Occlusion aware sensor fusion for early crossing pedestrian detection

Andras Palffy [a,1], Julian F. P. Kooij [a,2] and Dariu M. Gavrila [a,3]

***Abstract*— Early and accurate detection of crossing pedestrians is crucial in automated driving to execute emergency manoeuvres in time. This is a challenging task in urban scenarios however, where people are often occluded (not visible) behind objects, e.g. other parked vehicles. In this paper, an occlusion aware multi-modal sensor fusion system is proposed to address scenarios with crossing pedestrians behind parked vehicles. Our proposed method adjusts the detection rate in different areas based on sensor visibility. We argue that using this occlusion information can help to evaluate the measurements. Our experiments on real world data show that fusing radar and stereo camera for such tasks is beneficial, and that including occlusion into the model helps to detect pedestrians earlier and more accurately.**

I. INTRODUCTION

Densely populated urban areas are challenging locations for automated driving. One of the most critical reasons for that is the high number of Vulnerable Road Users (VRUs): pedestrians, cyclists, and moped riders. VRUs' locations are loosely regulated and they can change their speed and heading rapidly, which makes their detection and tracking complicated. At the same time, they are at high risk in case of a potential collision. E.g., of the approximately 1.3 million road traffic deaths every year, more than half are VRUs [1]. One particularly dangerous scenario is when a pedestrian crosses in front of the vehicle: 94% of injured pedestrians between 2010 and 2013 in the US were hit after such behaviour [2].

Detecting crossing pedestrians as early as possible is crucial since it leaves more time to execute emergency braking or steering. To plan such a manoeuvre, precise location and trajectory of the VRU is also needed. Thus, a pedestrian detection system has two aims: early detection and accurate tracking of the VRU. Intelligent vehicles have several sensors to cope with this task: cameras [3], [4], [5], LIDARs, [6], and radars [7], [8], [9] have been used. Fusing different type of sensors, e.g. radar with camera [10] or LIDAR with camera [11] can add reliability and redundancy to such systems.

Unfortunately, pedestrian detection is often complicated in urban scenarios by severe occlusions, e.g. by parked vehicles. Consider the scene in Figure 1. The ego-vehicle (white) is passing a parked vehicle (blue). A pedestrian, partly or fully occluded by the parked vehicle is stepping out onto the road in front of the approaching ego-vehicle. Such a behaviour of the VRU is also called *darting-out* [2]. This situation is particularly dangerous since neither the driver nor the pedestrian has clear view of the other road user. The sensors

a) Intelligent Vehicles Group, Delft Technical University, The Netherlands; 1) A.Palffy@tudelft.nl; 2) J.F.P.Kooij@tudelft.nl; 3) D.M.Gavrila@tudelft.nl

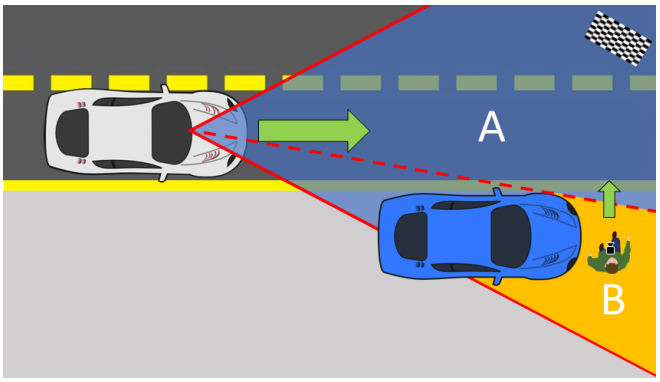


Fig. 1: Darting-out scenario: a pedestrian steps out from behind a parked car (blue), which blocks the line-of-sight of the ego-vehicle (white). Occlusion free/Occluded area is marked with A/B. The checkerboard and the camera on the VRU are only used for ground truth collection.

equipped in the ego-vehicle are also influenced by the parked car, since it blocks their direct line-of-sight to the VRU as well. However, the level of this influence is highly dependent on the sensor type. E.g., a camera may still see the upper part of the pedestrian behind a small car but cannot see them at all behind a truck, or a van. Radars, on the other hand, are able to detect reflections of a pedestrian even in full occlusion via multi-path propagation [12]. I.e., the reflected radar signal can 'bounce' from other parked cars, or the ground underneath the occluding vehicle. Such indirect reflections are weaker than direct ones [12]. However, they still could provide valuable information of a darting-out pedestrian.

