Detecting darting out pedestrians with occlusion aware sensor fusion of radar and stereo camera

Palffy, Andras; Kooij, Julian F.P.; Gavrila, Dariu M.

DOI
10.1109/TIV.2022.3220435

Publication date
2023

Document Version
Final published version

Published in
IEEE Transactions on Intelligent Vehicles

Detecting Darting Out Pedestrians With Occlusion Aware Sensor Fusion of Radar and Stereo Camera

Andras Palffy, *Member, IEEE*, Julian F. P. Kooij, *Member, IEEE*, and Dariu M. Gavrila, *Member, IEEE*

***Abstract*—Early and accurate detection of crossing pedestrians is crucial in automated driving in order to perform timely emergency manoeuvres. However, this is a difficult task in urban scenarios where pedestrians are often occluded (not visible) behind objects, e.g., other parked vehicles. We propose an occlusion aware fusion of stereo camera and radar sensors to address scenarios with crossing pedestrians behind such parked vehicles. Our proposed method adapts both the expected rate and properties of detections in different areas according to the visibility of the sensors. In our experiments on a real-world dataset, we show that the proposed occlusion aware fusion of radar and stereo camera detects the crossing pedestrians on average 0.26 seconds earlier than using the camera alone, and 0.15 seconds earlier than fusing the sensors without occlusion information. Our dataset containing 501 relevant recordings of pedestrians behind vehicles will be publicly available on our website for non-commercial, scientific use.**

***Index Terms*—Advanced driver assistance systems, millimeter wave radar, object detection, radar detection.**

I. Introduction

ABOUT 23% of the 1.35 million traffic fatalities world-wide involve pedestrians [1]. Automated driving has the potential to significantly reduce these traffic deaths, yet the sensor-based detection and tracking of pedestrians from a moving vehicle remains challenging. Pedestrians have a wide variation in appearance, can quickly alter their course, and can step onto the road at pretty much any location.

Intelligent vehicles can use multiple sensors to cope with this task: cameras [2], [3], [4], radars [5], [6], [7] and LiDARs [8], [9]. Fusing different sensors, e.g., camera with radar [10] or camera with LiDAR [11], can increase the reliability and redundancy of such systems. In this paper, we consider the fusion of a (stereo) camera with a radar. These are low-cost sensors with complementary strengths that are well established in driver assistance context on the market. Cameras provide color/texture information at a fine horizontal and vertical resolution. Radar sensors provide accurate depth information, can directly measure the radial velocities and are more robust to adverse weather and lighting conditions.

Manuscript received 2 August 2022; revised 9 October 2022; accepted 25 October 2022. Date of publication 8 November 2022; date of current version 20 March 2023. This work was supported by Dutch Science Foundation NWO-TTW through SafeVRU Project under Grant 14667. *(Corresponding author: Dariu M. Gavrila.)*

The authors are with the Intelligent Vehicles Group, TU Delft, 2628 CD Delft, The Netherlands (e-mail: d.m.gavrila@tudelft.nl).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TIV.2022.3220435.

Digital Object Identifier 10.1109/TIV.2022.3220435

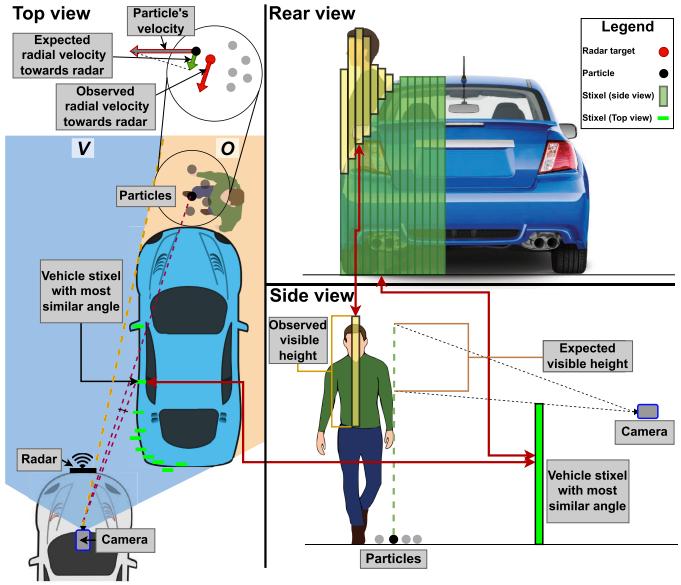

Fig. 1. Darting out scenario: a pedestrian steps out from behind a parked car (blue) which blocks the line-of-sight of the ego-vehicle (white). We propose to detect such pedestrians with the fusion of stereo camera and radar in an occlusion aware way, i.e., first building an occlusion model of the environment and then expecting fewer and different detections (e.g. shorter visible parts of the pedestrian) from the occluded regions (O) than from the visible, unoccluded ones (V).

Pedestrian sensing is often complicated in urban scenarios by occlusions, such as by parked vehicles. A substantial 26% of the accidents with crossing pedestrian analyzed in [12] involved some form of visual occlusion. In fact, this case is so important that the consumer advocacy group Euro NCAP designates a special test scenario for it, titled "Running Child from Nearside from Obstruction" [13]. This case of a pedestrian *darting out* [14] is illustrated in Fig. 1. It is particularly dangerous because neither a human driver nor the pedestrian have initially a clear, direct view of the other. Similarly, in an automated driving setting, a parked vehicle would block direct line-of-sight from the sensors of the ego-vehicle to the pedestrian. However, the extent of this blockage depends on the sensor's type and on the size and shape of the occlusion.

A camera may see the upper body of a pedestrian behind a passenger car, while a person behind a larger vehicle, such as a truck or a van may be invisible to the sensor. On the other hand, commercially available 2+1D radars, which provide two spatial dimensions (range and azimuth) and one dimension for Doppler (radial velocity), are often able to detect the reflections

of a pedestrian even in complete occlusion due to multipath propagation [15], [16]. That is, the reflected radar signal may "bounce" off other parked cars or the ground beneath the occluding vehicle and reach the ego-vehicle's sensor even if there is no direct line-of-sight. Such indirect reflections are weaker and occur less frequently than direct ones [15], but they could still provide valuable information about a potentially darting out pedestrian.

Since both camera and radar sensors are affected by occlusions, their fusion preferably requires an occlusion model that describes how many detections to expect from each sensor in the differently occluded areas of the scene (e.g., to expect fewer detections behind cars). In addition, the occlusion model could also provide information about the expected properties of such detections, e.g., that the visible part of a partially occluded pedestrian may be smaller than an unoccluded one. The stereo camera is a suitable sensor to create this occlusion model because it provides rich and dense textural and depth information that can help accurately detect and model the occluding vehicle itself.

In this paper, we present a Bayesian occlusion aware sensor fusion system designed to detect darting out pedestrians. We show that incorporating an occlusion model into such a sensor fusion system helps to detect darting out pedestrians earlier; thus precious time is gained to initiate emergency braking or steering, if needed. While we consider the fusion of (stereo) camera and radar, the framework is suitable to integrate other sensors, e.g., LiDAR.

The paper is structured as follows. In Section II we discuss previous work. Then, in Section III, we present our generic occlusion aware Bayesian multi-sensor fusion filter. Details of how this filter was implemented with radar and stereo camera sensors, and applied to darting out scenarios are discussed in Section IV. Section V describes the dataset that was created and used for this work. In Section VI, we present our experiments and results, which are discussed in-depth in Section VII. Finally, Section VIII concludes the paper.

II. Related Work

In this section, we first discuss camera-, radar-, and fusion based methods for pedestrian detection, with a focus on darting out scenarios and occluded pedestrians. Afterwards, we give an overview of some widely used methods for both object tracking and for modeling the environment considering occlusions. Finally, we review the available automotive datasets and their usability for this research.

A. Camera Based Approaches

Cameras are often used for pedestrian detection as they provide rich information while being relatively inexpensive. In recent years, convolutional neural networks (CNNs) and deep learning methods [3], [4] dominate in this field.

The problem of occlusion is widely recognized, e.g., many benchmarks define separate metrics for different levels of occlusion [4], [17]. For an overview of camera based methods that consider occlusions, see [18]. Several approaches aimed to explicitly account for occlusions by learning a set of component detectors and fusing their results to detect partially occluded pedestrians [19], [20], [21], [22]. More recently, researchers proposed special loss functions [23], [24] or top-down approaches [25] to jointly estimate the state of close-by pedestrians occluding each other, introduced hard negative mining to increase the occlusion tolerance of networks [26], or proposed to explicitly collect more training data of partially occluded pedestrians [4] to address the problem. However, none of these methods used a global scene model to describe the occlusions that may affect the number and attributes of detections.

B. Radar Based Approaches

Radars have been used to detect road users in a variety of ways, including clustering algorithms [27], [28], convolutional neural networks [6], [29] or point cloud processing neural networks [30], [31]. A radar based multi-class classification system (addressing pedestrians and pedestrian groups) was presented in [27]. [5] and [32] both aim to distinguish pedestrians from vehicles based on features such as size and velocity profiles of the objects using radar. Some methods also used radars to detect pedestrians in darting out or similar situations. [33] presented a tracking method using track-before-detection and particle filtering. The system was also tested in scenes of the pedestrian entering and exiting an occluded region behind a car. The radar was able to provide measurements even in the occlusion. However, the occlusion itself was not considered, and although they compared the performance to a camera based detection system, no fusion occurred. In [15], a binary classification system of pedestrians and static objects was presented that uses low-level radar data as input and extracts hand-crafted features. The system was evaluated using darting out scenarios, but no sensor fusion was used, nor was occlusion investigated as a possible source of information. [16] exploited that radar signals often "bounce" off large flat surfaces. They showed that it is possible to detect moving road users outside the direct line-of-sight with reflected radar measurements by using building facades or parked vehicles as relay walls. In [34], the authors explicitly addressed the detection of fully occluded, darting out pedestrians with radar. They designed an experimental setup with a static radar sensor in an indoor area (behind a corner) and an outdoor area (behind a van). Movement of the occluded pedestrian is then classified by clustering into different behavior types, such as walking towards, walking out of it, and walking inside the occluded region. None of these methods considered occlusion as a source of information, and none of them compared or fused camera and radar sensors to detect darting out pedestrians in realistic environments, i.e., from a moving ego-vehicle.

C. Fusion Based Approaches

Sensor fusion was extensively researched to provide more robust perception solutions either via model-based (mathematical, e.g., Kalman or particle filters, evidence modeling) [10], [35], [36], or data-driven approaches (e.g., with neural networks) [37], [38]. In this subsection, we focus on fusion systems that use radar, with particular attention to whether and how these systems address occluded pedestrians. A Kalman-filter

based pedestrian tracking system using camera and radar was introduced in [10] for indoor, static applications. To deal with the frequent occlusion of the lower body, the authors trained their camera based detector to detect only the upper body of pedestrians, but they did not explicitly model occlusions. In [39], LiDAR and radar were fused to detect pedestrians in a static experimental setup. First, a binary occlusion map of the scene was created by detecting occluding objects with LiDAR. This map was then used to select which sensors to use for detection: both sensors for unoccluded regions, and purely radar for occluded regions, exploiting its the multipath property. In [36], all three sensors were combined in a multi-class system for detecting moving objects, including pedestrians, in an intelligent vehicle setup using an occupancy grid representation. The LiDAR was used as the main sensor to detect moving objects, while camera and radar were mainly used for classification. The influence of occlusions was not considered. None of the fused systems found were developed for use in intelligent vehicles to address darting out scenarios, or considered occlusion as a source of information beyond helping sensor selection.

