Ojala, Risto; Vepsäläinen, Jari; Hanhirova, Jussi; Hirvisalo, Vesa; Tammi, Kari

Novel Convolutional Neural Network-Based Roadside Unit for Accurate Pedestrian Localisation

Published in:
IEEE Transactions on Intelligent Transportation Systems

DOI:
10.1109/TITS.2019.2932802

Published: 01/09/2020

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Ojala, R., Vepsäläinen, J., Hanhirova, J., Hirvisalo, V., & Tammi, K. (2020). Novel Convolutional Neural Network-Based Roadside Unit for Accurate Pedestrian Localisation. *IEEE Transactions on Intelligent Transportation Systems*, *21*(9), 3756-3765. Article 8793228. https://doi.org/10.1109/TITS.2019.2932802

Novel Convolutional Neural Network-based Roadside Unit for Accurate Pedestrian Localisation

Risto Ojala, Jari Vepsäläinen, Jussi Hanhirova, Vesa Hirvisalo and Kari Tammi, *Member, IEEE*

***Abstract*—Hazardous situations may easily be caused by limited visibility at urban traffic intersections due to buildings, fences, flora and other obstacles. Thus, drivers approaching an intersection have limited reaction time when other obscured road users, such as pedestrians and cyclists, appear unexpectedly. Previous research has been conducted on applications warning drivers of approaching out-of-sight vehicles. However, less focus has been on the detection and awareness applications revealing the presence of pedestrians. We propose a novel system that displays the driver real-time locations and types of hidden road users at traffic intersections. A roadside unit is installed in the infrastructure which sends safety-critical object data to the vehicle, supporting the real-time decision-making of the driver. The roadside unit consists of a monovision camera streaming video to a computing unit which performs object detection and distance measurements on the detected objects. This paper validates the capability of the proposed system of localising a pedestrian, and also examines its sensitivity to installation and detection errors. The results show that the accuracy of the proposed system is suitable for the intended application. However, an error in the vertical angle of the roadside unit camera caused an exponential error in the distance approximation in respect to the measured distance. The detection accuracy was noticed to decrease at long distances and in dark surroundings. Moreover, in order to reduce the effect of the presented errors, the camera should be installed as high as possible without hindering its detection capabilities.**

***Index Terms*—Cameras, intelligent transportation systems, machine vision, neural networks, object detection, vehicle safety.**

I. Introduction

Traffic intersections are hazardous areas as pedestrians and cyclists intermingle with vehicle traffic. These different types of road users must often cross each other's trajectories, leading to potentially dangerous situations. Even if common traffic rules are followed, collisions can occur, and the consequences can be severe. One fifth of all fatal traffic accidents in the U.S. occur in traffic intersections alone [1]. A similar percentage of road fatalities is reported to occur in traffic junctions in the EU [2]. These fatalities are partly caused by the surroundings of intersections being obscured by obstacles, such as buildings, fences and flora, in many densely populated regions. Naturally, limited vision reduces the reaction time of road users in rapidly changing situations. In addition, disturbances in the perception of a vehicle driver, such as difficult weather conditions, electronic devices and fatigue, can lead to mishaps and accidents. Thus, by increasing the reaction time and enhancing the perception of the driver at intersections, the safety of all road users could be significantly increased.

Therefore, any information on hidden or obscured traffic users and their locations would provide drivers with more time to adjust to the immediate situation, thus helping them to interpret the intentions of others. Drivers would be able to prepare to give way to approaching road users before they come into view, avoiding possible collisions. This is especially important when vulnerable road users such as pedestrians and cyclists are present, since they can quickly and unexpectedly appear in the view and attempt to cross the road. Additionally, drivers with the right of way could also more easily identify situations where the rules are about to be violated. One method for acquiring information of obscured subjects is real-time object detection. Information gathered by a roadside unit (RSU) located at the intersection can be transmitted utilising a wireless vehicle-to-infrastructure (V2I) communication.

V2I is an area of extensive research. Traditionally, research on V2I applications has mainly focused on acquiring statistical data from vehicles [3][4][5]. These data are beneficial for traffic design, traffic control and road maintenance applications. V2I has also been applied in safety applications [6][7][8][9]. A typical approach for such applications is an RSU that monitors traffic and warns drivers of possibly dangerous situations. In addition, vehicle-to-vehicle (V2V) communications have been investigated [10][11] which allow the sharing of safety-related information between vehicles; however, this information does not typically include the positions of other traffic users, such as pedestrians, cyclists and unconnected vehicles.

This paper considers a novel approach as illustrated in Fig. 1 where the information is delivered from the infrastructure to the

Submitted for review on 07.05.2019. This work was supported by the Henry Ford Foundation Finland.

R. Ojala is with the Department of Mechanical Engineering, Aalto University, Espoo, 02150 Finland (e-mail: risto.j.ojala@aalto.fi).

J. Vepsäläinen is with the Aalto University Department of Mechanical Engineering, Espoo, 02150 Finland (e-mail: jari.vepsalainen@aalto.fi).

J. Hanhirova is with the Department of Computer Science, Aalto University, Espoo, 02150 Finland (e-mail: jussi.hanhirova@aalto.fi).

V. Hirvisalo is with the Department of Computer Science, Aalto University, Espoo, 02150 Finland (e-mail: vesa.hirvisalo@aalto.fi).

K. Tammi is with the Department of Mechanical Engineering, Aalto University, Espoo, 02150 Finland (e-mail: kari.tammi@aalto.fi).

vehicle. In our approach, a V2I connection is used to acquire machine vision data of objects which are located out of the driver's line of sight. The safety-critical object data are sent from the RSU to the vehicle and transformed to the vehicle's coordinate frame utilising its navigation data. Hidden objects are classified and highlighted for the driver on a screen to support their real-time decision-making. Object detection technology is common in modern Advanced Driver Assistance Systems (ADASs) which offer considerable potential for increasing traffic safety [12]. However, their detection capabilities are usually limited to the sensors mounted on the car, whereas the proposed solution takes advantage of the different point of view provided by the RSU. In this paper, we study the pedestrian localisation accuracy of the proposed RSU and analyse its sensitivity to different installation and detection errors.

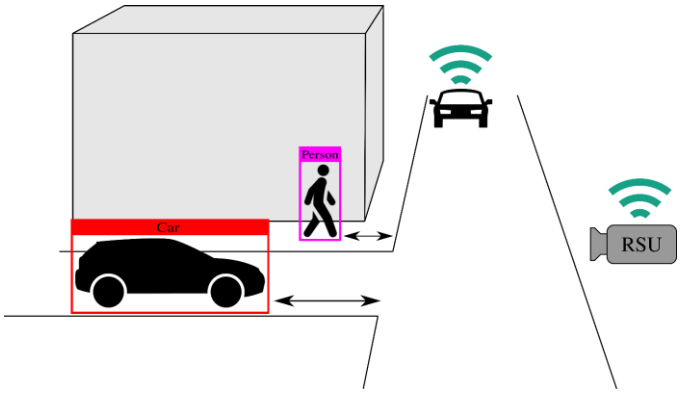

Fig. 1. Basic concept of the proposed system.