Detecting the occluding object (i.e. the parked car, van, or truck) and estimating its size and position is already addressed in the literature, e.g. in [13]. This information can be used to create an occlusion model of the environment which describes what kind of and how many detections are reasonable to expect in differently occluded regions of the scene. We argue that incorporating this occlusion model in a sensor fusion framework helps to detect darting-out pedestrians earlier and more accurately, and we show that camera and radar are suitable choices for such a system.

II. RELATED WORK

Pedestrian detection in intelligent vehicles is a widely researched topic. An extensive survey on vision based pedestrian detection can be found in [14]. In recent years, Convolutional Neural Networks and Deep Learning methods are often applied for pedestrian detection tasks [15].

Various camera based pedestrian detection systems explicitly address the problem of occlusion. E.g., [4] proposes

to use a set of component-based classifiers. A generative stochastic neural network model is used to estimate the posterior probability of pedestrian given its components scores. [5] uses Gradient Patch and a CNN to learn partial features and tackle occlusion without any prior knowledge. In [16] HOG features were used to create an occlusion likelihood map of the scanning window. Regions with mostly negative scores are segmented and considered as occluded regions. [17] used a mixture-of-experts framework to handle partial occlusion of pedestrians. [18] applied motion based object detection and pedestrian recognition on stereo camera input to initiate emergency braking or evasive steering for darting-out scenarios. However, none of these methods use a global model of the scene to describe occlusions, which can change the detection quality or quantity at certain positions (e.g. no/fewer detections behind cars).

Using radars to detect VRUs is also interesting as they are more robust to weather and lighting conditions (e.g. rain, snow, darkness) compared to camera and LIDAR sensors. Several radar based pedestrian detection systems were described in the literature. A radar based multi-class classifier system (including pedestrian and group of pedestrians) was shown in [19]. [7] and [8] both aim to distinguish pedestrians from vehicles using features like size and velocity profiles of the objects using radar. In [20] a radar based pedestrian tracking is introduced using track-before-detection method and particle filtering. They also tested their system on tracks where the VRU enters and leaves an occluded region behind a car. The radar was able to provide measurements even in the occlusion. However, the occlusion itself was not considered.

Several sensor fusion methods were published with the aim of better pedestrian detection. Camera has been combined with LIDAR [11] using CNNs. A fused system of camera and radar is introduced in [10] for static indoor applications. All three sensors were fused in [21] in a multi-class moving object detection system. In [13], fusion of LIDAR and radar was used to detect pedestrians in occlusion in a static experimental setup. They used LIDAR to detect both unoccluded pedestrians and to map occluding objects to provide regions of interest for the radar.

Pedestrians are often tracked using some kind of Bayesian filter, e.g. Kalman Filters (KF) [22]. Another commonly used method is the Particle Filter [20], [23], which estimates the posterior distribution using a set of weighted particles. To fit our use-case (detection and tracking of a pedestrian) a filter should not only track an object, but also report a detection confidence that there is an object to track. [23], [24], [25], [26] give solutions to include this existence probability into Particle Filters.

Several datasets have been published to help the development and testing of autonomous vehicles, e.g. the well-known KITTI [27] or the recently published Eurocity dataset [28]. At the time of writing, NuScenes [29] is the only public automotive dataset to include radar measurements. Unfortunately, it does not include enough darting-out scenes for this research.

In conclusion, pedestrian detection was extensively researched using all three main sensors (camera, LIDAR, radar). However, to the best of our knowledge this is the first research to propose a fused pedestrian detection system in automotive setup to include knowledge of occluded areas into the model, and the first one to explicitly address the previously defined darting-out scenario using a fusion of stereo camera and radar.