D. Tracking

Pedestrians are often tracked with Kalman Filters both in camera based [40] and radar based [41] detection systems. Kalman filters can only model linear motion. Situations with possibly non-linear motion dynamics, e.g., a pedestrian who may or may not stop at the road side, can be handled by using an "extended" Kalman Filter, or by switching between multiple linear motion models with a switching dynamic system [40]. Another commonly used method for pedestrian tracking is the particle filter [33], [42], [43], [44] which estimates the posterior distribution over the state space using a set of weighted particles. Unlike Kalman Filters, a particle filter can handle non-linear motion dynamics, and can represent arbitrary, potentially multi-modal distributions. To satisfy our use case (detecting and tracking a pedestrian), a filter should not only track an object of interest (i.e., a pedestrian), but also report a probability that a pedestrian is present in the scene. [42], [45] give solutions to incorporate this existence probability into particle filters.

E. Environment Modeling

Modeling occluded areas in the environment is often done in bird's-eye view (BEV). A common approach is to aggregate range measurements from radar or LiDAR sensors into a 2D occupancy grid and then project "shadows" behind the extracted objects [46], [47]. Creating an environment model with camera information can lead to a faster process (i.e., it does not need to be accumulated) and provides more information about the nature of the occluding object (e.g., whether it is a car) due to the rich texture information. In [43], the goal was to explicitly model only the occlusions caused by (parked) vehicles. To this end, 2D detections in the image plane were fetched from the *car, bus, truck,* and *van* classes from the Single Shot Multibox Detector (SSD) [48]. Depth (i.e., distance from the ego-vehicle) was estimated by projecting the stereo point cloud into the camera view and taking the median distance of the points inside each bounding box. Using this depth, we back-projected each 2D box to the 3D space to get a "2.5D" detection: a line segment in BEV with length corresponding to the width of the projected box. Areas behind these detections were considered occluded, creating a binary map. While this solution resulted in fast processing time and contributed to earlier detection of darting out pedestrians in the experiments, it also had some drawbacks. By assigning a single distance to the entire occluding vehicle, parts of the vehicle closer/farther than that distance are incorrectly considered "regular" unoccluded/occluded (but still walkable) regions. However, a pedestrian cannot be physically present in either of these halves. Modeling occlusion with a bounding box also has limitations in width and height, e.g., a pedestrian may be more visible behind the shorter parts of a car than behind its tallest point, but these two cases are treated identically.

An alternative camera based approach to creating a more accurate occlusion model that is still computationally efficient may be to use stixels [49]. Stixels are rectangular column-wise group of pixels based on disparity information with the goal of reducing the complexity of the stereo point cloud. Since the original publication [49], researchers have integrated class information [50] and later instance information into stixels [51]. The latter are referred to as Instance Stixels and could be a well suited input for an occlusion model because they follow the shape of an occluding car (both in depth and width/height) and are still computationally efficient to compute and process. In addition, the same Instance Stixels representation can also serve as input to a pedestrian detection and tracking system by providing the location and height of the pedestrian.

F. Datasets

To study the detection of darting out pedestrians with the fusion of camera and radar, a dataset is needed that 1) contains measurements from both sensors and 2) contains hundreds of the scenario under study. Several datasets have been published to help the development and testing of autonomous vehicles, e.g., the well-known KITTI [17] or the EuroCity dataset [4]. In recent years, the number of datasets containing radar data has increased with different goals such as ego-localization [52], [53], object classification [54], or object detection [55]. At the time of writing, nuScenes [56], Zendar [57], Astyx [58], and View-of-Delft [31] are the only publicly available automotive datasets that include measurements from both a camera and a radar sensor (which provides Doppler data). However, a real-world (i.e. not scripted or directed) dataset will always have relatively few darting out examples and thus, none of these datasets are suitable for our research.

III. Proposed Approach

A. Overview and Contributions

The goal of this paper is to fuse radar and stereo camera (Fig. 2, blue and red dashed rectangles) by incorporating occlusion information to detect darting out pedestrians. To this end, we propose a generic Bayesian filter to fuse these sensors in an occlusion aware manner. This estimates not only the 2D position

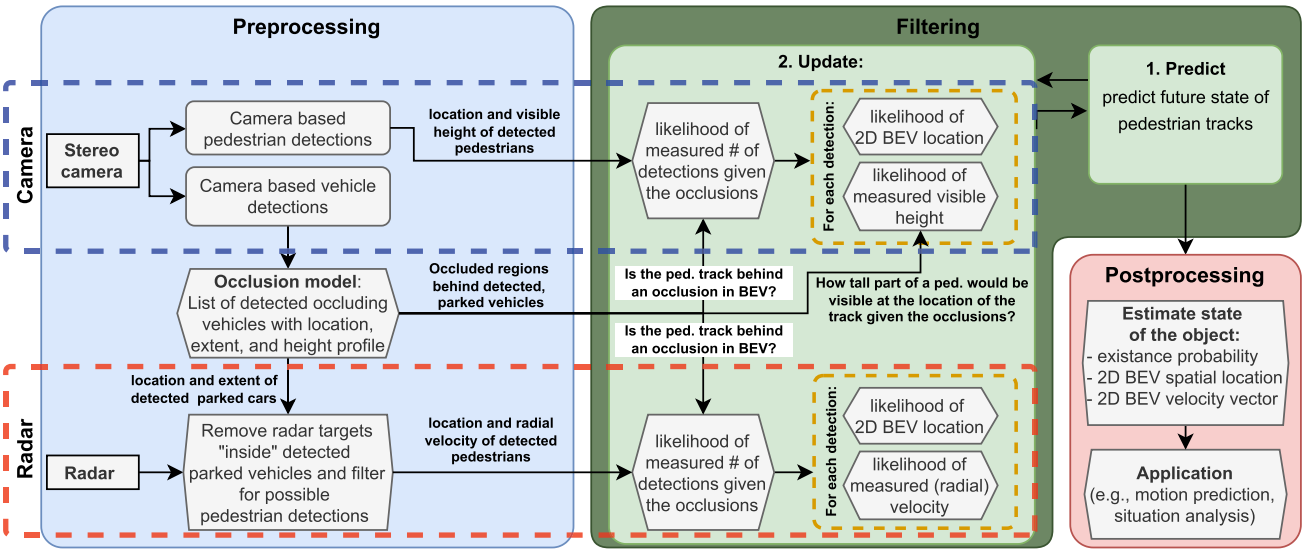

Fig. 2. Overview of our pipeline. Sensor data is processed to get the pedestrian detections and the occlusion model (Preprocessing). Future states of the object in the filter are predicted, and then their likelihood are updated with the detections considering the occlusion model (Filtering). The estimated existence probability and state of the object are calculated, which information can be used in subsequent applications (Postprocessing). Blue/red dashed boxes mark camera/radar specific steps that are described in Section IV.

and velocity of an object's center on the ground plane (i.e. BEV), but also the probability that the object of interest (i.e., a pedestrian) is present in the scene. The used state space will be discussed in details in Subsection III-B.

First, in a prediction step, we define the prior distribution of the filter given previous measurements (Fig. 2, Predict step). The distribution of predicted positions and velocities is defined for three cases: a new object entering the scene, an object leaving the scene, and finally a tracked object remaining in the scene. Please refer to Subsection III-C for details.

After the prediction step, the new detections are fetched from each sensor (Fig. 2, Preprocessing) and incorporated in the update step (Fig. 2, Update step), which is discussed in details in Subsection III-D. We assume conditional independence of the sensors given the true state of the object, thus we can perform the update with their set of detections individually whenever they arrive, even if the sensors operate asynchronously at different frame rates. We describe here the update in a generic way and define sensor specific details, e.g. measurement models, later in Section IV.

When updating with any of the sensors, its K detections (as determined by its measurement model) are fed into the filter. This updates the likelihood of the hypotheses in two ways. First, the likelihoods of measuring K detections with this sensor are calculated. Since the number of detections depends on the position of the object, we can incorporate information from an occlusion model here. That is, our system adjusts the expected number of detections to the visibility of a position and expects more/less detections at unoccluded/occluded locations, see Fig. 1. Second, we also consider the unique capabilities of the sensors. That is, we estimate the likelihood of the attribute of the detection based on the estimated state of the object. Here we could use, for example, the velocity measurement of a radar or the classification confidence of a camera. We can also evaluate the size of the visible part of a pedestrian given its assumed occlusion condition.

The occlusion model can be retrieved from a single or from a combination of sensors, or from an independent source. In this paper, it will be provided by the stereo camera, see Section IV.

Finally, after the filtering, the object's probability of existence and state (i.e., 2D BEV center location and velocity) can be estimated and used in subsequent processing steps, e.g., in predicting future positions or evaluating the dangerousness of the scene (Fig. 2, Postprocessing).

Our contributions are as follows.

1) We propose a generic occlusion aware multi-sensor Bayesian filter for object detection and tracking.
2) We apply the proposed filter as a radar and stereo camera based pedestrian detection and tracking system on challenging darting out scenarios. We show that incorporating occlusion information and the radar sensor into our model helps detect darting out pedestrians earlier while keeping the number of false alarms low when the pedestrian stays behind the car.
3) We share our dataset[1] containing more than 500 relevant scenarios with camera, radar, LiDAR, and odometry data.

This work builds upon our previous conference publication [43], where we initially proposed an occlusion-aware Bayesian filter for darting out pedestrians based on stereo camera and radar. This work features an improved sensor measurement model (incorporation of additional attributes besides location, see Subsections IV-B and IV-C). Among others, the occlusion extent is now more accurately represented by a height profile derived from instance segmentation rather than by a bounding

[1]The dataset will be made freely available at https://intelligent-vehicles.org/datasets/ to academic and non-profit organizations for non-commercial, scientific use.

box derived from an object detector (see Subsection II-E). In terms of validation, this work features a significantly enlarged dataset and added experimentation.

B. State Space and Notations

Now we discuss the mathematical formulation of our proposed generic occlusion aware, multi-sensor Bayesian filter without sensor related specifics. Let the space $\mathcal{T}$ consist of a 2D (lateral and longitudinal) position and velocity, and a binary flag marking if the tracked object (e.g., a pedestrian) exists. Let $\mathbf{h}$ be a state vector in $\mathcal{T}$ (vectors are written in boldface):

$$\mathcal{T} : \mathcal{R} \times \mathcal{R} \times \mathcal{R} \times \mathcal{R} \times \{0, 1\}, \tag{1}$$

$$\mathbf{h} \in \mathcal{T},\ \mathbf{h} = (\mathbf{x}, \mathbf{v}, \mathcal{E}), \tag{2}$$

where $\mathbf{x} = (x, y)$ and $\mathbf{v} = (v_x, v_y)$ are the object's 2D BEV position and velocity vectors on the ground plane, and $\mathcal{E}$ represents the existence probability. I.e., $\mathcal{E} = 1$ means there is a pedestrian in the scene and $\mathcal{E} = 0$ represents its absence.