II. State-of-the-art

V2I and V2V are popular fields of study as they form the core of an intelligent transportation system (ITS). V2V applications have been thoroughly researched for safety purposes, and there have been multiple patents [13][14][15] on such applications in the recent years. Intersection safety has been studied by Ibanez-Guzman et al. [10], who developed a V2V solution that warned drivers of upcoming vehicles. The performance of such systems at urban intersections was examined by Rashdan et al. [11]. They measured the amount of losses in short-range communication between two vehicles, caused by obstacles such as buildings between them. Among safety applications V2V has presented possibilities in traffic monitoring [16] and control [17][18], which are also common areas of V2I research.

V2I has proved to be effective in monitoring and controlling traffic, which are important means for reducing travel times and fuel consumption. In order to monitor traffic congestions, Barrachina et al. [3] developed a V2I-method for estimating traffic densities in a certain area with a 3.04 % accuracy, which is valuable information for traffic control applications. Optimized traffic flow was researched by Cai et al. [4], who simulated adaptive traffic signals that operated on travel-time approximations computed from vehicle positions. Compared to optimized fixed-time traffic signals, their model saved up to 11 % of an intersection's travel time in a high traffic scenario. Similar traffic flow case was studied by Ubiergo et al. [5], whose method was to give each vehicle approaching the traffic signals an individual speed limit. Their simulations achieved a 15 % reduction in traffic delays and 8 % savings in fuel consumption. In addition to successful research in these areas, V2I has also been investigated for safety applications.

Different V2I-based solutions have been proposed for collision detection systems aimed at increasing intersection safety. A system for preventing crashes and optimizing traffic flow in unsignalised intersections has been implemented by Milanés et al. [6]. Their approach was to determine all vehicles approaching the intersection, and then suggest target speeds to the drivers in order to allow safe passage. Another infrastructure-based collision warning system has been proposed by Basma et al. [7]. In their solution, magnetic sensors detected positions and speeds of vehicles approaching an intersection, and drivers were alerted if a collision seemed probable. Collision prediction has also been researched with a machine vision camera by Atev et al. [8]. Using a roadside camera, they were able to estimate the sizes and positions of visible vehicles and predict collisions between them over a short time interval. Their method for detecting vehicles was based on an assumption of a static background, so that areas that differed from the reference background were detected as targets. In addition to systems anticipating collisions between vehicles, there has been research on preventing collisions involving pedestrians. Artail et al. [9] have proposed an RSU that could inform drivers of upcoming pedestrians, localising them by requesting the positions of nearby cell phones from the mobile network. Applications such as these require equipment to be installed in the road infrastructure, and therefore co-operation with local transportation authorities is necessary.

Governments around the world have participated in the development of V2I and V2V solutions. MEC-View, an on-going ITS project funded by German Federal Ministry for Economic Affairs and Energy, evaluates improving automated driving utilising infrastructure-based sensor information [19]. Furthermore, European Commission has co-funded an ITS research project SAFESPOT [20]. Among other accomplishments, the project has developed RSUs capable of dynamically mapping surrounding areas with equipment such as laser sensors and cameras [21]. One of their applications using these data is "Intelligent Cooperative Intersection Safety System" (IRIS). The system is capable of warning drivers about possible red-light violations, crossing pedestrians and cyclists when turning right over a crosswalk, and approaching vehicles when taking a left turn [22]. United States Department of Transportation has funded research on ITSs as well. Among concepts meant for similar purposes as those of SAFESPOT, they have also developed a concept for "Stop Sign Gap Assist" (SSGA) [23]. SSGA helps drivers coming from a minor road through a stop-controlled intersection by informing them if a car is approaching on the main road. These government efforts indicate that intersection safety is a relevant field of research, and there exists a legitimate demand for related applications.

V2V, V2I and other traffic applications apply different methods for localising road users. Often these applications are designed to take advantage of GPS data provided by on-board units (OBUs), as high quality GPS receivers can obtain position information at an accuracy of a few metres [24]. However, GPS accuracy suffers from signal blockage caused by large objects

such buildings and trees. Other measurement devices can also be combined with GPS in order to create a more reliable system. A GPS/Dead Reckoning system consisting of a GPS receiver, a gyroscope and an odometer has been studied by Lee et al. [25]. Their tests carried out with an autonomous vehicle at an experimental site resulted in an RMSE of 2.11 metres in the longitudinal direction. Khattab et al. [26] have researched a GPS-free vehicle positioning RSU that used dedicated short-range communications (DSRC) for two-way time of arrival calculations, combined with an on-board inertial measuring unit. The system was evaluated with simulations, and it was able to provide the distance along the road with an average error of slightly below 2 metres. Another common alternative to GPS localising are vision-based solutions.

Different stereo vision and monovision approaches have been researched for obtaining distances to objects of interest in traffic scenarios. A stereo vision system for obstacle positioning has been studied by Nedevschi et al. [27]. Their algorithm created a 3D reconstruction of the surrounding scene, classifying grouped points as objects. They found the system to be highly accurate, with measurement errors lower than 10 cm at 10 m and approximately 30 cm at 45 m. Ibarra-Arenado et al. [28] have proposed a method for on-board monovision distance measurements, that was based on detecting vehicle license plates. As the size of a license plate is standardised, it can be used to derive the distance to a vehicle. They validated the method by conducting tests among urban traffic. With measuring distances between 0.5 and 10 metres, the highest relative error of their system on a cloudy day was 2 %. A typical roadside surveillance camera has been tested for monovision positioning of vehicles in a road environment in the SAFESPOT project [21][29]. Their experiments managed a mean error of 1.53 metres across all trials. The system recognised vehicles with a motion detection algorithm and computed the distance to them using a geometric transformation from the image plane to the ground plane. An image-to-ground-plane transformation was also used in the work of Atev et al. [8]. We aim to continue their research by using a similar method.

We present an in-depth study of the localisation competency of a monovision-based RSU. Our algorithm for computing the distance of detected objects is based on a geometric transformation that has previously been used in monovision systems. However, our approach takes advantage of a modern real-time object detection software capable of detecting all types of road users. Our aim is to combine previous methods into a practical approach, instead of presenting original technical means. The ultimate goal is to take a step in a new direction in the field of ITSs by developing and studying a platform, where alternative to a warning, drivers receive the real-time locations of other vehicles as well as cyclists and pedestrians when they are out of sight. The challenges and advantages of an object detection based RSU will also provide a viewpoint to the development and deployment of object detection software.

III. Methods

A. Machine vision

Commonly used methods for detecting road users include laser scanners [30][31], radars [32][33] and vision-based systems [34][35]. Laser scanners and radars offer accurate distance measurements along with their detection capabilities, but these systems tend to be better fit for detecting obstacles rather than detecting and classifying road users. Vision-based systems typically have a high detection rate for all road users in favourable conditions, as images contain plenty of information to perform detection on. Therefore, a vision-based solution was chosen for the application.

