification, however, in end-to-end learning, it is hard to determine which layers perform the feature-extraction task and which layers carry out the classification.

3.2 Data Collection

The data was collected using a prototype consisting of the following modules:

- Walker (Invacare Banjo P452E/3)
- Leddar M16

- USB Webcam HD C270 from Logitech
- Notebook (Lenovo Y720-15IKB)

Figure 2: Prototype used for data collection and testing

A customised structure was added to the walker to carry the sensors, as shown in Figure 2. A housing for the camera was fabricated using a 3D printer. The Leddar housing was integrated with a rechargeable power supply. Both of the sensors were connected to the notebook using the USB interface.

3.3 Camera dataset

In order to train our CNN network, a dataset was needed to represent the task at hand (curb stone detection from a pedestrian's point of view). As there was no prior dataset of this kind available, a new dataset had to be constructed from scratch. This dataset took into account different light and weather conditions, as the majority of the relevant accidents occur in conditions with reduced visibility [Sta13a], as well as different road environments.

Because the system is developed for use in the city of Berlin, the dataset incorporates the different types of pavements and curb stones found in the streets of Berlin. In order to achieve this, it was necessary to carry out an analysis of the existing pavement structures and how common they are. For further details, see [Qur18a].

In order to train the network, the dataset was divided into two classes labelled "positive" and "negative". Images labelled as "positive" show scenarios where users walk towards the road, whereas "negative" images depict scenarios where users are walking along the pavement parallel to the road, as shown in Figures 3 and Figure 4 respectively.

Figure 3: A few examples of the positive labelled images

Figure 4: A few examples of the negative labelled images

3.4 Leddar dataset

Since this problem was tackled as a binary classification problem, two types of data were also collected for the Leddar data, namely "positive" and "negative". The "positive" class relates to situations where the Leddar is facing the curb and the user intends to cross the road. For the "negative" class, data was collected about situations the Leddar was not facing the curb and the person had no intention of crossing the road. We used 16 channels of the Leddar which generate the point cloud data. Data was collected from each channel and a profile was made which gave us an array of size $(16, 1)$. Each array represents one data sample. The data was collected on the streets of Berlin with the help of the walker mentioned in section 3.2. While collecting the data, multiple aspects were considered (e.g. height of curb stone, angle of approach, distance from curb stone, parked cars, etc.).

3.5 Training Of Algorithm

3.5.1 Selection Of The Data

In order to train the network, it was important to select the adequate data from both of the sensors. The data

from the Leddar and the camera were collected separately, which means there is no connection between the data streams of the Leddar and the camera. We chose 40,000 frames from each sensor. The data from each sensor was divided into positive and negative databases at a 50:50 ratio. The dataset was then further divided into two parts: training and validation, at 70% and 30% respectively. The data from the camera was augmented using data augmentation techniques by adding rotations, artificial shifts, zooming and sheer effect, etc. to overcome the overfitting problems and also to teach the network the various conditions which present themselves in real-life situations (e.g. insufficient illumination, motion blur, etc.). However, the data from the Leddar was not augmented as each channel gives a number and adding artificial noise entirely changed the outcome.

3.5.2 Network Architecture

After the selection of data, experiments were conducted to find the best suitable network architecture. The training was started with the simplest case and the difficulty was increased gradually to monitor the performance of the network. In the beginning, only datasets without obstacles and objects were introduced as a positive category. Afterwards, the difficulty was increased by adding different factors such as leaves, obstacles and various weather conditions. Similarly, for the "negative" class, the difficulty level was increased by adding a range of different pavements and angles.

After extensive experimentation with the aim of achieving maximum accuracy, the network consisted of the following configuration. The final CNN architecture contained 5 layers with strided convolutions in all the convolution layers with a size of 2×2 and with a kernel size of 3×3. The ANN architecture had 4 layers with a varying number of neurons. Three fully connected layers are used after concatenating the features from CNN and ANN networks. These layers were then trained on the combined features from both networks (i.e. CNN and ANN). The complete network architecture can be seen in Figure 5.