We define a Bayesian filter for detection and tracking which estimates the posterior state distribution $P(\mathbf{h}_t|\mathcal{Z}_{1:t})$ given all measurements $\mathcal{Z}_{1:t}$. The filter operates on-line, integrating new measurements into a posterior using Bayes' theorem:

$$P(\mathbf{h}_t|\mathcal{Z}_{1:t}) \propto P(\mathcal{Z}_t|\mathbf{h}_t) \cdot P(\mathbf{h}_t|\mathcal{Z}_{1:t-1}), \tag{3}$$

where $\mathcal{Z}_t$ is the set of all sensor detections at current time t. Here the prior distribution $P(\mathbf{h}_t|\mathcal{Z}_{1:t-1})$ for time t is obtained by applying a state transition probability on the previous posterior, and integrating over the previous state $\mathbf{h}_{t-1}$ following the Chapman-Kolmogorov equation:

$$P(\mathbf{h}_t|\mathcal{Z}_{1:t-1}) = \int P(\mathbf{h}_t|\mathbf{h}_{t-1}) \cdot P(\mathbf{h}_{t-1}|\mathcal{Z}_{1:t-1}) d\mathbf{h}_{t-1}. \tag{4}$$

We are thus required to define the state transition distribution $P(\mathbf{h}_t|\mathbf{h}_{t-1})$ for the filter's prediction step, and measurement likelihood function $P(\mathcal{Z}_t|\mathbf{h}_t)$ for the update step, which we will derive in the following subsections. Note that the posterior contains the expected existence probability of a pedestrian in the scene:

$$P(\mathcal{E}_t|\mathcal{Z}_{1:t}) = \iint P(\mathbf{h}_t|\mathcal{Z}_{1:t}) d\mathbf{x}_t d\mathbf{v}_t. \tag{5}$$

C. Prediction Step

The state transition distribution is factorized into two terms:

$$P(\mathbf{h}_t|\mathbf{h}_{t-1}) = P(\mathcal{E}_t|\mathbf{h}_{t-1}) \cdot P(\mathbf{x}_t, \mathbf{v}_t|\mathcal{E}_t, \mathbf{h}_{t-1}). \tag{6}$$

The first term estimates the object presence flag $\mathcal{E}$. A new object can appear with a probability of p_n. Unlike p_n, $p_s(\mathbf{h}_{t-1})$, the probability that an object stays in the scene depends on the previous state $\mathbf{h}_{t-1}$, because the position of the object affects the probability that it will suddenly leave the region of interest. Using these, we can determine the probability of $\mathcal{E}$ given the previous state $\mathbf{h}_{t-1}$ for entering (new), not present and not entering, staying, and leaving objects respectively:

$$P(\mathcal{E}_t = 1|\mathcal{E}_{t-1} = 0, \mathbf{h}_{t-1}) = p_n, \tag{7}$$

$$P(\mathcal{E}_t = 0|\mathcal{E}_{t-1} = 0, \mathbf{h}_{t-1}) = 1 - p_n, \tag{8}$$

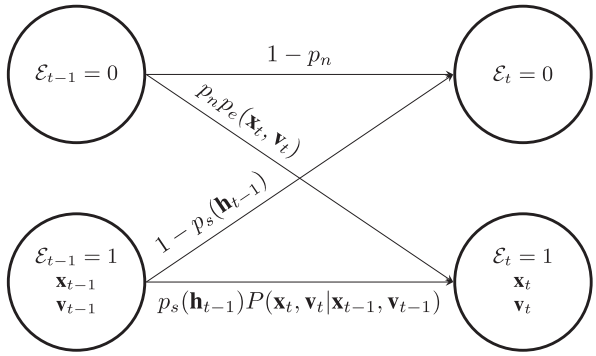

Fig. 3. Transition of states. $\mathcal{E}_t = 0$ denotes the lack of object, and $\mathcal{E}_t = 1$ denotes the presence of an object with the configuration of $\mathbf{x}_t$, $\mathbf{v}_t$ at timestamp t.

$$P(\mathcal{E}_t = 1|\mathcal{E}_{t-1} = 1, \mathbf{h}_{t-1}) = p_s(\mathbf{h}_{t-1}), \tag{9}$$

$$P(\mathcal{E}_t = 0|\mathcal{E}_{t-1} = 1, \mathbf{h}_{t-1}) = 1 - p_s(\mathbf{h}_{t-1}). \tag{10}$$

In case an object is present ($\mathcal{E}_t = 1$), the values of $\mathbf{x}$ and $\mathbf{v}$ are distributed as follows for entering and staying objects respectively:

$$P(\mathbf{x}_t, \mathbf{v}_t|\mathcal{E}_t = 1, \mathcal{E}_{t-1} = 0, \mathbf{h}_{t-1}) = p_e(\mathbf{x}_t, \mathbf{v}_t), \tag{11}$$

$$P(\mathbf{x}_t, \mathbf{v}_t|\mathcal{E}_t = 1, \mathcal{E}_{t-1} = 1, \mathbf{h}_{t-1}) = P(\mathbf{x}_t, \mathbf{v}_t|\mathbf{x}_{t-1}, \mathbf{v}_{t-1}).$$

For this last term, we use a constant velocity dynamic model similar to [40], with a normally distributed acceleration noise $\mathbf{a} \sim N(0, \Sigma_a)$:

$$\mathbf{v}_t = \mathbf{v}_{t-1} + \mathbf{a}\Delta t, \tag{12}$$

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \mathbf{v}_{t-1}\Delta t + \frac{1}{2}\mathbf{a}\Delta t^2. \tag{13}$$

Through the introduction of the binary flag $\mathcal{E}$, the full state transition can be regarded as a state machine, see Fig. 3.

D. Update Step

Now we describe the likelihood $P(\mathcal{Z}_t|\mathbf{h}_t)$. We follow the common assumption of conditional independence for our sensors, thus the single-sensor update step described here can be applied independently to each. The sensor s returns K detections at once: $\mathcal{Z} = \{z^1, \ldots z^K\}$. Each detection z^k contains a 2D BEV location and some additional attributes: $z^k = [z_{pos},\ z_{attr}]$. To include occlusion awareness, our measurement model introduces several auxiliary variables, with conditional dependencies as shown in the graphical model of Fig. 4. These variables and their distributions will be introduced in the next paragraphs, where we first distinguish between the expected number of detections, which differentiates our occlusion aware from the naive approach, and then the likelihood term for a single measurement z^k.

a) Detection rates: The total number of detections (K) is the sum of foreground (K^F) and background (K^B) detections: $K = K^F + K^B$. If we consider detections as conditionally independent events occurring during a fixed interval, it is natural to model the number of foreground (true positive) and background (false positive) detections with two Poisson distributions. Let

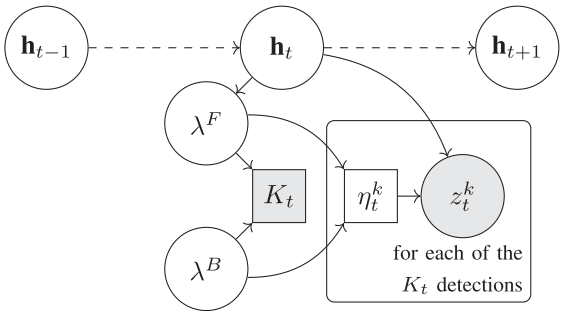

Fig. 4. Graphical model of the probabilistic dependencies in a single time slice t with K detections $\mathcal{Z}_t = \{z_t^1, \dots z_t^K\}$. $\mathbf{h}_t$ is the state vector and λ^B, λ^F are the expected detection rates. The binary flag η_t^k denotes if the k^{th} detection z_t^k comes from foreground or background. Discrete/real variables are shown with square/circle nodes. Observed variables are shaded.

us denote the corresponding detection rates with $\lambda^F(\mathbf{x}, \mathcal{E})$ and λ^B for the foreground and background detections respectively. The values of K^B, K^F follow Poisson distributions, $K^B \sim Pois(\lambda^B)$ and $K^F \sim Pois(\lambda^F)$, with scalar parameters λ^B and λ^F. The total number of detections K is then also Poisson distributed:

$$P\left(K|\lambda^B, \lambda^F\right) = Pois\left(\lambda^B + \lambda^F\right). \tag{14}$$

We distinguish between our novel occlusion aware filter by the way it determines the value of the foreground detection rate, as opposed to the naive approach. In the proposed *Occlusion Aware Filter (OAF)* approach, the number of foreground detections depends both on the object's presence and location. A benefit of Poisson distributions is that we can incorporate the occlusion information here with a spatially dependent rate parameter, i.e., more true detections are expected if the pedestrian is unoccluded (i.e. visible) than if the pedestrian is occluded:

$$\lambda^F = \begin{cases} \lambda^F_{unocc} & \text{if } \mathbf{x} \in V, \\ \lambda^F_{occ} & \text{if } \mathbf{x} \in O, \end{cases} \qquad \text{(OAF)} \tag{15}$$

where λ^F_{unocc}, λ^F_{occ} indicate the expected detection rates in unoccluded (V), occluded (O) areas respectively, see Fig. 1. The extent of these areas will be determined by the implementation-specific environment occlusion model.

In contrast, a naive (i.e., not occlusion aware) filter assumes that λ^F is constant, targeting the more typical unoccluded case:

$$\lambda^F = \lambda^F_{unocc}. \qquad \text{(naive approach)} \tag{16}$$

Our occlusion aware filter behaves the same as a naive one in unoccluded cases, but in occluded positions it adapts its expected rate λ^F.

b) Measurement likelihood: Derived from the properties of Poisson distributions, the number of false and true positive detections given K are distributed as Binomial distributions parametrized by the ratio of λ^B and λ^F. Thus, the probability of a detection z^k being foreground/background is (given K number of detections):

$$P\left(\eta^k = 1|\lambda^B, \lambda^F\right) = \frac{\lambda^F}{\lambda^F + \lambda^B}, \tag{17}$$

$$P\left(\eta^k = 0|\lambda^B, \lambda^F\right) = \frac{\lambda^B}{\lambda^F + \lambda^B}, \tag{18}$$

where the binary flag η^k denotes if the k^{th} detection z^k comes from the tracked object, i.e., is a true positive detection. Since every detection is conditioned on $\mathcal{E}$ and η^k latent variables, we have to define the likelihood function $P\left(z^k|\mathcal{E}, \mathbf{x}, \mathbf{v}, \eta^k\right)$ for the following cases: $(\mathcal{E} = 1, \eta^k = 1)$, $(\mathcal{E} = 1, \eta^k = 0)$, $(\mathcal{E} = 0, \eta^k = 0)$, which stand for the true positive and for the false positive cases, with and without a present pedestrian, respectively.