Stereo vision is a common approach for photographic distance measurements, yet monovision is applied for lighter computing and simpler installation. Hence, the distance approximations are based on assumptions of the environment and optic equations of the camera lens. Monovision distance measurements have previously provided accurate results in traffic implementations [28][29]. A stereo vision system could possibly be more accurate for the intended application, yet it would also be more expensive. A second camera would increase the cost of the RSU and make the installation more demanding, as the cameras' relative geometry would have to be calibrated. Off-the-shelf stereo vision systems are often compact and therefore have limited range. Most importantly, a stereo vision would require significantly more computational power due to either performing object detection on an additional camera or locating a detected object in the other camera view by mapping the corresponding pixels.

B. Convolutional Neural Networks

Real-time object detection was achieved with a Convolutional Neural Network (CNN). CNNs consist of input and output layers, and multiple hidden layers that process input data to output information. Object detection CNN models take an image as input and output detected object class and bounding box information. Architecture and inner details of layers vary between different CNN models. CNN computation is guided by model weights learned iteratively using annotated training data, which defines what a CNN model is able to detect. CNN models can be optimized and fine-tuned for specific use cases [36].

Operation of a CNN is typically computationally demanding and trade-offs between speed and accuracy must often be considered [37]. One-stage detectors, such as YOLOv3 [38], [39] and RetinaNet [40], offer different trade-offs between speed and accuracy as Regional Convolutional Neural Networks (R-CNNs) [41][42][43], which first apply a CNN to find regions with probable objects, and then apply a second CNN to analyse these regions.

One-stage detectors are considered for the RSU due to their real-time performance with modern GPUs. Their speed and accuracy properties are typically modified by adjusting the size of the input layer, which affects the amount of information the CNN analyses from an image [37].

C. Measurement system

The measurement system was built using off-the-shelf components. Main hardware components were a GPU equipped computer and a USB camera. Software components included a

YOLOv3 CNN model executed on top of the Darknet [44] runtime, and a distance measurement algorithm. YOLOv3 is in this paper abbreviated to YOLO for simplicity. During operation, YOLO received input images from the USB camera, and output detected object class and bounding-box information. This information was processed by the distance measurement algorithm, which output the estimated distance to the target.

Specifications for the used computer, camera and version of YOLO are provided in Table I, Table II and Table III, respectively. A consumer grade camera was chosen as it includes on-board video compression, which is important for the RSU to reduce the amount of computational power that has to be installed alongside the camera. Additionally, the camera emphasises that the concept is not dependent on expensive vision equipment. Proper image lighting was ensured by using the automatic exposure mode of the camera.

TABLE I
COMPUTER SPECIFICATIONS

GPU	Nvidia Geforce GTX 1080 Ti 11 GB
CPU	Intel Core i7-8700K 3.7 GHz
RAM	2 x 16 GB 2400 MHz DDR4

TABLE II
CAMERA SPECIFICATIONS

Camera model	Logitech C920 HD Pro
Documented focal length	3.67 mm
Diagonal field of view	78°
Capture resolution	1920 x 1080
Auto-exposure	On

TABLE III
YOLO MODEL SPECIFICATIONS

YOLO version	3
Input size	1024 x 1024
Frame rate	8 frames per second (FPS)
Framework	Darknet
Training dataset	COCO [47]

A large input size was used for YOLO to achieve higher object detection accuracy. Higher bounding-box accuracy allows examining the maximal distance estimation potential of the system. However, this accuracy is acquired at the expense of detection throughput. In the intended application, a suitable balance with distance estimation accuracy and throughput needs to be determined. The suitable balance is affected by intersection specific factors such as the speed limit in the area, lighting conditions, and average number of road users.

The basic principle of YOLO object detection is depicted in Fig. 2. YOLO meshes the given image into an S x S grid, which is proportional to the input size defined by the user. Each cell in the grid is responsible for detecting an object that has its centre point located inside the cell, detecting a defined number of possible bounding boxes simultaneously. Along the location and size of the box, they also return the class of the object and the probability that identified object is found inside. Boxes with probabilities under an adjustable threshold are filtered away, so that only the highest scoring predictions are eventually left.

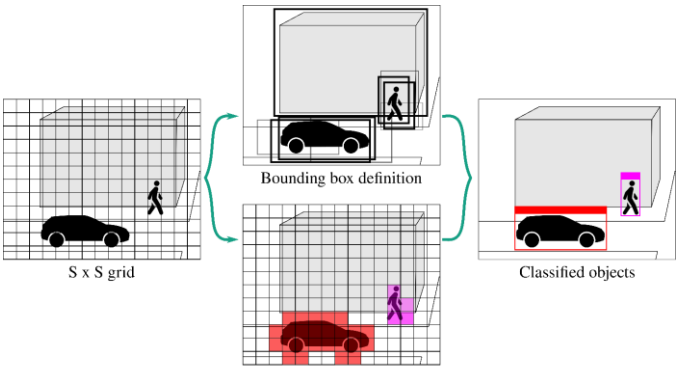

Fig. 2. YOLO object detection procedure.

D. *Distance measurement algorithm*

Computing the distance to a detected object in a monovision camera view requires additional information regarding the scenario, since depth cannot be perceived from a single point of view. Hence, the applied distance measurement algorithm assumes all detected road users to be located on the ground, and the ground to be even. The algorithm measures distance as a component of the object's position in the ground plane, in the horizontal direction the camera is facing. Therefore, the distance represents the component of the object's position that is parallel to the road, given that the installed camera is facing the direction of the road of interest. On a relatively straight road, this distance measurement is sufficient for describing the position of a road user, presuming that the side of the road the road user is located on can be determined from the images. Applying basic trigonometry and optic equations, the distance d can be defined as

$$d = h\frac{1 - \frac{s}{l}tan(\alpha)}{tan(\alpha) + \frac{s}{l}}, \quad (1)$$

which can be derived from the geometry presented in Fig. 3. The horizontal angle of the camera is denoted by α and the distance between the camera lens and sensor is denoted by l. Since measured objects are expected to touch the ground, the height h of the camera is also the height between the camera and the bottom of the measured object. The bottom of the bounding box formed by YOLO is presumed the lowest point of the object. The position of the bottom of the bounding box in the image is provided by YOLO as a ratio of the height of the image it receives. Since the image is physically formed on the active area of the camera sensor, the same ratio applies there.

By knowing the height of the camera sensor, the ratio can be used to compute the physical distance s on the sensor to the bottom part of the bounding box. These parameters provide sufficient information for distance measurements when operating under the previously stated assumptions.

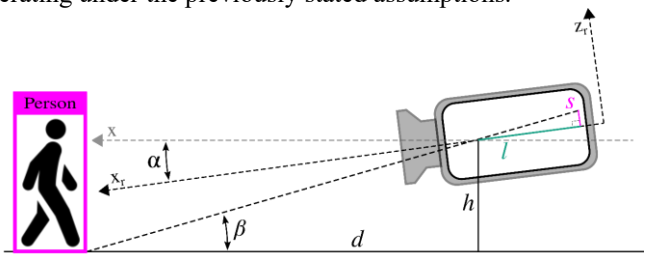

Fig. 3. Illustration of the geometry during the measuring process.