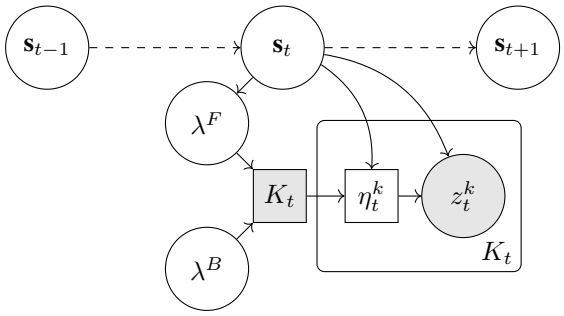


Fig. 2: Graphical model of the probabilistic dependencies in a single time slice t with K_t detections $\mathcal{Z}_t = \{z_t^1, \dots z_t^{K_t}\}$. $\mathbf{s}_t$ is the state vector and λ^B, λ^F are the expected detection rates. The binary flag η_t^k denotes if the k^{th} detection z_t^k comes from foreground or background. Discrete/real variables are shown with square/circle nodes. Observed variables are shaded.

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 camera based pedestrian detection and tracking system on challenging darting-out scenarios, and show that including occlusion information into our model helps to detect these occluded pedestrians earlier and more accurately.

III. Method

In this section, we will present our occlusion aware sensor fusion framework. For modeling the state space and transitions we will follow [23]. A graphical model of the method is shown at Figure 2. Subscripts are timestamps, superscripts are either indices or mark background/foreground.

Let the space $\mathcal{S}$ be a 2D (lateral and longitudinal) position and velocity, and a binary flag marking if the tracked object (e.g. a VRU) exists. Let $\mathbf{s}$ be a state vector in $\mathcal{S}$:

$$\mathcal{S} : \mathbb{R}^2 \times \mathbb{R}^2 \times \{0, 1\}, \tag{1}$$

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

where $\mathbf{x}$ and $\mathbf{v}$ are the object's position and velocity vectors on the ground plane, and $\mathcal{E}$ is its the binary presence flag. $\mathcal{E}$ will be used to represent the existence probability. I.e., $\mathcal{E}=1$ means there is a pedestrian in the scene and $\mathcal{E}=0$ means its absence. We define a Bayesian filter for detection and tracking, for which we need the prior distribution $P(\mathbf{s}_t|\mathcal{Z}_{1:t-1})$, and measurement likelihood function $P(\mathcal{Z}_t|\mathbf{s}_t)$, where $\mathcal{Z}_t$ is the set of all sensor detections at timestamp t.

A. Prediction step

In this subsection, we derive the prior distribution $P(\mathbf{s}_t|\mathcal{Z}_{1:t-1})$ for time t. To get this we need the previous state and state transition $P(\mathbf{s}_t|\mathbf{s}_{t-1})$. To handle the temporal

changes in the presence flag $\mathcal{E}$, we introduce two functions and a parameter to our model. An object stays in the scene with a probability of $p_s(\mathbf{s}_{t-1})$. A new object can appear with a probability of p_n, and its entering position is distributed as $p_e(\mathbf{s}_t)$. Using these, we can set the probabilities of the $\mathcal{E}_t$ flag states given the previous state $\mathbf{s}_{t-1}$:

$$P(\mathcal{E}_t{=}1|\mathcal{E}_{t-1}{=}0, \mathbf{x}_{t-1}, \mathbf{v}_{t-1}) = p_n \tag{3}$$
$$P(\mathcal{E}_t{=}0|\mathcal{E}_{t-1}{=}0, \mathbf{x}_{t-1}, \mathbf{v}_{t-1}) = 1 - p_n \tag{4}$$
$$P(\mathcal{E}_t{=}1|\mathcal{E}_{t-1}{=}1, \mathbf{x}_{t-1}, \mathbf{v}_{t-1}) = p_s(\mathbf{s}_{t-1}) \tag{5}$$
$$P(\mathcal{E}_t{=}0|\mathcal{E}_{t-1}{=}1, \mathbf{x}_{t-1}, \mathbf{v}_{t-1}) = 1 - p_s(\mathbf{s}_{t-1}). \tag{6}$$