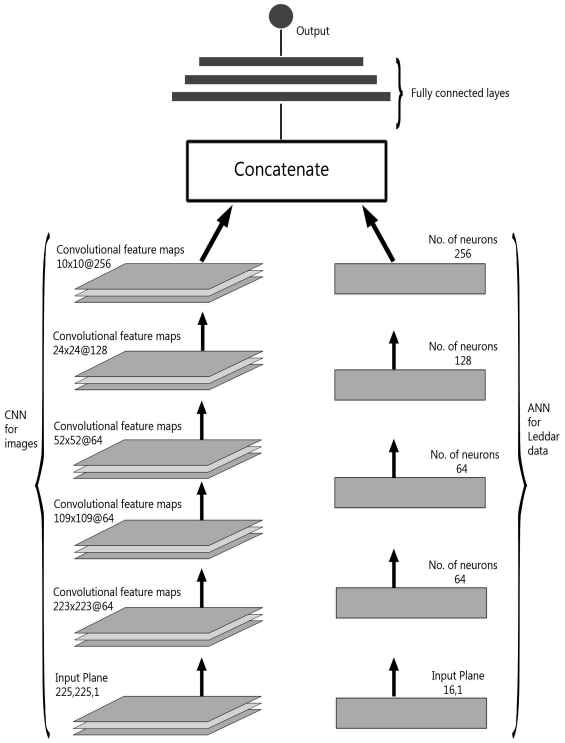


Figure 5: Network architecture

4 RESULTS

The efficacy of the model was determined through the accuracy and loss of training and validation. These values indicated the system's ability to learn the underlying features of the data. A validation accuracy of 99.04% and a validation loss of 0.043 were observed. The epoch by epoch analysis is shown in Figure 6. Another way to observe the efficacy of the model is the through confusion matrix. This confusion matrix demonstrates how many times the algorithm was not able to predict the correct label. The confusion matrix was plotted with 10,000 test samples and can be seen in Figure 7.

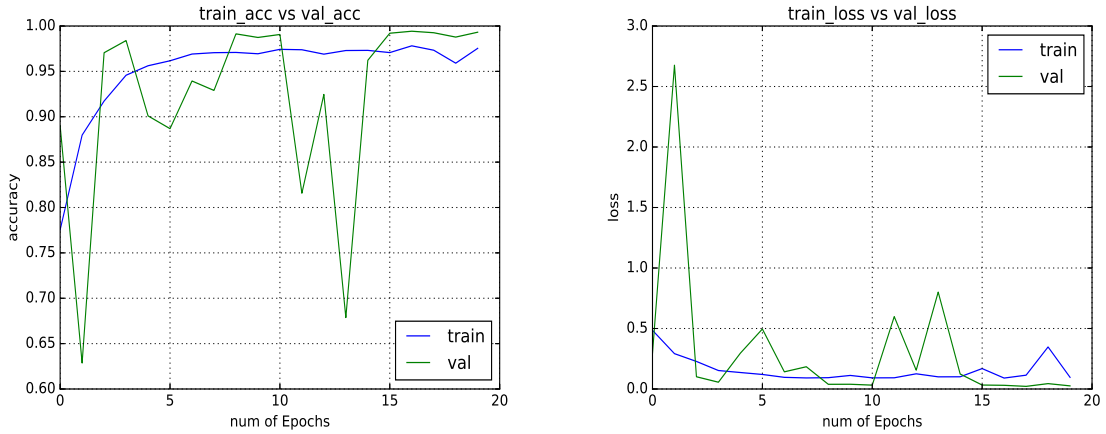
The Lenovo Y720-15IK notebook, which has NVIDIA GTX 1050 GPU, was used for real-time testing. The system runs at 22 Frames Per Second (FPS). In order to assess the performance of the system, we adopted the F1 score as an evaluation criterion for curb detection, which is calculated using precision and recall. The results are listed in Table 1.