In case the algorithm parameter values are incorrect, or the assumptions are violated, the measurement process is naturally prone to errors. These errors can be mathematically modelled utilising geometries similar to those presented in Fig. 3. For example, in a scenario where the ground is uneven, resulting in the object not being located on the ideal ground plane, the measured distance d_m can be expressed as

$$d_m = \frac{h_m}{h_a} d_a \,. \tag{2}$$

The parameter used for the height of the camera is denoted by h_m, whereas the actual values for the distance and the height between the lowest point of the target and the camera are denoted by d_a and h_a, respectively.

E. Experiment protocol

The following experiments were conducted to validate the feasibility of the presented distance measurement method. The camera was placed at a known height h in a laboratory with a level floor. It was then positioned at the desired angle by placing a point representing the angle on the opposite wall, and tilting the camera so that the centre of the image was on the point. Measurement waypoints were marked on the floor, representing the real-life distances. The distances from the camera to the marker were measured with a laser meter, which had 1 mm accuracy. For each distance measurement, a person was standing still with their toes on the marking, while the camera measured the distance for five seconds. The mean of the measurements was recorded, which consisted of 39 to 42 samples, due to slight variations in the computational speed. These measurements were then replicated with different camera heights and angles. In addition to verifying the system's ability to measure distances, the system's sensitivity to installation errors was studied one factor at a time. This was achieved by measuring distances with known errors in the vertical angle α and horizontal angle γ, which are both illustrated in Fig. 4. Before the recorded experiments were performed, the system was calibrated to determine necessary parameters for the distance calculations.

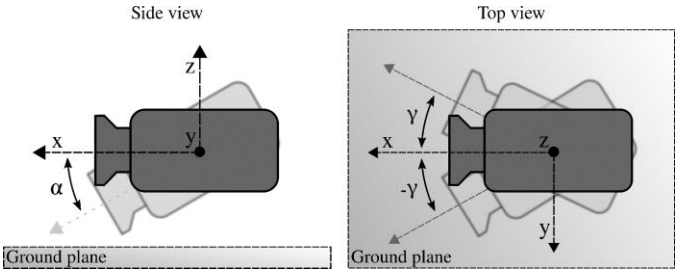

Fig. 4. Vertical and horizontal camera angles depicted.

Detailed information of the height of the sensor's active area s_{max} and the distance from the lens to the sensor l was essential for computing the distance of an object using Eq. 1. The distance l depends on the focus setting and was therefore an unknown value. Furthermore, since there was no certainty of the area the sensor utilises for forming an image, the parameters l and s_{max} were both initially unknown. However, their relation could be conveniently measured, which is here referred to as λ. This was sufficient for the equation computing the distance to the object. As the camera was placed at the location from which the measurements were to be taken, the distance to the point where the view intersected the ground was measured. This setting is depicted in Fig. 5, and according to the geometrical rule of similar triangles, the ratio λ could be defined as

$$\lambda = \frac{s_{max}}{l} = \frac{h}{d}, \tag{3}$$

where the height of the camera is denoted by h, and the distance to the point where the view intersects the ground is denoted by d.

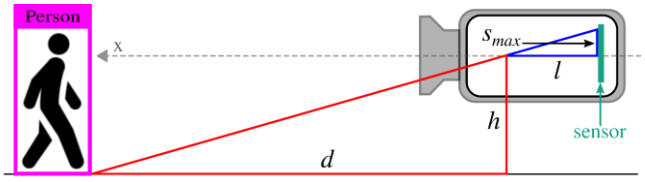

Fig. 5. Definition of λ based on the rule of similar triangles.

The defined value λ was validated with distance measurements, and eventually adjusted by a few percent to optimise the accuracy. In addition to focus, distorted imaging may affect the core optics of the camera. It is typically caused by poor quality lenses, or intentionally applied to record wide-angle photographs and videos without stretching occurring in the edges of the image.

Camera lenses often have radial distortion, which warps captured images. Such distortion would cause irregularities in the distance measurement. Therefore, the distortion of the used camera was investigated with the chessboard calibration of the OpenCV library [45]. No lens distortion was found present, and this was further confirmed by comparing test sets of calibrated and uncalibrated distance measurements which were seemingly identical. Hence, in this case it was deemed unnecessary to compensate for image distortion in the distance measurement.

IV. Results

The experiments provide data displaying the accuracy of the proposed system in multiple scenarios. Firstly, the distance measurement with the algorithm is validated. Secondly, the system's sensitivity to installation errors in the camera angle are studied. Both the horizontal and vertical angle deviations are investigated and analysed. Thirdly, the accuracy achieved with our distance measurement is compared to the theoretical accuracy of a high-end stereo vision camera. Finally, estimations are presented for the values of possible error components affecting the initial validation distance measurements, and their effect in a general application is analysed. All experiments are recorded with the camera at a height of 2.16 m and a λ-ratio of 0.39.

A. Distance measurement accuracy

The distance measurements were conducted from 6 m to 22.5 m, with measurement intervals of 25 cm. The camera was set at an α-angle of 0°. The distance measurements proved to be accurate, as can be seen from the results depicted in Fig. 6.

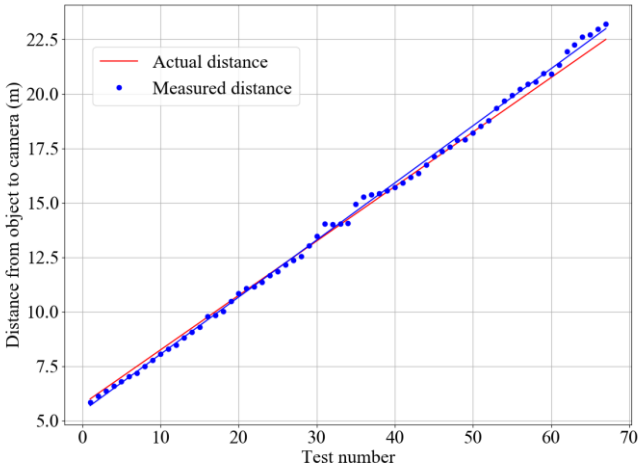

Fig. 6. Measured distances compared to the actual distances.

The proposed system managed a RMSE of 0.32 m, with the maximum absolute error being 0.87 m. These values were made proportional by dividing each error with the respective actual distance of the measurement. These relative errors resulted in a RMSE of 2.1 %, and the maximum absolute relative error was 4.0 %.

B. Comparison to stereo vision

In order to evaluate the proposed system's accuracy, the errors in the results presented in Fig. 6 were compared to the theoretical error of Point Grey's Bumblebee XB3, a high-end stereo vision system. The theoretical error of the stereo vision system was computed with a tool provided by Point Grey [46], using the most favourable values for the situation: baseline of 24 cm, lens focal length of 6 mm and stereo resolution width of 1024 pixels. The errors of the different systems were relatively close to one another, and the comparison of absolute errors is displayed in Fig. 7.