In case of a present object ($\mathcal{E}_t{=}1$) the $\mathbf{x}_t$ and $\mathbf{v}_t$ values are distributed as:

$$P(\mathbf{x}_t, \mathbf{v}_t|\mathcal{E}_t{=}1, \mathcal{E}_{t-1}{=}0, \mathbf{s}_{t-1}) = p_e(\mathbf{x}_t, \mathbf{v}_t) \tag{7}$$
$$P(\mathbf{x}_t, \mathbf{v}_t|\mathcal{E}_t{=}1, \mathcal{E}_{t-1}{=}1, \mathbf{s}_{t-1}) = P(\mathbf{x}_t, \mathbf{v}_t|\mathbf{x}_{t-1}, \mathbf{v}_{t-1}).$$

For this last term, we assume a linear dynamic model with a normally distributed acceleration noise.

Thus, we can finally write the full state transition as:

$$P(\mathbf{s}_t|\mathbf{s}_{t-1}) = P(\mathcal{E}_t|\mathbf{s}_{t-1}) \cdot P(\mathbf{x}_t, \mathbf{v}_t|\mathcal{E}_t, \mathbf{s}_{t-1}). \tag{8}$$

B. Update step

Now we describe the likelihood $P(\mathcal{Z}_t|\mathbf{s}_t)$. We assume conditional independence for our sensors, thus the update step here is described for one sensor. The sensor returns K_t detections at once: $\mathcal{Z}_t = \{z_t^1, \dots z_t^{K_t}\}$. Each detection z_t^k contains a 2D location. The total number of detections (K_t) is the sum of foreground (K_t^F) and background (K_t^B) detections: $K_t = K_t^B + K_t^F$. We model the number of foreground (true positive) and background (false positive) detections with two Poisson distributions. Let us denote the detection's rates with λ^B and $\lambda^F(\mathbf{x}_t, \mathcal{E}_t)$ for the background and foreground detections respectively. $K_t^B,\ K_t^F$ are then distributed as Poisson distributions parametrized by $\lambda^B,\ \lambda^F$:

$$K_t^B \sim Pois(\lambda^B), \tag{9}$$
$$K_t^F \sim Pois(\lambda^F(\mathbf{x}_t, \mathcal{E}_t)). \tag{10}$$

Together, the number of detections K_t is distributed as:

$$P(K_t|\mathbf{x}_t, \mathcal{E}_t) \sim Pois(\lambda^B + \lambda^F(\mathbf{x}_t, \mathcal{E}_t)). \tag{11}$$

Note that the number of true positive (foreground) detections depends both on the object's presence and location, thus we can incorporate occlusion information here. E.g., more true detections are expected if the pedestrian is unoccluded than if the pedestrian is occluded.

$$\lambda^F(\mathbf{x}_t, \mathcal{E}_t{=}1) = \begin{cases} \lambda_{unocc} & if\ \mathbf{x}_t \in A, \\ \lambda_{occ} & if\ \mathbf{x}_t \in B, \end{cases} \tag{12}$$

where $\lambda_{unocc},\ \lambda_{occ}$ stand for the expected detection rates in unoccluded (A), occluded (B) areas respectively (see Figure 1). In a not occlusion aware (naive) filter, λ^F is constant and assumes the unoccluded case:

$$\text{naive approach: } \lambda^F(\mathbf{x}_t, \mathcal{E}_t{=}1) = \lambda_{unocc}, \tag{13}$$

as occlusion is not incorporated. Our occlusion aware filter (OAF) behaves the same as a naive one in unoccluded cases, but in occluded positions it adapts its expected rate λ^F.