Unlike [43], in which we only considered location of detections, here we compose this likelihood function with two parts: a spatial component (i.e., likelihood of detection's location) and an attribute component (likelihood of such a detection at that location). We assume that true positive (foreground) detections are spatially distributed around the object's position $\mathbf{x}$ described by some distribution $L_F(z_{pos}|\mathbf{x})$, and that false (background) detections are distributed as described by some distribution $L_B(z^k)$. Similarly, we define the attribute likelihood functions, $A_F(z_{attr}|\mathbf{x}, \mathbf{v})$ and $A_B(z_{attr})$ for true and false detections, but also conditioned on velocity. Then:

$$\begin{aligned} P\left(z^k|\mathcal{E} = 1, \mathbf{x}, \mathbf{v}, \eta^k = 1\right) &= L_F(z_{pos}|\mathbf{x}) \cdot A_F(z_{attr}|\mathbf{x}, \mathbf{v}), \\ P\left(z^k|\eta^k = 0\right) &= L_B(z_{pos}) \cdot A_B(z_{attr}). \end{aligned} \tag{19}$$

Finally, the complete likelihood of a single measurement is:

$$\begin{aligned} P\left(z^k|\mathcal{E}, \mathbf{x}, \mathbf{v}, \lambda^B, \lambda^F\right) &= P\left(z^k|\mathcal{E}, \mathbf{x}, \mathbf{v}, \eta^k = 1\right) \\ &\cdot P\left(\eta^k = 1|\lambda^B, \lambda^F\right) + P\left(z^k|\mathcal{E}, \mathbf{x}, \mathbf{v}, \eta^k = 0\right) \\ &\cdot P\left(\eta^k = 0|\lambda^B, \lambda^F\right). \end{aligned} \tag{20}$$

Since all K detections are conditionally independent given $\mathbf{x}$ and $\mathcal{E}$, and λ^B and λ^F are determined through position $\mathbf{x}$ and areas V plus O, the full measurement likelihood becomes:

$$P(\mathcal{Z}_t|\mathbf{h}_t) = P\left(K|\lambda^B, \lambda^F\right) \cdot \prod_{k=1}^{K} P\left(z^k|\mathbf{h}_t, \lambda^B, \lambda^F\right). \tag{21}$$

IV. Implementation

First, we describe how the Bayesian filter was implemented with a particle filter. Then, we discuss how the attribute likelihood function was implemented for the two sensors. A summary of the model parameters is given in Table I.

A. Particle Filtering

For inference, we use a particle filter to represent the posterior distribution in our model by a set of samples (i.e., particles). Unlike, say, a multiple-model Kalman Filter, it is straight-forward to include information about occlusion in the particle filter, i.e. particles in occluded areas are treated differently than those in unoccluded areas, and to represent uniform initial uncertainty over the bounded occlusion region.

Furthermore, such a system is easy to scale for the available hardware resources by changing the number of particles.

TABLE I
LIST OF MODEL PARAMETERS AND THEIR EXPERIMENTAL VALUE SETTINGS

Parameter	Short description	In our experiments
$p_s(\mathbf{h}_t^{(i)})$	Probability for a ped. to stay	0.95 in ROI
p_n	Probability of an entering ped.	0.2
$p_e(\mathbf{h}_t)$	Distrib. of $\mathbf{h}_t$ for an entering ped.	Uniform in ROI
λ^F_{unocc}	Exp. # of detections (unoccluded)	1/1.5 for cam./rad.
λ^F_{occ}	Exp. # of detections (occluded)	0.1/0.3 for cam./rad.
λ^B	Exp. # of background detections	0.05/0.1 for cam./rad.
$L_B(z_{pos})$	Spatial likelihood (background)	Uniform in ROI
$L_F(z_{pos}\|\mathbf{x})$	Spatial likelihood (foreground)	Eq. (32), Eq. (34)
$A_B(z_{attr})$	Attribute likelihood (background)	Eq. (33), Eq. (36)
$A_F(z_{attr}\|\mathbf{x},\mathbf{v})$	Attribute likelihood (foreground)	Eq. (33), Eq. (36)
Σ_{cx}, Σ_{rx}	std. dev. used in L_F	$0.2\ m, 0.3\ m$
$\Sigma^F_{ch}, \Sigma^F_{rv}$	std. dev. used in A_F	$0.7\ m, 0.8\ m/s$
$\Sigma^B_{ch}, \Sigma^B_{rv}$	std. dev. used in A_B	$1.5\ m, 3\ m/s$
W_{speed}	Exp. distrib. of particle speeds	$N(p, \Sigma_w)$
W_{dir}	Exp. distrib. of particle orientation	Uniform in $\pm 22.5°$

To include the existence probability in the filter, we follow [42]. Of N particles, the first one (index 0) will represent all hypotheses with non-present pedestrian, called the negative particle. The remaining $N-1=N_s$ particles (called the positive ones) represent the cases of a present pedestrian:

$$\mathcal{E}_t = 0 \rightarrow w_t^{(0)}, \tag{22}$$

$$\mathcal{E}_t = 1 \rightarrow \left(\mathbf{h}_t^{(i)}, w_t^{(i)}\right)\ for\, i = 1 \dots N_s. \tag{23}$$

where $\mathbf{h}_t^{(i)} = \left[\mathbf{x}_t^{(i)}, \mathbf{v}_t^{(i)}, \mathcal{E}_t^{(i)} = 1\right]$ is the state of the ith particle, $w_t^{(i)}$ is the weight assigned to it, and $\mathcal{E}_t^{(i)} = 1$ marks that these N_s particles represent hypotheses of a present pedestrian. Thus, the estimated probability of a non-present/existing pedestrian given all detections is the normalized weight of the first particle/summed weights of all the others, see Eq. (5):

$$P(\mathcal{E}_t = 0|\mathcal{Z}_{1:t}) = w_t^{(0)},\ P(\mathcal{E}_t = 1|\mathcal{Z}_{1:t}) = \sum_{i=1}^{N_s} w_t^{(i)}. \tag{24}$$

To obtain the estimated state of the pedestrian, we use the weighted average of the particles along the hypothesis space:

$$\widetilde{\mathbf{h}}_t = \left[\widetilde{\mathbf{x}}_t, \widetilde{\mathbf{v}}_t, \widetilde{\mathcal{E}}\right] = \sum_{i=1}^{N_s} w_t^{(i)} \cdot \mathbf{h}_t^{(i)}, \tag{25}$$

where $\widetilde{\mathbf{x}}_t = (\widetilde{x}_t, \widetilde{y}_t)$ is the estimated position, $\widetilde{\mathbf{v}}_t = (\widetilde{v}_{x,t}, \widetilde{v}_{y,t})$ is the estimated velocity vector of the pedestrian, and $\widetilde{\mathcal{E}}$ is the estimate of the pedestrian being present, see Eq. (24).

1) Initialization: Particles' positions are initialized uniformly across the Region of Interest (ROI). Their velocity is drawn from normal distribution $W_{speed} \sim N(p, \Sigma_w)$ around slow walking pace $p = 1\ m/s$ and their orientation is drawn from a uniform distribution W_{dir} between $\pm 22.5°$, where $0°$ is the orientation perpendicular to the movement of the ego-vehicle, pointing towards the road.

2) Prediction Step: The input of the prediction step are N_s uniformly weighted particles representing the present pedestrian, and one particle representing the $\mathcal{E}_t = 0$ hypothesis.

Predicted variables are marked with ^ sign. First, we estimate the next weight of the negative particle as follows:

$$P(\mathcal{E}_t = 0|\mathcal{Z}_{1:t-1}) = \hat{w}_t^{(0)} = \frac{w_{np}}{w_{np} + w_p}, \tag{26}$$

where w_p, w_{np} are the cumulative weights of present, and not present predicted states using Eqs. (7)–(10):

$$w_p = p_n \cdot w_{t-1}^{(0)} + \sum_{i=1}^{N_s} \left(p_s\left(\mathbf{h}_t^{(i)}\right)\right) w_{t-1}^{(i)}, \tag{27}$$

$$w_{np} = (1 - p_n) \cdot w_{t-1}^{(0)} + \sum_{i=1}^{N_s} \left(1 - p_s\left(\mathbf{h}_t^{(i)}\right)\right) w_{t-1}^{(i)}. \tag{28}$$

Afterwards, we sample N_s new positive particles, which are either a mutation of an existing particle moved by the dynamic model, or a completely new (entering) one, see Eq. (11). An existing particle stays in the scene with probability $p_s(\mathbf{h}_t^{(i)})$, or is replaced by a new one with probability of $1 - p_s(\mathbf{h}_t^{(i)})$:

$$\mathbf{h}_{t-1}^{(i)} \rightarrow \begin{cases} \hat{\mathbf{h}}_t^{(i)} \sim P\left(\mathbf{h}_t|\mathbf{h}_{t-1}^{(i)}\right) & \text{if moved particle,} \\ \hat{\mathbf{h}}_t^{(i)} \sim p_e(\mathbf{h}_t) & \text{if new particle.} \end{cases} \tag{29}$$

All weights of the predicted positive particles are then set uniformly:

$$\hat{w}_t^{(i)} = \frac{1 - \hat{w}_t^{(0)}}{N_s} \forall\, i = 1 \dots N_s. \tag{30}$$

3) Update Step: Particles are updated by new detections using the measurement likelihood Eq. (21):

$$w_t^{(i)} \propto \hat{w}_t^{(i)} \cdot P\left(\mathcal{Z}_t|\hat{\mathbf{h}}_t^{(i)}\right). \tag{31}$$

Details of the attribute likelihood calculations are discussed later in Subsection IV-B and IV-C. After the update, all weights are renormalized. To avoid sample degeneracy, we resample the positive particles if the Effective Sample Size (ESS) drops below a threshold [59].

B. Use of Stereo Camera Data

The camera sensor data is used for two purposes: 1) to update our filter with camera based pedestrian detections and 2) to update our occlusion model, see Fig. 2, top. For both tasks, we use the Instance Stixel representation [51]. Stixels [49] are rectangular upright sticks in the 3D space, perpendicular to the estimated ground plane. With the extension of [51], each stixel has the following parameters: a 3D position of their bottom, a height, a class label (among others: *car, bus, truck, person, sky*) and an instance id. In this way, objects of interest (pedestrians and occluding vehicles) are represented by a loose set of stixels connected by their class and instance information. Unlike the bounding box representation used in [43], these stixels better describe the shape and extent of objects in both bird's-eye and camera perspectives (e.g., varying visible height of cars) while keeping the processing load low. First we filter the stixels to keep only those from the relevant classes: *pedestrian* stixels as input for the particle filter and vehicle stixels (i.e. from *car*, *truck*, and *bus* classes) to update the occlusion model.

1) Update of the Occlusion Model: The stixels of vehicles that are close enough (i.e., at least one of their stixels is in ROI) are fitted with a bird's-eye view 2D rectangle to model the position and extent of the parked vehicles. The fitting is done with plausible minimum widths and lengths to avoid unrealistically small car assumptions. We consider the projected region behind the farther end of these car models as *occluded* as shown on Figs. 1 and 7. We also store the set of stixels for each car to calculate the height of the occlusion for later use, see below.

2) Update of the Filter: The pedestrian stixels are grouped by their instance id. Then, the average 2D BEV position of the stixels and their largest height range in meters (i.e., the difference between the lowest and highest stixels ends) are computed to create a pedestrian detection for the filter: $z = [z_{pos},\, z_{attr} = z_{height}]$. The position z_{pos} is then used in the spatial component, which is modeled with a normal distribution, with standard deviation Σ_{cx}:

$$L_F(z_{pos}|\mathbf{x}_t^{(i)}) = N(z_{pos}|\mathbf{x}_t^{(i)}, \Sigma_{cx}). \tag{32}$$

The height z_{height} is used to calculate the attribute likelihood $A_F(z_{attr}|\mathbf{x}_t^{(i)}, \mathbf{v}_t^{(i)})$. We consider the likelihood of observing a pedestrian with visible z_{height} at the location $\mathbf{x}_t^{(i)}$ of each particle, given the current occlusion model. First, we compute the expected observable height $\widetilde{h}_t^{(i)}$ for each occluded particle by looking up the car stixel with the most similar angle to it, see Fig. 1. Then, the height of this stixel is scaled by the distance of the particle to get how tall objects would be occluded by the stixel/parked car at the particle's location. Afterwards, the expected observable height $\widetilde{h}_t^{(i)}$ is the difference between the occluded height and the expected height of a pedestrian m_{height}. For example, behind a tall van we expect to see no part of a pedestrian ($\widetilde{h}_t^{(i)} = 0$), while at an unoccluded location the full height of the pedestrian should be visible $\widetilde{h}_t^{(i)} = m_{height}$.