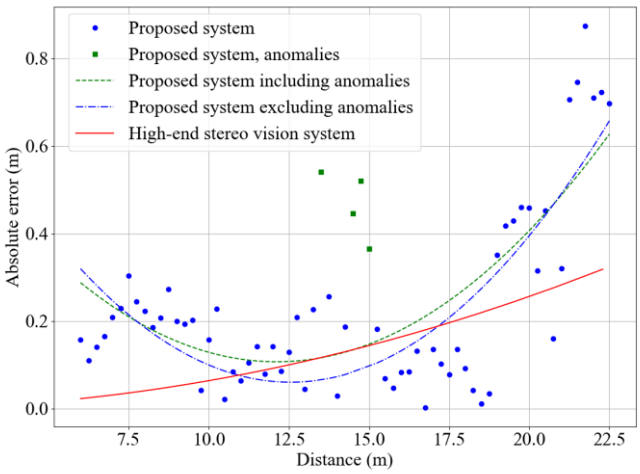

Fig. 7. Absolute error of the proposed system compared to the theoretical error of a high-end stereovision system.

A separate trend line was regressed to the results ignoring the unusually high errors present around the 14 m mark, as they were anomalies. A possible reason for the anomalies was floor T-slot patterns affecting the detection of a person's shoes, as shown in Fig. 8. The system managed to provide results more accurate than the theoretical ones of the stereo vision camera in the 15.5 m to 18.5 m area but showed generally worse results at short-range and long-range. Accuracy in the 15.5 m to 18.5 m area was higher due to exceptional bounding box placement occurring in the area for no apparent reason.

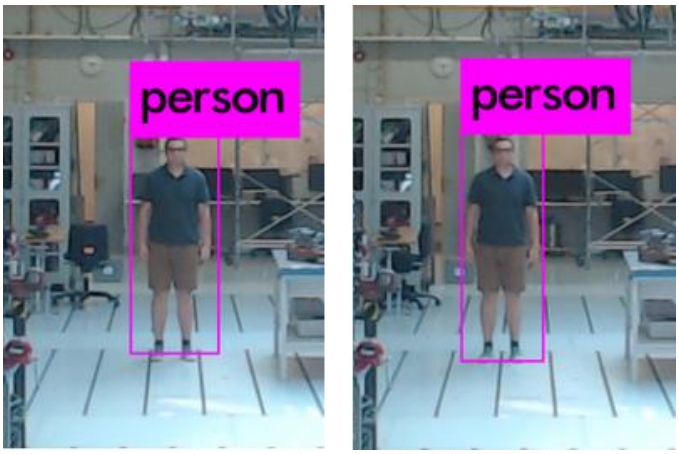

Fig. 8. Detection results visualised at 14 m with different positions relative to the T-slots.

Impact of incorrect bounding box placement was further examined by mathematically modelling the experiment scenario, and computing measurement error contributions of different bounding box offsets. Solely offsets of the bottom of the bounding box were considered, since the bottom is the only parameter of the bounding box affecting the results of the presented measurement method. Several plausible pixel offsets were estimated for the experiment scenario, and their measurement error contributions are presented in Fig. 9.

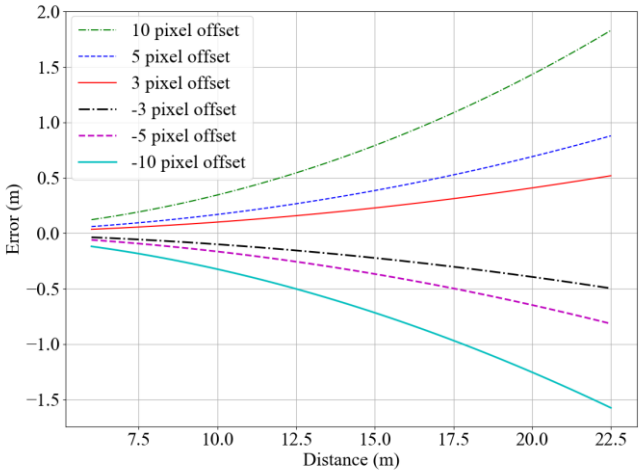

Fig. 9. Contribution of offset of the bottom of the bounding box.

Pixel offset of the bottom of the bounding box contributed an exponentially growing error relative to the measurement distance. Absolute value of the error was found dependent on offset direction, with a negative offset downwards in the image causing smaller absolute error than equal offset upwards in the positive direction. The magnitude of error contribution of pixel offset proved substantial, as previously indicated by the higher measurement errors occurring in the 14 m area with the incorrect bounding box placement. Since the measurement errors in proximity to the 14 m mark correspond to offsets between five and ten pixels, the considered range of pixel offsets was arguably reasonable.

C. *Sensitivity to vertical angle displacement*

The effect of error in the vertical α-angle was analysed by tilting the camera downwards from the $\alpha = 0°$ position, yet performing the calculations as if it still was in the original position. The results are presented in Fig. 9 for the studied errors α-error $\in [1°, 2°, 3°]$. The measurements were carried out in a range of 6 m to 20 m, at 2 m intervals.

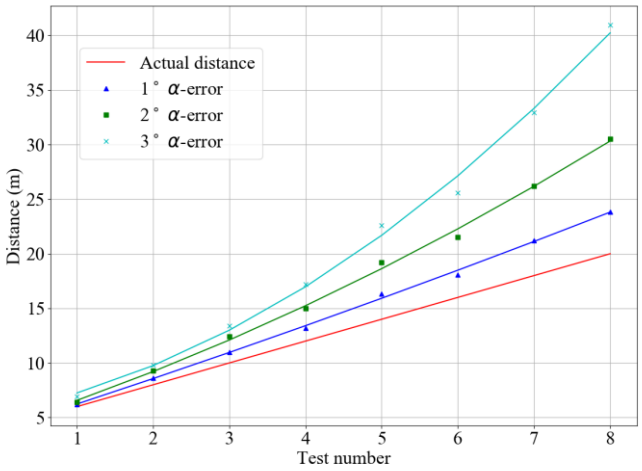

Fig. 10. Contribution of errors in the α-angle.

Error in the α-angle had a critical impact on the distance measurements. As the α-error grew, the distance error became increasingly exponential. With an α-error of 3°, the distance measurement supplied values twice as long as the actual distance at the 20 m mark. The measurements provided values longer than the actual distances since the angle was shifted downwards, causing the person to appear higher in the images. A shift upwards would have resulted in the measured values being shorter than the actual distance.

D. *Sensitivity to horizontal angle displacement*

The system's sensitivity to errors in the horizontal γ-angle was studied in a similar manner to the α-angle. Measurements were recorded with γ-error $\in [5°, 10°, 15°]$. A γ-error of 5° had an insignificant effect on the measured distances and was therefore left out of the results visualised in Fig. 10.