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

$$P(\eta_t^k{=}1|\mathcal{E}_t, \mathbf{x}_t, K_t) = \frac{\lambda^F(\mathbf{x}_t, \mathcal{E}_t)}{\lambda^F(\mathbf{x}_t, \mathcal{E}_t) + \lambda^B}, \tag{14}$$
$$P(\eta_t^k{=}0|\mathcal{E}_t, \mathbf{x}_t, K_t) = \frac{\lambda^B}{\lambda^F(\mathbf{x}_t, \mathcal{E}_t) + \lambda^B}, \tag{15}$$

where the binary flag η_t^k denotes if the k^{th} detection z_t^k comes from the tracked object, i.e. is a true positive detection.

Now we have to define the likelihood function $P(z_t^k|\eta_t^k, \mathbf{x}_t)$ for true positive ($\eta_t^k{=}1$) and false positive cases ($\eta_t^k{=}0$). We assume that true positive detections are distributed around the object's position $\mathbf{x}_t$ described by some distribution $L(z_t^k|\mathbf{x}_t)$, and that false detections are distributed as described by some distribution $D(z_t^k)$:

$$P(z_t^k|\mathbf{x}_t, \eta_t^k{=}1) = L(z_t^k|\mathbf{x}_t), \tag{16}$$
$$P(z_t^k|\eta_t^k{=}0) = D(z_t^k). \tag{17}$$

Hence the whole likelihood of one measurement is given:

$$\begin{aligned} P(z_t^k|\mathcal{E}_t, \mathbf{x}_t, K_t) = {} & P(z_t^k|\mathbf{x}_t, \eta_t^k{=}1) \cdot P(\eta_t^k{=}1|\mathcal{E}_t, \mathbf{x}_t, K_t) \\ & + P(z_t^k|\eta_t^k{=}0) \cdot P(\eta_t^k{=}0|\mathcal{E}_t, \mathbf{x}_t, K_t). \end{aligned} \tag{18}$$

We assume that all K_t detections are conditionally independent given $\mathbf{x}_t$ and $\mathcal{E}_t$, thus we can update with them individually using (11):

$$P(\mathcal{Z}_t|\mathbf{s}_t) = \prod_{k=1}^{K_t} P(z_t^k|\mathcal{E}_t, \mathbf{x}_t, K_t) \cdot P(K_t|\mathcal{E}_t, \mathbf{x}_t). \tag{19}$$

Existence probability of a pedestrian in the scene given all measurement is then:

$$P(\mathcal{E}_t|\mathcal{Z}_{1:t}) = \iint P(\mathbf{s}_t|\mathcal{Z}_{1:t}) d\mathbf{x}_t d\mathbf{v}_t. \tag{20}$$

IV. IMPLEMENTATION

In this section, we discuss the implementation of the proposed framework in our experiments.

A. Particle filtering

For inference, we use a particle filter to represent the posterior distribution in our model with a set of samples (i.e. particles), because it is straightforward to include occlusion information. I.e., particles in occluded areas are handled differently than those in unoccluded areas.

To include existence probability into the filter, we will follow [23]. The method is briefly explained in this paragraph. From N particles the first one (index 0) is assigned to all the hypotheses with non-present pedestrian, called the negative particle. The remaining $N - 1 = N_s$ particles (called the

positive ones) represent the case of a present pedestrian:

$$\mathcal{E}_t{=}0 \rightarrow w_t^{(0)}, \tag{21}$$

$$\mathcal{E}_t{=}1 \rightarrow (\mathbf{s}_t^{(i)}, w_t^{(i)}) \; for \; i = 1 \dots N_s. \tag{22}$$

where $\mathbf{s}_t^{(i)}$ is the state of the i^{th} particle, and $w_t^{(i)}$ is its assigned weight. Thus, the probability of a non-present/present pedestrian given all detection is the normalized weight of the first particle/summed weights of all remaining, see (20):

$$P(\mathcal{E}_t{=}0|\mathcal{Z}_{1:t}) = w_t^{(0)}, \; P(\mathcal{E}_t{=}1|\mathcal{Z}_{1:t}) = \sum_{i=1}^{N_s} w_t^{(i)}. \tag{23}$$

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

1) Initialization

Particles' positions are initialized uniformly across the Region of Interest (ROI). Their velocity is drawn from normal distribution around walking pace, and their direction is uniformly drawn from an angular region between $\pm 22.5°$, where $0°$ is the direction perpendicular to the movement of the ego-vehicle.