Finally, we model both $A_F(z_{attr}|\mathbf{x}_t^{(i)}, \mathbf{v}_t^{(i)})$ and $A_B(z_{attr})$ as zero mean normal distributions with standard deviations Σ_{ch}^F and Σ_{ch}^B:

$$\begin{aligned} d_{height} &= z_{height} - \widetilde{h}_t^{(i)}, \\ A_F\left(z_{attr}|\mathbf{x}_t^{(i)}, \mathbf{v}_t^{(i)}\right) &= N\left(d_{height}|0, \Sigma_{ch}^F\right), \\ A_B(z_{attr}) &= N\left(d_{height}|0, \Sigma_{ch}^B\right). \end{aligned} \tag{33}$$

C. Use of Radar Data

Radar data is solely used as an input to our pedestrian detection filter. For an overview of radar specific steps, see Fig. 2, bottom. Our equipped radar outputs a sparse point cloud of reflections called *radar targets*. Each point has two spatial dimensions, range r and azimuth α, and a third dimension referred to as Doppler, which is the radial velocity v_{rel} of the target relative to the ego-vehicle. First, we perform ego-motion compensation for v_{rel}. That is, by eliminating the motion of the sensor that comes from both the translational and rotational movement of the ego-vehicle we get the *compensated radial velocity*, a signed scalar value denoted by v_r, describing the ego-motion compensated (i.e., absolute) radial velocity of the point. In a next step, we filter the reflections based on their *RCS* and v_r, i.e. we remove targets with very weak reflections or too low velocities to only keep ones that could potentially originate from a darting pedestrian. We also eliminate radar targets that are located in the rectangles fitted to the parked cars since a pedestrian cannot be present there, but the high reflectivity of these cars could yield a moving radar target in case of a faulty ego-motion compensation. The remaining reflections are considered as detections for the filter: $z = [z_{pos},\, z_{attr} = z_{vel} = v_r]$.

The position z_{pos} is then used in the spatial component and modeled with a normal distribution analogous to the camera, with standard deviation Σ_{rx}:

$$L_F(z_{pos}|\mathbf{x}_t^{(i)}) = N(z_{pos}|\mathbf{x}_t^{(i)}, \Sigma_{rx}). \tag{34}$$

In addition to the spatial distribution, the radar also has an attribute likelihood component $A_F(z_{attr}|\mathbf{x}_t^{(i)}, \mathbf{v}_t^{(i)})$. We consider the likelihood of observing a radar reflection with the measured radial velocity z_{vel} given the location and velocity of each particles. Let us define $\mathbf{los} = \mathbf{x}_t^{(i)} - \mathbf{x_{radar}}$ as the line-of-sight vector pointing from the radar to the particle. We calculate the expected radial velocity $\widetilde{v}_{r,t}^{(i)}$ as the particle's velocity $\mathbf{v}_t^{(i)}$'s projection to this vector (i.e., its radial component):

$$\widetilde{v}_{r,t}^{(i)} = \frac{\mathbf{los} \cdot \mathbf{v}_t^{(i)}}{\|\mathbf{los}\|}. \tag{35}$$

Finally, we model both $A_F(z_{attr}|\mathbf{x}_t^{(i)}, \mathbf{v}_t^{(i)})$ and $A_B(z_{attr})$ as zero mean normal distributions with standard deviations Σ_{rv}^F and Σ_{rv}^B:

$$\begin{aligned} d_{vel} &= z_{vel} - \widetilde{v}_{r,t}^{(i)}, \\ A_F(z_{attr}|\mathbf{x}_t^{(i)}, \mathbf{v}_t^{(i)}) &= N\left(d_{vel}|0, \Sigma_{rv}^F\right), \\ A_B(z_{attr}) &= N\left(d_{vel}|0, \Sigma_{rv}^B\right). \end{aligned} \tag{36}$$

V. Dataset

Our dataset was captured by our prototype vehicle [60] in Delft, the Netherlands. We recorded the output of a Continental 400 radar mounted behind the front bumper (2+1D: range, azimuth, Doppler, ∼13 Hz, ∼100 m range, ∼120° field of view), an IDS stereo camera (1936 × 1216 px, ∼10 Hz, 35 cm baseline) mounted on the windshield, a Velodyne HDL-64 LiDAR (64 layers, ∼10 Hz) scanner on the roof, and the ego-vehicle's odometry (Spatial Dual GNSS/INS/AHRS sensor and wheel odometry fused via an Unscented Kalman Filter, ∼30 Hz). All sensors were jointly calibrated following [61]. While the LiDAR data is not used in this paper, it will be made available for future work.

The dataset contains 501 recordings, each with a length between 8–20 seconds. In each recording, the ego-vehicle approaches or passes (at least) one parked vehicle with a pedestrian behind it. All recordings were performed in a real environment, with driving speeds suitable for the environment (mean: 4.0 m/s, std.: 0.57 m/s). The pedestrian either steps out from behind the parked vehicle ("*darting*" or "*walking*" sequences) or remains there ("*staying*" sequences). Participants were instructed which

Fig. 5. Examples of darting out pedestrians from our dataset.

Fig. 6. Example of annotated frames on a walking out sequence. We marked the first frames where (a) the pedestrian's head, (b) the body center, (c) one of the feet, and (d), full body is visible.

action to perform next, but were free to choose their walking speed during *darting*, or their activity (imitating e.g. phone call, bagging groceries, slight movement) during *staying* recordings. See Fig. 5 for examples of darting out pedestrians. Fifteen subjects with different heights participated in the experiment (mean: 178 cm, standard deviation: 8.5 cm). In total, more than 100 different parked vehicles were used as occlusion, ranging from passenger cars (partial occlusion) to vans (full occlusion).

The resulting dataset contains 249 walking and staying 252 sequences. For each sequence, we manually annotated its type (darting or staying), the pedestrian's height, the occluding vehicle's type (car or van) and some environment conditions (e.g., harsh lighting, leaves on the ground, etc.). We have also marked the first timestamps where a) the head, b) the body center, c) one of the feet, and d) the entire body of the pedestrian is visible, see Fig. 6. This allows a temporal alignment of the sequences and a better understanding of the visual occlusion in the case of different occluding vehicles.

VI. Experiments

In our experiments, we investigate how the fusion of stereo camera and radar sensors, and the incorporation of occlusion information help to detect darting out pedestrians. For this purpose, we compare the following methods: *naive camera*, *naive fusion*, *OAF camera*, and *OAF fusion*, where "naive"/"OAF" stands for naive/occlusion aware filtering. The *naive camera* and *naive fusion* are methods that use only the camera/both sensors to update the filter in a naive way, see Eq. (16). Similarly, *OAF camera* and *OAF fusion* use only camera/both sensors to update, but in an occlusion aware way, i.e., they are "occlusion aware filters", see Eq. (15). All four methods above use Instance Stixels (IS) as camera based pedestrian detections to update the filter, while *OAF camera* and *OAF fusion* also use Instance Stixels to model the occlusions. To study the benefits of the improvements introduced in this paper, we compare the methods above with the fusion method from our previous publication [43]: OAF^{SSD} *fusion*. This is also an occlusion aware filter fusing both sensors, similar to *OAF fusion*, but it uses the output of the Single Shot Detector (SSD) instead of Instance Stixels (IS) as camera based method. Further, in contrast to the other methods, OAF^{SSD} *fusion* does not use the attribute likelihood components introduced in Subsection IV-B and IV-C, only the spatial component. Note that unlike IS, SSD provides detections as bounding boxes, not involving the height profile of the cars. Hence, the height related attribute likelihood component would not be possible to calculate with SSD. An overview of the compared methods is given in Table II. Both the "Filtering" and the "Postprocessing" module from Fig. 2 (including the presented application example) run at a processing speed of over 500 Hz for all methods with 1000 particles in an optimized Python based implementation using the Robot Operating System (ROS) on a high-end system PC (64 GB RAM, TITAN X (Pascal) GPU, Intel Xeon CPU E5-1560 CPU). This brings a negligible overhead compared to the camera based detection modules (off-the-shelf implementation of SSD and IS, including the occlusion model) running around 14 Hz, and the radar related preprocessing steps running at over 200 Hz.

Our framework has a set of parameters and distributions that should be tuned to the characteristics (type, accuracy, noise, etc.) of the user's sensors. A brief overview of these can be found in Table I. In this research, the parameters were empirically tuned on the distinct dataset used in [43] and during in-vehicle experiments, and visually validated on the first few sequences of the new dataset. ROI was defined as a 4.5 m wide, 14 m long rectangle in front of the ego-vehicle. For the camera, we use $\lambda^{F}_{unocc} = 1$ because detection is reliable in this range in

TABLE II
Overview of the Compared Methods With Whether They Use Radar, Type of Camera Based Method (IS: Instance Stixels, SSD: Single Shot Detector), Whether They are Occlusion Aware, and Whether They Implement the Attribute Likelihood Components

Method	Radar	Camera based method	Occlusion awareness	Attribute likelihood component
naive camera	✗	IS	✗	✓
OAF camera	✗	IS	✓	✓
naive fusion	✓	IS	✗	✓
OAF^{SSD} *fusion* [43]	✓	SSD	✓	✗
OAF fusion (proposed)	✓	IS	✓	✓

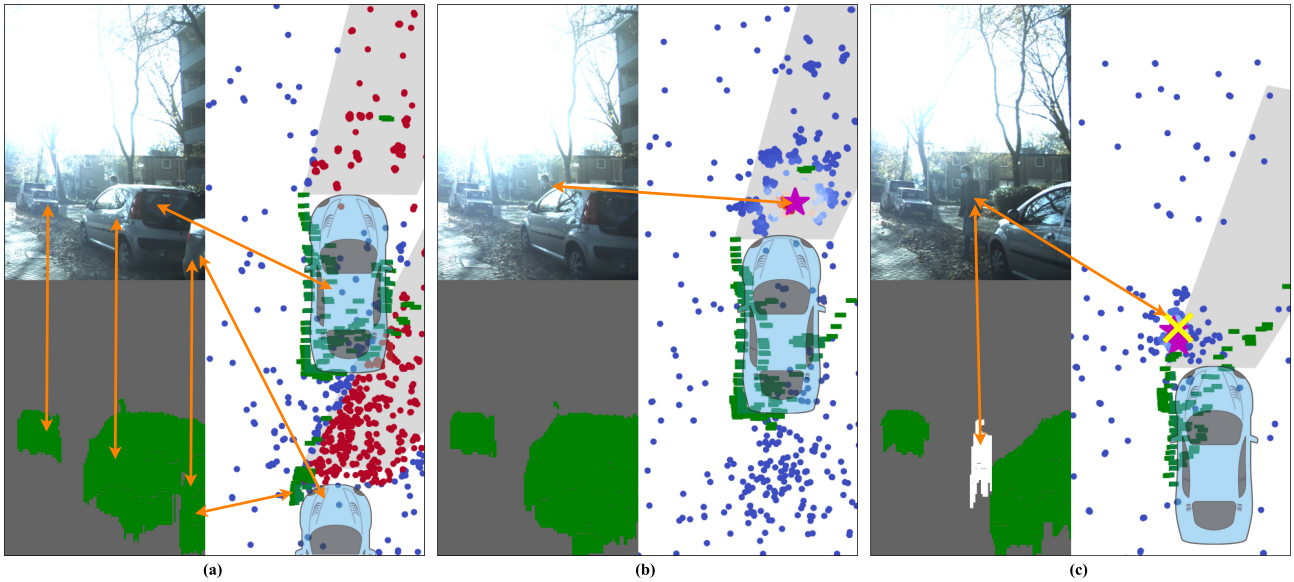

Fig. 7. Camera views (top left), stixel images (bottom left) and top views (right) of the scene at consecutive timestamps using *OAF fusion*. Vehicle/pedestrian stixels are shown with green/white colors (bottom left). Vehicle stixels are also shown as short green lines on the top view, representing the outlines of the detected parked cars. Occlusion (greyish areas) is calculated as the "shadow" of these cars. Initially, the particles (blue to red, for small to high relative weights) have higher density and weights in occluded regions (a) and converge on the pedestrian's position after being detected first by the radar (magenta star, (b)) and then by the camera sensor (yellow 'x', (c)). All views are cropped for easier visibility. Orange arrows connect corresponding objects in different views.