Labels	Precision	Recall	F1 score
Positive	1.00	0.99	0.99
Negative	0.99	1.00	0.99

Table 1: Precision, recall and F1 score of the system

5 CONCLUSION

Based on the analysis of the target group, we deduced that the sensors chosen to detect the curb stone should be lightweight and inexpensive. In order to train the networks, the datasets were constructed from scratch, taking into account various factors such as light, weather and structural combinations. These datasets will be extended in future in order to account for a wider range of scenarios. Both of these datasets can be used as independent entities in other applications and systems and are therefore valuable irrespective of



(a) Accuracy graph (b) Loss graph

Figure 6: Simulation results for the network

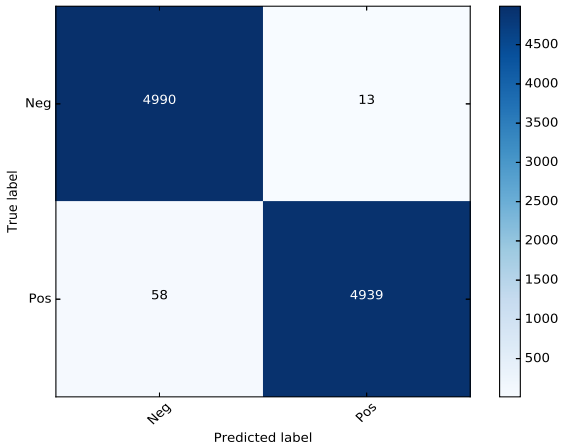


Figure 7: Confusion matrix for 10,000 test images

the algorithm. Additionally, a novel approach for the multi-sensor data fusion using the end-to-end learning technique is presented where two algorithms, CNN and ANN, were used to efficiently detect the curb stone in various scenarios. The fusion network was trained on a small amount of data, which in turn requires less training time. However, despite the minimal amount of data, the network was able to generalise well and detected the curb stone with an efficiency of more than 99%. The system worked reliably in different conditions, with very little time required for prediction. We believe that end-to-end learning can be effective for multi-sensor data fusion. This technique proved to be very fast and reliable as there was no need for hand-crafted rules or labels for the data, saving a tremendous amount of time. By fusing CNN and ANN using end-to-end learning algorithm, we proved that end-to-end learning is also capable of handling multiple algorithms at once. In future we will train the different algorithms to compare the accuracy and also observe whether end-to-end learning is able to handle more complex architectures. Finally, a field study will be conducted with the target group to evaluate the performance of the systems. We will also investigate the performance of the users in the detection of hazards as well as their acceptance and trust of the system.

6 REFERENCES

[Ryt06a] Rytz, M., .Senioren und Verkehrssicherheit. VCS Verkehrs-Club der Schweiz, Bern, 2006.

[Bun13a] Bundesamt, S., Unfallentwicklung auf Deutschen Strassen 2012: Begleitmaterial zur Pressekonferenz am 10. Juli 2012 in Berlin, 2013.

[Bra15a] S. A. B. Brandenburg, Verkehrsstatistiken. 2015.

[Wic16a] Wiczorek, R., Siegmann, J., and Breitinger, F. Investigating the impact of attentional declines on road-crossing strategies of older pedestrians. in Proceedings of the Human Factors and Ergonomics Society Europe, pp. 155-169, 2016.

[Pro17a] Protzak, J. and Wiczorek, R. On the Influence of Walking on Hazard Detection for Prospective User-Centered Design of an Assistance System for Older Pedestrians., i-com, vol. 16, no. 2, pp. 87-98, 2017.

[Dix07a] Dixon, S. R., Wickens, C. D., and McCarley, J. S. On the independence of compliance and reliance:Are automation false alarms worse than misses?, Human Factors, vol. 49, no. 4, pp. 564-572, 2007.

[Mad06a] Madhavan, P., Wiegmann, D. A., and Lacson, F. C. Automation failures on tasks easily performed by operators undermine trust in automated aids., Human Factors, vol. 48, no. 2, pp. 241-256, 2006.

[Bli95a] Bliss, J. P., Gilson, R. D., and Deaton, J. E. Human probability matching behaviour in

response to alarms of varying reliability., Ergonomics, vol. 38, no. 11, pp. 2300-2312, 1995.

[Kel15a] Kellner, M., Hofmann, U., Bouzouraa, M. E., and Stephan, N. Multi-cue, model-based detection and mapping of road curb features using stereo vision. in 2015 IEEE 18th International Conference on Intelligent Transportation Systems, pp. 1221-1228, 2015.