2) Predict 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. First, we estimate the next weight of the first particle as the following.

$$P(\mathcal{E}_t{=}0|\mathcal{Z}_{1:t-1}) \sim \hat{w}_t^{(0)} = \frac{w_{np}}{w_{np} + w_p} \tag{24}$$

where $w_p, \; w_{np}$ are the cumulative weights of present, and not present predicted states using equations (3) - (6):

$$w_p = (p_n) w_{t-1}^{(0)} + \sum_{1}^{N_s} (p_s(\mathbf{s}_t^{(i)})) w_{t-1}^{(i)}, \tag{25}$$

$$w_{np} = (1 - p_n) w_{t-1}^{(0)} + \sum_{1}^{N_s} (1 - p_s(\mathbf{s}_t^{(i)})) w_{t-1}^{(i)}. \tag{26}$$

Afterwards, we sample N_s new positive particles, each is either a mutation of an existing particle moved by the dynamic model, or a completely new (entering) one (7). An existing particle stays in the scene with probability $p_s(\mathbf{s}_t^{(i)})$, or is replaced by a new one with probability of $1 - p_s(\mathbf{s}_t^{(i)})$:

$$\mathbf{s}_{t-1}^{(i)} \rightarrow \begin{cases} \hat{\mathbf{s}_t}^{(i)} \sim P(\mathbf{s}_t|\mathbf{s}_{t-1}^{(i)}) & \text{if moved particle,} \\ \hat{\mathbf{s}_t}^{(i)} \sim p_e(\mathbf{s}_t) & \text{if new particle.} \end{cases} \tag{27}$$

All positive particles weights are normed and set uniformly:

$$\hat{w}_t^{(i)} = \frac{1 - \hat{w}_t^{(0)}}{N_s} \; for \; i = 1 \dots N_s. \tag{28}$$

3) Update step

Particles are updated by new detections based on (19):

$$w_t^{(i)} \propto \hat{w}_t^{(i)} \cdot P(\mathcal{Z}_t|\hat{\mathbf{s}_t}^{(i)}). \tag{29}$$

After the update, all weights are normalized. To avoid sample degeneracy, we resample the positive particles if the Effective Sample Size (ESS) drops below a threshold as in e.g. [30].

Parameter	Short description	In our experiments
$p_s(\mathbf{s}_t^{(i)})$	Probability for a particle to stay	$0.95/0.0$, in/out ROI
p_n	Probability of an entering pedestrian	0.2
λ_{unocc}	Exp. # of detections (occluded)	sensor specific
λ_{occ}	Exp. # of detections (unoccluded)	sensor specific
λ^B	Exp. # of false detections	sensor specific
$D(z_t^k)$	Exp. distribution of noise	Uniform in the ROI
$L(z_t^k\|\mathbf{x}_t)$	Exp. distribution of true detections	$N(z_t^k\|\mathbf{x}_t, \Sigma_s)$
Σ_s	Covariance matrix of true detections	sensor specific

TABLE I: List of model parameters and functions chosen by the user in the framework. See text for sensor specific values.

Our framework can be 'tuned' by the following parameters and functions. The function $p_s(\mathbf{s}_t^{(i)})$ returns that how likely it is for that particle's hypothesis to stay in the scene. p_n is the chance of an entering pedestrian. Their ratio tunes how sceptical is the system about the presence of a pedestrian. $D(z_t^k), \; L(z_t^k|\mathbf{x})$ tune the expected spatial distribution of false, true positive detections. The parameter λ^B gives the assumed rate of false positive detections. Finally, $\lambda_{unocc}, \; \lambda_{occ}$ are the expected rates of occurring true positive detections in unoccluded, occluded areas respectively. The ratio of these tunes how sceptical is the system about a measurement at that position in the first place. For a short conclusion of these parameters see Table I.