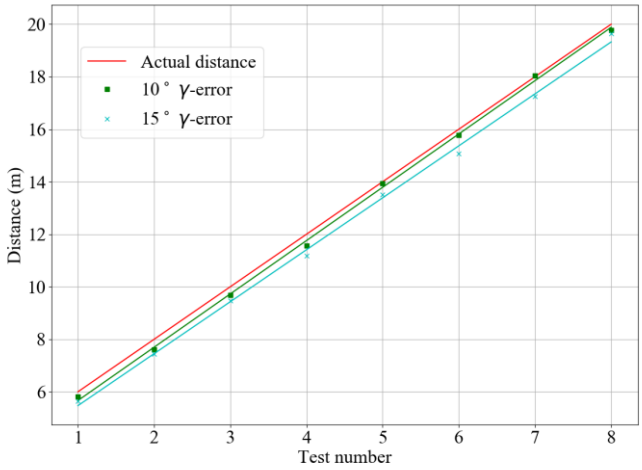

Fig. 11. Contribution of errors in the γ-angle.

Received results were shorter than the actual values, as the system is designed to measure the distance component in the direction of the horizontal angle of the camera. Equivalent results would have been received if the γ-angle had been shifted in the negative direction. The system was robust towards errors in the horizontal γ-angle, and even an error as wide as 15° did not cause a crucial error in the distance approximation.

E. *Summary*

To further examine the total error present in the distance measurements, possible sources of error were identified, and their relative quantities were estimated and documented in Table IV. Estimation was mostly based on geometrical models of the measurement scenario, and the sources of error were studied one factor at a time. Distance errors were computed using estimated maximum error source values for the presented experiments and a general RSU application. Since detection error was difficult to quantify in the measurements, it was deduced by subtracting the other errors from the total error. Estimation of maximum bounding box offset in a general application was based on an assumption that lower bodies of pedestrians are entirely visible in the images. Therefore, no tremendous offset in the bottom of the bounding box should occur.

TABLE IV
ESTIMATED ERROR CONTRIBUTION OF DIFFERENT FACTORS, DISPLAYED AS RELATIVE ERROR OF THE MEASURED DISTANCE

Factor	Experiments	General application
Distortion	**Low** Distortion was negligible.	**Low** Distortion can be corrected with software.
Horizontal angle γ	**Low** • γ-error 0.3° • Relative distance error 0.001 %	**Low** • γ-error 3° • Relative distance error 0.1 %
Height	**High** • Height error 1 cm • Camera height 2.16 m • Relative distance error 0.5 %	**High** • Height error 20 cm • Camera height 4 m • Relative distance error 5 %
Vertical angle α	**High** • α-error 0.1° • Camera height 2.16 m • Nearly linear relative distance error, 0.6 % at 6 m and 1.9 % at 22.5 m	**High** • α-error 0.4° • Camera height 4 m • Nearly linear relative distance error, 1.9 % at 9 m and 7.6 % at 40 m
Detection	**High** • Detection caused a maximum relative distance error of 3.6 %.	**High** • Camera height 4 m • α-angle 10° • Pixel offset 10 • Nearly linear relative distance error, 1.9 % at 9 m and 7.8 % at 40 m.

V. DISCUSSION

The proposed system managed to measure distances at a sufficient level for a vision-based method, as can be seen in the comparison to a high-end stereovision camera in Fig 7. However, the measurements were performed on a level laboratory floor, which is not the case on a real road. If the measurements were to be carried out on a road, the inclination and bumps would violate the assumption of known height between the camera and the target's lowest point. Furthermore, the relative error caused by varying height is the relation between the assumed and the actual height, which is shown in Eq. 2. Therefore, the higher the camera is located, the smaller are the errors caused by an uneven road, since the relative change in height caused by bumps and slopes is lower. This implies that the camera should be placed as high as possible. However, then objects close by might be left out of the field of view, and longer distances may hinder the detection capabilities of YOLO.

At longer distances, objects become seemingly smaller and start to blend into the background, making it more difficult for YOLO to detect them. Even if an object is detected, the detection may not be perfectly accurate, leading to a bounding box that either contains the object only partially, or is oversized for the detected object. When performing the tests, it was noticed that as the measurement distances started reaching the 20 m mark, YOLO excluded person's feet of the bounding box. This caused the measurements to show values higher than the actual distance, as can be seen in Fig. 6. It was noted that the colour of the person's shoes affected the detection, with shoes blending into the floor being left out, and brightly coloured shoes being included in the detection. At shorter distances, YOLO tended to detect bounding boxes that were larger than the person, leading to distance measurements that were shorter than the actual distance, which can be seen in Fig. 6. Errors such as these could possibly be reduced by emphasising bounding box placement accuracy during the training process of a CNN. Range was not the only factor affecting the detection, and it was noticed that a relatively small change in the background could cause a significant anomaly in the measurement.

The larger errors around the 14 m distance in the experiments were as well likely caused by YOLO leaving the person's shoes out. When further investigated, it was found that when the person's feet were next to the laboratory floor's t-slots, YOLO did not recognise the shoes and left them out of the bounding box. With feet placed farther from the T-slots, YOLO managed to perform the detection more accurately. These two scenarios are shown in Fig. 8 at a distance of 14 m. Slight detection failures should not notably disturb the distance measurements in the intended application. As a detected road user is moving, their nearby background is constantly changing, and therefore anomalies such as colour patterns on the road will not affect the overall localisation. Therefore, these errors were left out of the trend line in Fig. 7. Most likely, errors in the vertical α-angle will cause noteworthy deviations in distance measurements.

As shown in Fig. 9, the α-angle has to be accurately known in order to receive valid distance measurements. This means that the camera needs to be precisely installed and calibrated, and it must be attached to something rigid, which will not shift or bend overtime. To avoid false measurements and the possibly dangerous scenarios they might cause, the system could be trained to interpret the truthfulness of its results. In case of continuous measurements that are unlikely for the specific installation, the system would report a possible malfunction. A gyroscope could also actively ensure that the camera is well aligned. This way the system would be able to adjust its measurements to its current position and thus become notably more robust towards shifts in the vertical angle. In the horizontal γ-angle, the system seemed robust by itself, which can be seen in Fig. 10. It should be noted that the system measures the distance component in the ground plane, in the horizontal direction the camera is facing. Therefore, if the camera is pointed in the direction of a straight road, the received values represent distance components along the road, regardless where the detected object is located horizontally. Another important factor in the measurements is the focus of the camera, which affects the value for the ratio λ.

In order to receive valid results with the distance measurement algorithm, the value for the ratio λ must be correct. This requires setting the camera focus to a certain value, limiting the distances at which objects appear sharp in images. Lenses with short focal lengths are well suited for this, since using a single focus setting, they can cover the range from shorter distances to infinity than lenses with long focal lengths. However, shorter focal length results in a wider angle of view, limiting the pixels available for depicting the road at longer distances. With fewer available pixels, the detection typically cannot operate as efficiently, resulting in errors in the localisation. A lens with a long focal length provides more

accurate detection results for longer distances, yet has to be placed farther away from the intersection in order to keep nearby targets in focus. Placing the camera farther increases the localisation errors caused by other factors, such as differing height, and errors in the camera angle, as can be interpreted from the values provided in Table IV. Consequently, the localisation of road users in proximity to the intersection might be less accurate compared to a closer placed camera with a shorter focal length.