unoccluded regions. λ^F_{occ} is set to 0.1 for occluded locations. Few false positives occur in the ROI, so λ^B is set to 0.05. For radar, we set $\lambda^F_{unocc} = 1.5$ for unoccluded positions, since multiple reflections are often received from the same pedestrian. In occlusions, we set $\lambda^F_{unocc} = 0.3$ as we still expect some reflections due to the multipath propagation. An average rate of $\lambda^B = 0.1$ is expected for radar, as false positives occur more often than with camera due to e.g. incorrect ego-motion compensation.

A. Estimated Existence Probability in Dangerous Situations

In our first experiment, we ran the methods on all walking sequences and recorded the reported existence probabilities as in Eq. (5). The sequences were temporally aligned by marking the first moment when the pedestrian's body center was visible as $t = 0$, see Fig. 6. Then, for each timestamp, and for each method, we calculate the mean estimated probability by averaging over all walking sequences as in [43], see Fig. 8. In general, the inclusion of radar helps to detect the pedestrian earlier. I.e., any chosen threshold of probability is reached earlier by the three fusion methods (*naive fusion*/*OAF fusion*/*OAF^{SSD} fusion*) using both sensors, than by the methods using only the camera. For example, on average, the threshold $P(\mathcal{E}_t = 1|\mathcal{Z}_{1:t}) = 0.5$ is reached 0.26 seconds earlier by *OAF fusion* than by *naive camera*. When examining only smaller occluding vehicles (i.e., cars), this time gain increases to 0.30 s. In contrast, for sequences with a van as an occlusion, the measured gain is only 0.12 s.

The previously discussed threshold of $P(\mathcal{E}_t = 1|\mathcal{Z}_{1:t}) = 0.5$ is reached 0.15 seconds earlier by our proposed occlusion aware fusion *OAF fusion* than by the naive method *naive fusion*. *OAF fusion* also reports higher probabilities at all times when the pedestrian is occluded ($t < 0$). *OAF^{SSD} fusion* reaches the same

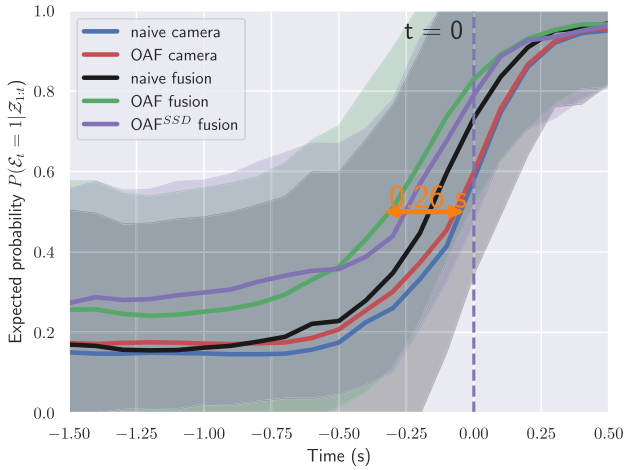

Fig. 8. Estimated probabilities of a pedestrian being present, averaged over all walking sequences, with standard deviation around the mean for fusion methods. $t = 0$ is the first moment when the pedestrian's body center was visible. The addition of radar results in earlier detection than using the camera alone.

threshold later than *OAF fusion* by 0.06 seconds, but still earlier than *naive fusion*.

We also examine the sequences individually, and calculate the time difference between the reported probabilities of *naive camera* and *OAF fusion* to be over 0.5. A histogram of the gained reaction times can be found in Fig. 9. In the large majority ($\sim$68%) of dangerous scenarios *OAF fusion* gains some additional reaction time over *naive camera*.

In Fig. 7, we show an example of a walking scene to demonstrate how *OAF fusion* behaves when there has been no prior detection, and then when first the radar and then the camera has detected the pedestrian.

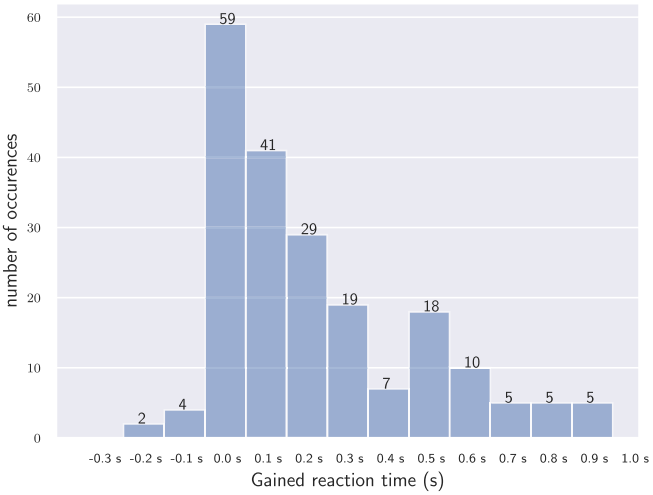

Fig. 9. Histogram of the gained reaction times. Time difference is calculated between the moments *naive camera* and *OAF fusion* reaches the threshold $P(\mathcal{E}_t = 1|\mathcal{Z}_{1:t}) = 0.5$. For clarity, here we only show the sequences where both methods reach the threshold within the time window of $[-1\,s,\ 0.5\,s]$.

B. Distinguishing Dangerous and Non-Dangerous Scenarios

Similar to [2] we classify the scene into two classes: $c = darting$ (there is a darting pedestrian, dangerous scenario) or $c = non-darting$ (there is no pedestrian, or he/she is not darting). To do this, we estimate the probability $P(\mathcal{E}_t = 1|\mathcal{Z}_{1:t})$ of a present pedestrian of any kind (staying or darting) by Eq. (24). We also estimate whether the assumed-to-be present pedestrian darts out creating a dangerous scenario $P(c = darting|\mathcal{Z}_{1:t}, \mathcal{E} = 1)$ based on the estimated state of the pedestrian $\widetilde{\mathbf{h}}_t$, see Eq. (25). The pedestrian is assumed to be darting if he/she is already on the road in front of the ego-vehicle: $\widetilde{x}_t > dangerousPos$ (axis is perpendicular to the movement of the ego-vehicle, increases towards the road), or he/she has a lateral velocity component large enough to assume he/she will be on the road later: $\widetilde{v}_{x,t} > dangerousSpeed$. Similarly, we assume that the pedestrian will not dart out if he/she is far enough from the road: $\widetilde{x}_t < safePos$, or their lateral velocity is close to zero/pointing away from the road: $\widetilde{v}_{x,t} < safeSpeed$. Probabilities for values between these limits ($dangerousPos > \widetilde{x}_t > safePos$ and $dangerousSpeed > \widetilde{v}_{x,t} > safeSpeed$) are linearly interpolated. We evaluate the probability of darting based on these two conditions (spatial and velocity) independently, and then take the maximum of the two values for safety. Finally, the probability of a present, darting pedestrian is calculated by multiplying the two probabilities: $P(\mathcal{E}_t = 1,\ c = darting|\mathcal{Z}_{1:t}) = P(\mathcal{E}_t = 1|\mathcal{Z}_{1:t}) \cdot P(c = darting|\mathcal{Z}_{1:t}, \mathcal{E} = 1)$. In the staying scenarios, the pedestrian's body center was not always visible during the recording as the pedestrian may have remained hidden completely. Hence, unlike for walking scenes, we marked the last moment the occluding vehicle was still visible as $t = 0$, to represent the moment when the ego-vehicle passes the occluding, parked vehicle. For each timestamp, we average the probability of a darting pedestrian $P(\mathcal{E}_t = 1,\ c = darting|\mathcal{Z}_{1:t})$ for walking and staying scenes separately, see Fig. 10. For the walking cases, the fusion methods (*naive fusion*, *OAF fusion* and *OAF^{SSD} fusion*) report higher probabilities of a darting pedestrian earlier, see Fig. 10(a). When evaluating the non-dangerous staying scenarios, all methods report a small, but moderately increased probability of a darting pedestrian in the moments before the occluding vehicle and the staying pedestrian are passed, and significantly decreased probabilities after the drive-by, see Fig. 10(b).

VII. Discussion

The benefits of including radar in darting out pedestrian detection has been shown in Subsection VI-A, where all fusion methods reacted earlier than the methods using only the camera. Such an earlier detection may mean additional reaction time in case of a dangerous situation. One cause is that radars can often detect pedestrians behind parked vehicles, as their reflected radar signal may be able to propagate under the occluding vehicle and reach the sensor. Some of the gains could also be the result of cases when the camera was not able to detect the already visible pedestrian (e.g. caused by harsh lighting), but the radar was. Our radar is mounted on the front of the ego-vehicle, as is common in the industry, see Fig. 1. This could provide a slightly better viewing angle and also contribute to the earlier detections.

The reaction time gained was significantly greater for smaller occluding vehicles such as cars than for vans (0.30 s vs 0.12 s for threshold $P(\mathcal{E}_t = 1|\mathcal{Z}_{1:t}) = 0.5$). This difference may have been caused by the length of these vehicles. Vans tend to be longer than cars, which can affect the propagation of the radar signal under the vehicle. This suggests that it may be beneficial to also estimate the length of the parked vehicle and explicitly include it in the fusion pipeline (i.e., expect fewer reflections behind longer vehicles).