[Sod16a] Sodhi, D., Upadhyay, S., Bhatt, D., Krishna, K. M., and Swarup, S. Crf based method for curb detection using semantic cues and stereo depth. in Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing, p. 41, 2016.

[Fer14a] Fernandez, C., Izquierdo, R., Llorca, D. F., and Sotelo, M. A. Road curb and lanes detection for autonomous driving on urban scenarios. in IEEE 17th International Conference on Intelligent Transportation Systems, pp. 1964-1969, 2014.

[Kel14a] Kellner, M., Bouzouraa, M. E., and Hofmann, U. Road curb detection based on different elevation mapping techniques. in IEEE Intelligent Vehicles Symposium, pp. 1217-1224, 2014.

[Enz13a] Enzweiler, M., Greiner, P., Knoppel, C., and Franke, U. Towards multi-cue urban curb recognition. in IEEE Intelligent Vehicles Symposium, pp. 902-907, 2013.

[Sei13a] Seibert, A., Hahnel, M., Tewes, A., and Rojas, R. Camera based detection and classification of soft shoulders, curbs and guardrails. in IEEE Intelligent Vehicles Symposium, pp. 853-858, 2013.

[Sie11a] Siegemund, J., Franke, U., and Förstner, W. A temporal filter approach for detection and reconstruction of curbs and road surfaces based on Conditional Random Fields. in IEEE Intelligent Vehicles Symposium, pp. 637-642, 2011.

[Oni11a] Oniga, F. and Nedevschi, S. Curb detection for driving assistance systems:A cubic spline-based approach. in IEEE Intelligent Vehicles Symposium, pp. 945-950, 2011.

[Hu12a] Hu, T. and Wu, T. Roadside curb detection based on fusing stereo vision and mono vision. in Fourth International Conference on Machine Vision (ICMV 2011): Computer Vision and Image Analysis; Pattern Recognition and Basic Technologies, 83501H, 2012.

[Sie10a] Siegemund, J., Pfeiffer, D., Franke, U., and Förstner, W. Curb reconstruction using conditional random fields. in 2010 IEEE Intelligent Vehicles Symposium, pp. 203-210, 2010.

[Oli15a] Olivier, P. Leddar optical time-of-flight sensing technology: A new approach to detection and ranging:A new approach to detection and ranging., 2015.

[Boj16a] Bojarski, M. et al. End to End Learning for Self-Driving Cars., CoRR, vol. abs/1604.07316, 2016.

[Qur18a] Qureshi, H. S., Glasmachers, T., and Wiczorek, R. User-Centered Development of a Pedestrian Assistance System Using End-to-End Learning. in 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 808-813, 2018.

[Oni08a] Oniga, F., Nedevschi, S., and Meinecke, M. M. Curb detection based on a multi-frame persistence map for urban driving scenarios. in 2008 11th International IEEE Conference on Intelligent Transportation Systems, pp. 67-72, 2008.

[Enz13b] Enzweiler, M., Greiner, P., Knoppel, C., and Franke, U. Towards multi-cue urban curb recognition. in IEEE Intelligent Vehicles Symposium, pp. 902-907, 2013.

[Sta13a] Stab Des Polizeipräsidenten. (2013) Sonderutnersuchung fußgängerverkehrsunfälle in berlin 2013.

[Lim09a] Limbourg, M. and Matern, S., Erleben, Verhalten und Sicherheit älterer Menschen im Strassenverkehr: Eine qualitative und quantitative Untersuchung (MOBIAL). Köln: TÜV Media, 2009.

[Pri16a] Prinet, V., Wang, J., Lee, J., and Wettergreen, D. 3D road curb extraction from image sequence for automobile parking assist system. in 2016 IEEE International Conference on Image Processing (ICIP), pp. 3847-3851, 2016.

[mol04a] Mollenkopf, H. et al. Social and behavioural science perspectives on out-of-home mobility in later life:Findings from the European project MOBILATE., European Journal of Ageing, vol. 1, no. 1, pp. 45-53, 2004.