VI. Conclusion

Localisation methods are a vital area of ITS research, as they provide the base information many applications operate on. In order to locate road users, an RSU based on monovision CNN object detection was proposed. The object detection enabled a novel type of system to observe road users concealed from drivers by obstacles. Such a system's capabilities were evaluated by performing distance measurements using a person as the target in a laboratory setting. The accuracy of the system was validated and compared to a state-of-the-art stereo vision solution. Furthermore, the system's sensitivity to errors in the horizontal and vertical angles were studied and analysed. Overall, the results were in line with the expectations, and the proposed method showed robust operation with small errors.

The system proved capable of measuring distances accurately in the test range of 6 m to 22.5 m on a level floor. It fared well in comparison to the theoretical error of a stereo vision-based solution, which is intended for a similar purpose. When analysing the sensitivity, an error in the vertical α-angle of the camera was noticed to significantly impact the measured values, with the error growing exponentially to the measured distance. However, the horizontal γ-angle had little impact on the measurements, causing the measured values to appear slightly shorter than the actual distances. The proposed system showed substantial potential and could be a valuable addition to the existing distance measuring methods used in ITS.

Object detection-based distance measurements could significantly improve traffic safety, as these measurements include all types of traffic users: pedestrians, cyclists and vehicles. The proposed system enables mapping the locations of all traffic users present at an intersection. This information could then be presented to the driver, greatly increasing their situational awareness. The experiments presented in this paper were conducted inside with optimal lighting and visibility, level floor and a stationary target. Future work will focus on studying the system in a real-world traffic intersection. With a camera placed at an intersection, topics such as optimal camera height, detection capabilities, abnormalities and measurement accuracy can be investigated in depth with different types of CNNs. Emphasis will be on reliable operation of the unit in challenging conditions, including operation in poor visibility during difficult weather and night-time. Detection in poor visibility will be enhanced with methods such as fusing motion detection with the object detection and monitoring the intersection with an infrared camera. Additionally, fine-tuning the used CNN with traffic related and weather sensitive data will be crucial for optimal detection accuracy in all conditions. Detection accuracy and robustness could also be further increased using an ensemble of multiple parallel CNNs and majority voting on the detection results. Once there is a proper understanding of the proposed system's capabilities, it can be integrated as a part of a V2I system to improve traffic perception and safety.

References

[1] F. Provenzan, "Traffic Control - Connected Vehicle Technologies," *Econlite Group, Inc.*, 2015. [Online]. Available: http://www.econolitegroup.com/wp-content/uploads/2016/10/ConnectedVehiclesOverview_Econolite-20Nov15.pdf. [Accessed: 27-Nov-2018].

[2] European Road Safety Observatory, "Annual Accident Report 2017," *European Commission*, 2017. [Online]. Available: https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/pdf/statistics/dacota/asr2017.pdf. [Accessed: 23-Nov-2018].

[3] J. Barrachina *et al.*, "A V2I-Based Real-Time Traffic Density Estimation System in Urban Scenarios," *Wirel. Pers. Commun. J. Barrachina al., "A V2I-Based Real-Time Traffic Density Estim. Syst. Urban Scenar. Wirel. Pers. Commun., vol. 83, no. 1, pp. 259–280, Jul. 2015.*, vol. 83, no. 1, pp. 259–280, Jul. 2015.

[4] G. Geers, C. Cai, and Y. Wang, "Vehicle-to-infrastructure communication-based adaptive traffic signal control," *IET Intell. Transp. Syst.*, vol. 7, no. 3, pp. 351–360, Sep. 2013.

[5] G. A. Ubiergo and W.-L. Jin, "Mobility and environment improvement of signalized networks through Vehicle-to-Infrastructure (V2I) communications," *Transp. Res. Part C Emerg. Technol.*, vol. 68, pp. 70–82, Jul. 2016.

[6] V. Milanés, J. Villagrá, J. Godoy, J. Simó, J. Pérez, and E. Onieva, "An intelligent V2I-based traffic management system," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 1, pp. 49–58, 2012.

[7] F. Basma, Y. Tachwali, and H. H. Refai, "Intersection collision avoidance system using infrastructure communication," in *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, 2011, pp. 422–427.

[8] S. Atev, O. Masoud, R. Janardan, and N. Papanikolopoulos, "A collision prediction system for traffic intersections," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2005, pp. 169–174.

[9] H. Artail, K. Khalifeh, and M. Yahfoufi, "Avoiding car-pedestrian collisions using a VANET to cellular communication framework," in *2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC)*, 2017, pp. 458–465.

[10] J. Ibanez-Guzman, S. Lefevre, A. Mokkadem, and S. Rodhaim, "Vehicle to vehicle communications applied to road intersection safety, field results," *IEEE Conf. Intell. Transp. Syst. Proceedings, ITSC*, pp. 192–197, 2010.

[11] I. Rashdan, M. Schmidhammer, F. de Ponte Mueller, and S. Sand, "Performance Evaluation of Vehicle-to-

Vehicle Communication for Cooperative Collision Avoidance at Urban Intersections," in *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, 2017, pp. 1–5.

[12] J. Gwehenberger, M. Borrack, and C. Lauterwasser, "Retrospective and Prospective Analysis of the Effectiveness of Driver Assistance Systems with Increasing Degree of Automation," *Aachen Colloq. Automob. Engine Technol.*, 2018.

[13] K. Rubin and J. Betts-Lacroix, "V2V safety system using consensus," U.S. Patent 9,355,561, 2016.

[14] F. Ibrahim, "System and method for lane boundary estimation and host vehicle position and orientation," U.S. Patent 9,261,601, 2016.

[15] S. O. Han, "Adaptive cruise control system for vehicle using V2V communication and control method thereof," U.S. Patent 9,333,971, 2016.

[16] L.-W. Chen, Y.-C. Tseng, and K.-Z. Syue, "Surveillance on-the-road: Vehicular tracking and reporting by V2V communications," *Comput. Networks*, vol. 67, pp. 154–163, Jul. 2014.

[17] J. Lee and B. Park, "Development and Evaluation of a Cooperative Vehicle Intersection Control Algorithm Under the Connected Vehicles Environment," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 1, pp. 81–90, Mar. 2012.

[18] J. Rios-Torres and A. A. Malikopoulos, "A Survey on the Coordination of Connected and Automated Vehicles at Intersections and Merging at Highway On-Ramps," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 5, pp. 1066–1077, May 2017.

[19] "MEC-View, Mobile Edge Computing Based Object Detection for Automated Driving." [Online]. Available: http://mec-view.de/. [Accessed: 01-Nov-2018].

[20] "Safespot." [Online]. Available: http://www.safespot-eu.org/. [Accessed: 19-Jun-2018].

[21] M. Kutila, J. Laitinen, T. Lovas, and A. Barsi, "Final Report: Specifications for infrastructure-based sensing Part A - Sensing systems and Data fusion," *SAFESPOT*, 2007. [Online]. Available: http://www.safespot-eu.org/documents/SF_D2.3.2_Specifications_Part_A_v3.1.pdf. [Accessed: 19-Jun-2018].