The benefits of occlusion awareness become clear when we compare the naive methods (*naive camera*, *naive fusion*) with their occlusion aware pairs (*OAF camera*, *OAF fusion*). For example, *OAF fusion* reports a higher probability of a pedestrian being present than *naive fusion* at all times when the pedestrian is occluded ($t < 0$). The reason for this is twofold. First, *OAF fusion* is occlusion aware, and thus it "acknowledges" that parts of the scenes are occluded and cannot be properly observed, leading to uncertainty. That is, the absence or low number of detections from these areas is not considered hard evidence for the absence of a pedestrian, unlike in naive methods, e.g. *naive fusion*. Instead, particles behind occlusions are weighted higher compared to the unoccluded particles to represent this uncertainty, resulting in higher a priori awareness to these locations even before any detections occur, see Fig. 7, left. Similarly, this elevated a priori awareness of an occlusion aware method is also observable between *OAF camera* and *naive camera* for $t < 0$ moments. Such "caution" resembles the behavior of a human driver approaching highly occluded regions where pedestrians might be. Second, detections originating from these occluded regions are valued more than in the naive methods, because the number of detections received better fits the expectations in Eq. (17). As a result, the likelihoods are higher for the same detections than when processed by a naive method, e.g. *naive fusion*, see Eq. (21).

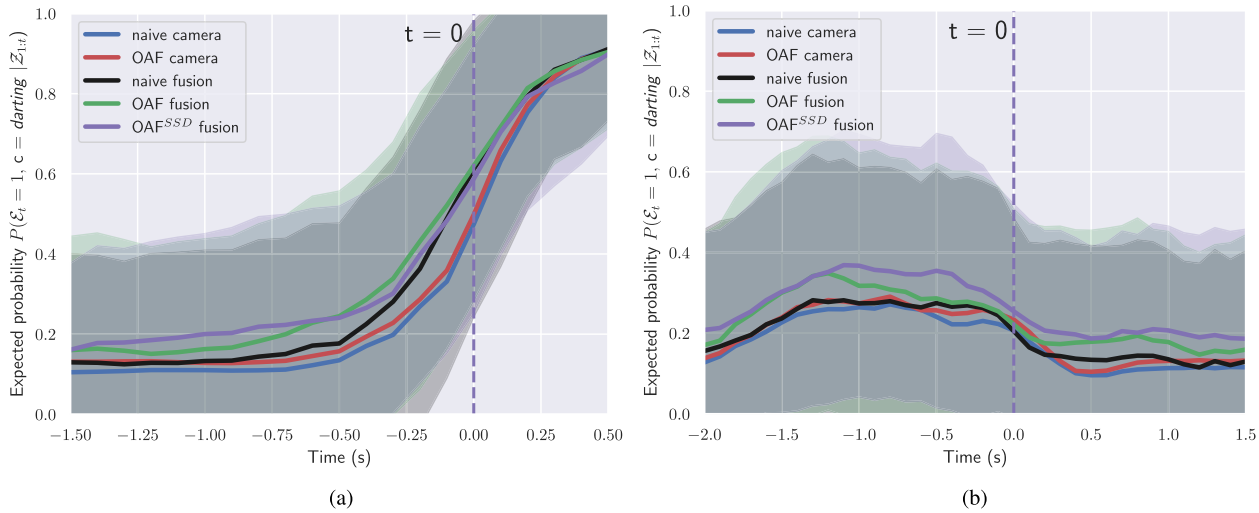

Fig. 10. Estimated probabilities of a darting pedestrian, averaged over all walking (a) and staying (b) sequences, with standard deviation around the mean for fusion methods. $t = 0$ is the first moment when the pedestrian's body center was visible for walking scenes, and the last moment when the occluding vehicle was visible for staying scenes. For walking scenes (a), the addition of radar results in earlier detection than using the camera alone. For staying scenes (b), all methods reports slightly increased, but still small probabilities of danger (i.e. darting) before the passing.

The occlusion aware fusion approach presented in this paper, *OAF fusion*, responded earlier to darting out pedestrians than its older version *OAF^{SSD} fusion* from [43]. The reason for this, we believe, is twofold. First, as described in Section II, the occlusion model used by *OAF^{SSD} fusion* was often inaccurate. A more accurate model of the occlusions (see Subsubsection IV-B1) helped to better evaluate measurements in this study. Although this occlusion model was created using stixels, this improvement could also be achieved using other methods to obtain more accurate occlusion information. Second, this work introduced the concept of attribute likelihood components. More specifically, for camera based detections, even small patches of detections were accepted as reasonable, valid measurements if they matched our occlusion model. This was also supported by the decision to use instance segmentation as input instead of standard object detection, since the former tends to provide more partial detections, which suits our use-case. In the case of radar, the attribute component meant comparing expected and observed radial velocities. This filters out unrealistic radar targets, which could originate from other road users or simply from noise. On the other hand, radar detections that matched our prior expectations of the object's motion were highly valued and increased the probability earlier. It is noteworthy, however, that *OAF^{SSD} fusion* still responded earlier than *naive fusion*, suggesting that even its simpler, SSD based occlusion model benefited more than the attribute likelihood components and the use of stixels as camera based detection. This means that, depending on the application and available resources, using a simpler occlusion model (i.e., SSD instead of IS) could be satisfactory with the benefit of reduced computational load.

In our second experiment, we presented an example application of our methods to distinguish dangerous and non-dangerous situations. For scenes where the pedestrian remains behind the car (i.e. not in danger), the estimated probabilities somewhat increase during the drive-by, but remain small. This observed increase can be explained by the way the particles are initialized with a walking speed and an movement orientation pointing to the road, which intentionally introduces a bias towards the darting hypothesis. Such increase in uncertainty about whether the sighted pedestrian will dart out is similar to the reaction of a human driver, who, having noticed a pedestrian in a similar situation, would also slow down/be more cautious for safety reasons. The occlusion aware fusion methods (*OAF fusion* and *OAF^{SSD} fusion*) show further increased caution due to perceived occlusions in the scene, which increase a priori uncertainty by design. *OAF fusion*, however, shows lower estimated probabilities for all $t < 0$ timestamps than *OAF^{SSD} fusion*, again suggesting that the new occlusion model based on Instance Stixels is superior to the one based on SSD, and follows the shape of the occlusions more closely.

All filters depend heavily on the quality of the inputs, especially from the camera, where we expect "high-end" detections (e.g. pedestrian instance stixels) from an off-the-shelf module, even under occlusion. The quality of camera based detections also affects the reliability of the filter over the occlusion model. Common errors arise from radar targets that are incorrectly reported as moving by the radar due to poor ego-motion estimation, and from camera based detections that mistake vertically shaped objects (e.g., trees) for pedestrians.

The proposed system can be further improved in several ways. For example, an additional use of the occlusion model would be to adjust the expected background noise for the radar. That is, instead of uniform distribution, it might be beneficial to increase the expected noise near parked vehicles with highly reflective metallic chassis, and decrease it in uncluttered regions.

Integrating additional sensors, (e.g., LiDAR) into our framework is straightforward. In particular, replacing or supporting the 2+1D radar used in this paper with a 3+1D radar similar to that used in [31] could be interesting for three reasons. First, the elevation information and increased density of the radar point

cloud could be used in a more advanced pedestrian classification step, as shown in [31]. Second, the elevation information could be further used in this particular use case by filtering the radar targets based on their elevation angle, leaving only those that are received from below the parked, occluding vehicle - as these targets could be the result of multi-path propagation. This step would help filter out false positive radar reflections that originate from the chassis of parked cars and not from occluded pedestrians. Third, in [31] the 3+1D radar has been shown to be capable of detecting both moving and parked vehicles. As such, it could contribute directly to the occlusion model and reduce or even eliminate the need for the camera sensor.

To generalize the filter for other road users, one has to adjust the prior velocity and RCS values, e.g., faster and more reflective targets should be expected from a cyclist. For the camera based detectors (IS, SSD), the expected class of object has to be changed. Multiple road users can also be tracked with the filter by modifying the state estimation step in Eq. (25) to expect more than one peak in the particle distribution. Consideration of objects other than vehicles as occlusions, e.g., walls, is also possible, and the observed visible height should be treated as in this study. However, the type of occlusion must be considered for radar, since multipath propagation is not possible if the occlusion has no space under it, such as walls.

Finally, we did state estimation in this research and showed quantitative benefits of both fusion and occlusion awareness. However, extending the scope to trajectory prediction, the gained reaction times detecting/predicting dangerous situations could be even greater.

VIII. Conclusions and Future Work

In this paper we proposed a generic occlusion aware multi-sensor Bayesian filter to detect occluded crossing pedestrians. To facilitate our and future research of these scenarios, we publish our dataset of more than 500 relevant scenarios with stereo camera, radar, LiDAR, and odometry data. We applied the proposed filter to camera and radar data using this dataset, and provided techniques to account for the unique characteristics of these sensors. Our results show that both the inclusion of radar sensor and occlusion information is beneficial for this use case, as pedestrians are detected earlier in dangerous walking scenarios. For example, the threshold of 0.5 for the estimated existence probability of a pedestrian in the scene is reached on average 0.26 seconds earlier by our occlusion aware fusion than by a naive camera only detector, and 0.15 seconds earlier than by the method that fuses the two sensors in a naive way.

We also showed in an application example of our filter that it can distinguish between dangerous and non-dangerous situations, which is necessary to avoid false alarms. In this task, too, the inclusion of the radar proved to be beneficial.

Future work may include a more precise expected distribution of background noise, improved scene classification by extending the scope for trajectory prediction, and the inclusion of further sensors, more particularly a 3+1D radar as discussed Section VII.

References

[1] World Health Organization, "Global status report on road safety," 2018. [Online]. Available: https://www.who.int/publications/i/item/9789241565684

[2] C. G. Keller and D. M. Gavrila, "Will the pedestrian cross? A study on pedestrian path prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 2, pp. 494–506, Apr. 2014.

[3] A. Brunetti, D. Buongiorno, G. F. Trotta, and V. Bevilacqua, "Computer vision and deep learning techniques for pedestrian detection and tracking: A survey," *Neurocomputing*, vol. 300, pp. 17–33, 2018.

[4] M. Braun, S. Krebs, F. Flohr, and D. M. Gavrila, "EuroCity persons: A novel benchmark for person detection in traffic scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1844–1861, Aug. 2019.

[5] S. Heuel and H. Rohling, "Pedestrian recognition in automotive radar sensors," in *Proc. IEEE Int. Radar Symp.*, 2013, pp. 732–739.

[6] A. Palffy, J. Dong, J. F. P. Kooij, and D. M. Gavrila, "CNN based road user detection using the 3D radar cube," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 1263–1270, Apr. 2020.

[7] O. Schumann, M. Hahn, J. Dickmann, and C. Wöhler, "Semantic segmentation on radar point clouds," in *Proc. IEEE Int. Conf. Inf. Fusion*, 2018, pp. 2179–2186.

[8] K. Granström, S. Reuter, M. Fatemi, and L. Svensson, "Pedestrian tracking using velodyne data — stochastic optimization for extended object tracking," in *Proc. IEEE Intell. Veh. Symp.*, 2017, pp. 39–46.

[9] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12689–12697.

[10] R. Streubel and B. Yang, "Fusion of stereo camera and MIMO-FMCW radar for pedestrian tracking in indoor environments," in *Proc. IEEE Int. Conf. Inf. Fusion*, 2016, pp. 565–572.

[11] J. Schlosser, C. K. Chow, and Z. Kira, "Fusing LIDAR and images for pedestrian detection using convolutional neural networks," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2016, pp. 2198–2205.

[12] B. Bartels and H. Liers, "Bewegungsverhalten von fußgängern im straßenverkehr," *FAT-Schriftenreihe*, vol. 268, no. 2, pp. 1–59, 2014 .

[13] European New Car Assessment Programme, "Test protocol - AEB VRU systems," 2020 [Online]. Available: https://cdn.euroncap.com/media/58226/euro-ncap-aeb-vru-test-protocol-v303.pdf

[14] R. Sherony and C. Zhang, "Pedestrian and bicyclist crash scenarios in the U.S.," in *Proc. IEEE Conf. Intell. Transp. Syst.*, 2015, pp. 1533–1538.

[15] A. Bartsch, F. Fitzek, and R. H. Rasshofer, "Pedestrian recognition using automotive radar sensors," *Adv. Radio Sci.*, vol. 10, pp. 45–55, 2012.

[16] N. Scheiner et al., "Seeing around street corners: Non-line-of-sight detection and tracking in-the-wild using doppler radar," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2065–2074.

[17] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.

[18] C. Ning, L. Menglu, Y. Hao, S. Xueping, and L. Yunhong, "Survey of pedestrian detection with occlusion," *Complex Intell. Syst.*, vol. 7, no. 1, pp. 577–587, 2021.

[19] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila, "Multi-cue pedestrian classification with partial occlusion handling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 990–997.

[20] Y. Tian, P. Luo, X. Wang, and X. Tang, "Deep learning strong parts for pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1904–1912.

[21] C. Zhou and J. Yuan, "Learning to integrate occlusion-specific detectors for heavily occluded pedestrian detection," in *Proc. Asian Conf. Comput. Vis.*, 2017, pp. 305–320.

[22] C. Zhou and J. Yuan, "Multi-label learning of part detectors for heavily occluded pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3506–3515.

[23] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion loss: Detecting pedestrians in a crowd," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7774–7783.

[24] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Occlusion-aware R-CNN: Detecting pedestrians in a crowd," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 657–674.

[25] M. Braun, F. B. Flohr, S. Krebs, U. Kreße, and D. M. Gavrila, "Simple pair pose - pairwise human pose estimation in dense urban traffic scenes," in *Proc. IEEE Intell. Veh. Symp.*, 2021, pp. 1545–1552.

[26] X. Wang, A. Shrivastava, and A. Gupta, "A-Fast-RCNN: Hard positive generation via adversary for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3039–3048.

[27] O. Schumann, M. Hahn, J. Dickmann, and C. Wöhler, "Comparison of random forest and long short-term memory network performances in classification tasks using radar," in *Proc. Sensor Data Fusion: Trends, Solutions, Appl.*, 2017, pp. 1–6.
[28] R. Prophet et al., "Pedestrian classification with a 79 GHz automotive radar sensor," in *Proc. IEEE Int. Radar Symp.*, 2018, pp. 1–6.
[29] R. Pérez, F. Schubert, R. Rasshofer, and E. Biebl, "Single-frame vulnerable road users classification with a 77 GHz FMCW radar sensor and a convolutional neural network," in *Proc. IEEE Int. Radar Symp.*, 2018, pp. 1–10.
[30] A. Danzer, T. Griebel, M. Bach, and K. Dietmayer, "2D car detection in radar data with PointNets," in *Proc. IEEE Conf. Intell. Transp. Syst.*, 2019, pp. 61–66.
[31] A. Palffy, E. Pool, S. Baratam, J. F. P. Kooij, and D. M. Gavrila, "Multi-class road user detection with 3+ 1D radar in the view-of-delft dataset," *IEEE Robot. Automat. Lett.*, vol. 7, no. 2, pp. 4961–4968, Apr. 2022.
[32] S. Heuel and H. Rohling, "Pedestrian classification in automotive radar systems," in *Proc. IEEE Int. Radar Symp.*, 2012, pp. 39–44.
[33] M. Heuer, A. Al-Hamadi, A. Rain, and M. M. Meinecke, "Detection and tracking approach using an automotive radar to increase active pedestrian safety," in *Proc. IEEE Intell. Veh. Symp.*, 2014, pp. 890–893.
[34] S. Hayashi, K. Saho, D. Isobe, and M. Masugi, "Pedestrian detection in blind area and motion classification based on rush-out risk using micro-doppler radar," *Sensors*, vol. 21, no. 10, 2021, Art. no. 3388.
[35] M. P. Muresan, I. Giosan, and S. Nedevschi, "Stabilization and validation of 3D object position using multimodal sensor fusion and semantic segmentation," *Sensors*, vol. 20, no. 4, 2020, Art. no. 1110.
[36] R. O. Chavez-Garcia and O. Aycard, "Multiple sensor fusion and classification for moving object detection and tracking," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 2, pp. 525–534, Feb. 2016.
[37] S. Chadwick, W. Maddern, and P. Newman, "Distant vehicle detection using radar and vision," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 8311–8317.
[38] J. Nie, J. Yan, H. Yin, L. Ren, and Q. Meng, "A multimodality fusion deep neural network and safety test strategy for intelligent vehicles," *IEEE Trans. Intell. Veh.*, vol. 6, no. 2, pp. 310–322, Jun. 2021.
[39] S. K. Kwon, E. Hyun, J.-H. Lee, J.-H. Lee, and S. H. Son, "Detection scheme for a partially occluded pedestrian based on occluded depth in lidar-radar sensor fusion," *Opt. Eng.*, vol. 56, no. 11, 2017, Art. no. 113112.
[40] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila, "Context-based pedestrian path prediction," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 618–633.
[41] A. Angelov, A. Robertson, R. Murray-Smith, and F. Fioranelli, "Practical classification of different moving targets using automotive radar and deep neural networks," *IET Radar, Sonar Navigation*, vol. 12, no. 10, pp. 1082–1089, 2018.
[42] S. Munder, C. Schnörr, and D. M. Gavrila, "Pedestrian detection and tracking using a mixture of view-based shape-texture models," *IEEE Trans. Intell. Transp. Syst.*, vol. 9, no. 2, pp. 333–343, Jun. 2008.
[43] A. Palffy, J. F. P. Kooij, and D. M. Gavrila, "Occlusion aware sensor fusion for early crossing pedestrian detection," in *Proc. IEEE Intell. Veh. Symp.*, 2019, pp. 1768–1774.
[44] A. Almeida, J. Almeida, and R. Araújo, "Real-time tracking of moving objects using particle filters," in *Proc. IEEE Int. Symp. Ind. Electron.*, 2005, pp. 1327–1332.
[45] Z. Radosavljević, D. Mušicki, B. Kovačević, W. C. Kim, and T. L. Song, "Integrated particle filter for target tracking in clutter," *IET Radar, Sonar Navigation*, vol. 9, no. 8, pp. 1063–1069, 2015.
[46] S. Hoermann, P. Henzler, M. Bach, and K. Dietmayer, "Object detection on dynamic occupancy grid maps using deep learning and automatic label generation," in *Proc. IEEE Intell. Veh. Symp.*, 2018, pp. 826–833.
[47] D. Nuss, T. Yuan, G. Krehl, M. Stübler, S. Reuter, and K. Dietmayer, "Fusion of laser and radar sensor data with a sequential monte carlo bayesian occupancy filter," in *Proc. IEEE Intell. Veh. Symp.*, 2015, pp. 1074–1081.
[48] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
[49] H. Badino, U. Franke, and D. Pfeiffer, "The stixel world - A compact medium level representation of the 3D-World," in *Lecture Notes in Comput. Sci.*, Berlin, Heidelberg, Germany: Springer, 2009, pp. 51–60.
[50] L. Schneider et al., "Semantic stixels: Depth is not enough," in *Proc. IEEE Intell. Veh. Symp.*, 2016, pp. 110–117.
[51] T. Hehn, J. Kooij, and D. Gavrila, "Fast and compact image segmentation using instance stixels," *IEEE Trans. Intell. Veh.*, vol. 7, no. 1, pp. 45–56, Mar. 2022.
[52] G. Kim, Y. S. Park, Y. Cho, J. Jeong, and A. Kim, "MulRan: Multimodal range dataset for urban place recognition," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 6246–6253.
[53] D. Barnes, M. Gadd, P. Murcutt, P. Newman, and I. Posner, "The oxford radar RobotCar dataset: A radar extension to the oxford RobotCar dataset," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 6433–6438.
[54] J. Bai, L. Zheng, S. Li, B. Tan, S. Chen, and L. Huang, "Radar transformer: An object classification network based on 4D MMW imaging radar," *Sensors*, vol. 21, no. 11, 2021, Art. no. 3854.
[55] O. Schumann et al., "RadarScenes: A. real-world radar point cloud data set for automotive applications," in *Proc. IEEE Int. Conf. Inf. Fusion*, 2021, pp. 1–8.
[56] H. Caesar et al., "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11618–11628.
[57] M. Mostajabi, C. M. Wang, D. Ranjan, and G. Hsyu, "High resolution radar dataset for semi-supervised learning of dynamic objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 450–457.
[58] M. Meyer and G. Kuschk, "Automotive radar dataset for deep learning based 3D object detection," in *Proc. IEEE Eur. Radar Conf.*, 2019, pp. 129–132.
[59] T. Li, S. Sun, T. P. Sattar, and J. M. Corchado, "Fight sample degeneracy and impoverishment in particle filters: A review of intelligent approaches," *Expert Syst. Appl.*, vol. 41, no. 8, pp. 3944–3954, 2014.
[60] L. Ferranti et al., "SafeVRU: A research platform for the interaction of self-driving vehicles with vulnerable road users," in *Proc. IEEE Intell. Veh. Symp.*, 2019, pp. 1660–1666.
[61] J. Domhof, J. F. P. Kooij, and D. M. Gavrila, "A joint extrinsic calibration tool for radar, camera and LIDAR," *IEEE Trans. Intell. Veh.*, vol. 6, no. 3, pp. 571–582, Sep. 2021.

Andras Palffy (Member, IEEE) received the M.Sc. degree in computer science engineering from Pazmany Peter Catholic University, Budapest, Hungary, in 2016, and the M.Sc. degree in digital signal and image processing from Cranfield University, Cranfield, U.K., in 2015. He is currently working toward the Ph.D. degree with Delft University of Technology, Delft, Netherlands, focusing on radar based vulnerable road user detection for autonomous driving. From 2013 to 2017, he was with Startup Eutecus, developing computer vision algorithms for traffic monitoring and driver assistance applications.

Julian F. P. Kooij (Member, IEEE) received the Ph.D. degree in artificial intelligence from the University of Amsterdam, Amsterdam, Netherlands, in 2015. In 2013, he was with Daimler AG worked on path prediction for vulnerable road users. In 2014, he joined Computer Vision Lab, Delft University of Technology (TU Delft), Delft, Netherlands. Since 2016, he has been with Intelligent Vehicles Group, part of the Cognitive Robotics Department, TU Delft, where he is currently an Associate Professor. His research interests include probabilistic models and machine learning techniques to infer and anticipate critical traffic situations from multi-modal sensor data.

Dariu M. Gavrila (Member, IEEE) received the Ph.D. degree in computer science from the University of Maryland, College Park, MD, USA, in 1996. From 1997, he was with Daimler R&D, Ulm, Germany, where he became a Distinguished Scientist. 2016, he moved to Delft University of Technology, Delft, Netherlands, where he since Heads the Intelligent Vehicles Group as a Full Professor. His research interests include sensor-based detection of humans and analysis of behavior, recently in the context of the self-driving cars in urban traffic. He was the recipient of the Outstanding Application Award 2014 and the Outstanding Researcher Award 2019, from the IEEE Intelligent Transportation Systems Society.