[22] P. J. Feenstra *et al.*, "Test and validation results," *SAFESPOT*, 2010. [Online]. Available: http://www.safespot-eu.org/documents/SF_D5.5.2_Test_and_validation_results_v1.4.pdf. [Accessed: 23-Nov-2018].

[23] D. R. Stephens, T. J. Timcho, E. Young, R. A. Klein, and J. L. Schroeder, "Accelerated vehicle-to-infrastructure (V2I) safety applications," *U.S. Department of Transportation*, 2012. [Online]. Available: https://rosap.ntl.bts.gov/view/dot/26496. [Accessed: 25-Jun-2018].

[24] "Global Positioning System (GPS) Standard Positioning Service (SPS) Performance Analysis Report," *Federal Aviation Administration*, 2017. [Online]. Available: http://www.nstb.tc.faa.gov/reports/PAN96_0117.pdf#page=22. [Accessed: 13-Nov-2018].

[25] B.-H. Lee, J.-H. Song, J.-H. Im, S.-H. Im, M.-B. Heo, and G.-I. Jee, "GPS/DR Error Estimation for Autonomous Vehicle Localization," *Sensors*, vol. 15, no. 8, pp. 20779–20798, Aug. 2015.

[26] A. Khattab, Y. A. Fahmy, and A. Abdel Wahab, "High Accuracy GPS-Free Vehicle Localization Framework via an INS-Assisted Single RSU," *Int. J. Distrib. Sens. Networks*, vol. 11, no. 5, p. 795036, May 2015.

[27] S. Nedevschi *et al.*, "High accuracy stereo vision system for far distance obstacle detection," *IEEE Intell. Veh. Symp. 2004*, pp. 292–297, 2004.

[28] L. A. Rentería, C. Torre-Ferrero, M. I. Arenado, and J. M. P. Oria, "Monovision-based vehicle detection, distance and relative speed measurement in urban traffic," *IET Intell. Transp. Syst.*, vol. 8, no. 8, pp. 655–664, 2014.

[29] J. Ehrlich et al., "Final Report: Results on test and validation," *SAFESPOT*, 2010. [Online]. Available: http://www.safespot-eu.org/documents/SF_D2.5.2_Results_on_test_and_validation_v1.15.pdf. [Accessed: 19-Jun-2018].

[30] D. Meissner, S. Reuter, and K. Dietmayer, "Real-time detection and tracking of pedestrians at intersections using a network of laserscanners," *Intell. Veh. Symp.*, pp. 630–635, 2012.

[31] D. Streller, K. Furstenberg, and K. Dietmayer, "Vehicle and object models for robust tracking in traffic scenes using laser range images," in *Proceedings. The IEEE 5th International Conference on Intelligent Transportation Systems*, pp. 118–123.

[32] H. Rohling, S. Heuel, and H. Ritter, "Pedestrian detection procedure integrated into an 24 GHz automotive radar," in *2010 IEEE Radar Conference*, 2010, pp. 1229–1232.

[33] M. Skutek, M. Mekhaiel, and G. Wanielik, "A precrash system based on radar for automotive applications," in *IEEE IV2003 Intelligent Vehicles Symposium. Proceedings (Cat. No.03TH8683)*, pp. 37–41.

[34] S. Sivaraman and M. M. Trivedi, "Looking at Vehicles on the Road: A Survey of Vision-Based Vehicle Detection, Tracking, and Behavior Analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 4, pp. 1773–1795, 2013.

[35] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2179–2195, 2009.

[36] J. Hanhirova, T. Kämäräinen, S. Seppälä, M. Siekkinen, V. Hirvisalo, and A. Ylä-Jääski, "Latency and throughput characterization of convolutional neural networks for mobile computer vision," in *Proceedings of the 9th ACM Multimedia Systems Conference on - MMSys '18*, 2018, pp. 204–215.

[37] J. Huang *et al.*, "Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3296–3297.

[38] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi,

"You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[39] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv*, 08-Apr-2018. [Online]. Available: http://arxiv.org/abs/1804.02767. [Accessed: 24-Nov-2018].

[40] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007.

[41] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[42] R. Girshick, "Fast R-CNN," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.

[43] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[44] J. Redmon, "Darknet: Open Source Neural Networks in C." [Online]. Available: http://pjreddie.com/darknet/. [Accessed: 18-Oct-2018].

[45] "OpenCV library." [Online]. Available: https://opencv.org/. [Accessed: 13-Aug-2018].

[46] Point Grey, "Point Grey Research Stereo Accuracy." [Online]. Available: https://www.ptgrey.com/Content/Images/uploaded/misc/modifiedstereoaccuracy.xls.

[47] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," in *Computer Vision – ECCV 2014*, 2014, pp. 740–755.

Risto Ojala is studying for a B.Sc. degree in mechanical engineering at Aalto University.

He is currently working as a Research Assistant in the Smart Mobility group alongside his studies. He has also participated in the teaching of Computer-aided design as a Course Assistant. He has had remarkable success in his studies, being included in the Dean's list for the academic year 2017-2018.

Jari Vepsäläinen received B.Sc. and M.Sc. degrees at Aalto University in 2014 and 2016.

During his studies, he concurrently worked as a Research Assistant in the Fluid Power laboratory. Jari is currently continuing his academic career as a Doctoral Candidate at Aalto University, a position he has held since 2016. During his doctoral studies, he has also been involved in course planning and assisting in teaching of mechatronics. He is also a student member of the IEEE.

Jussi Hanhirova received M.Sc. degree in 2015 from Aalto University.

He is now working on his PhD thesis at Aalto University Computer Science Department in the Embedded Software research group. His research interests are in methodologies for developing predictable real-time stream processing systems for applications of the Industrial Internet. He also works in close conjunction with the industry to get research results into practical use.

Vesa Hirvisalo received MSc, LicSc, and DSc degrees at Helsinki University of Technology in 1994, 1998, and 2004. He received pedagogical qualification (Higher Education Pedagogy Program) at Helsinki University of Technology, 2002.

He is a Senior University Lecturer at Aalto University Computer Science Department. He is the leader of the Embedded Software research group at Aalto University and an active member of the education faculty. He has authored over 30 peer-reviewed publications and has instructed and supervised over 100 academic thesis works. He has also worked together with the industry to get research results in practical use including the related patenting efforts. His main interests are cyber-physical systems for computationally demanding applications.

Kari Tammi received MSc, LicSc, and DSc degrees at Helsinki University of Technology in 1999, 2003, and 2007. He received a Teacher's pedagogical qualification at Häme University of Applied Sciences in 2017.

He has worked as an Associate Professor at Aalto University since 2015. Kari also serves the Finnish Administrative Supreme Court as Chief Engineer Counselor. Earlier, he worked as Research Professor, Research Manager, Team Leader and other positions at VTT Technical Research Centre of Finland (2000-2015). He was a post-doctoral researcher at North Carolina State University (USA, 2007-2008), and a researcher at CERN, the European Organization for Nuclear Research in 1997-2000. He has authored more than 70 peer-reviewed publications cited in more than 1500 other publications. Dr. Tammi serves as the deputy chair for IFTOMM Finland, and he is a member of IEEE and the Finnish Academy of Technology.