B. Sensors

Plausible values for the different λ and Σ_s values were chosen after examining the dataset. A ROI of $5 \; m \times 15 \; m \times 3 \; m$ was applied in front of the vehicle.

Our vision based detection's input is a stereo camera (1936×1216 px) mounted behind the windshield of the research vehicle. Detections are fetched from the Single Shot Multibox Detector (SSD) [31]. Depth is estimated by projecting the bounding boxes into the stereo point cloud computed by the Semi-Global Matching algorithm (SGM) [32], and taking the median distance of the points inside them. For camera, we used $\lambda_{unocc} = 1$ as SSD is reliable at this range in unoccluded regions. λ_{occ} was set to 0.1 in case of partial occlusion scenes, and 0 for fully occluded ones, as no visual clue is expected if the view is fully blocked. Few false positives occurred in the ROI, thus λ^B was set to 0.05. SSD is also used to compute occluded regions. Objects from *car, bus, truck,* and *van* classes are considered as occlusions. Using the projection described before, their 3D positions are calculated. Areas behind them are marked as occluded as shown on Figure 1 and Figure 3.

Our radar sensor is a Continental 400 mounted behind the front bumper. It outputs a list of reflections, each consisting of a 2D position, a Radar Cross Section (RCS) value, and a relative speed to the ego-vehicle. In a preprocessing step we used the RCS and speed values (after compensating for ego-motion) to filter the reflections. We expect $\lambda_{unocc} = 1.5$ detections for unoccluded positions, as often multiple reflections are received from the same pedestrian. Behind occlusions (vehicles), we still assume $\lambda_{unocc} = 0.3$ reflections because of the multi-path propagation described earlier. An average of $\lambda^B = 0.1$ false positive reflections was expected.

V. DATASET

Our dataset consists of 83 recordings. Each contains one pedestrian stepping out from behind a vehicle. See Figure 3 for an example sequence. In total, nine different occluding vehicles were used, of which four were passenger cars (partial occlusion) and five were vans (full occlusion), see Table II. Three persons, with different heights, cloths and body type, acted as pedestrians in our dataset. To record the position of the pedestrian we mounted an additional camera to the chest. A calibration board was placed in the field of view the car and pedestrian cameras as on Figure 1. 6-DoF positions of the sensors relative to the board were obtained by off-the-shelf pose estimation methods.

	Cars	Vans	Total
moving	19	27	48
standing	23	12	35
Total	**42**	**39**	**83**

TABLE II: Number of sequences of each type in the dataset.

VI. EXPERIMENTS

In the experiments we evaluated how the fusion of the two sensors performs compared to the individual sensors in aspects from Section I. Firstly, in subsection VI-A we examine the change of the probability existence over time. Secondly, in subsection VI-B, we discuss the spatial accuracy of the methods. We tested the following four methods: *camera*, *radar*, *naive fusion* and *OAF fusion*. The first three are naive filters using camera/radar/both sensors to update, see (13). The fourth method is the proposed occlusion aware fusion. All methods run with a processing speed of $\sim 10\ Hz$ using 1000 particles in an (un-optimized) Matlab environment on a high-end PC.

As stated by Eq. (13), *OAF fusion* works identically to the naive one in unoccluded positions. To validate this, we simulated a scene without any occlusions or detections, on which the existence probabilities reported by *OAF fusion* and *naive fusion* methods both converged to the exact same value (0.021). This means that comparing the methods is only reasonable when occlusions occur, thus this section will focus on occluded scenes only. Sequences of the dataset were aligned temporally by marking the last moment where the ground truth position is in the occluded region as $T_0 = 0$.

A. Existence probability

We executed the methods on all sequences and recorded the probability outputs as in equation (20). Subfigure 4a and Subfigure 4b show the results from the fully (vans) and partially occluded (cars) scenes.

The results show that detection in car scenarios happens sooner than in the cases of vans by 0.25-0.5 seconds. This is expected for the camera, as in contrast to vans, partial visual detections (e.g. head above the car) occur. In case of radar this was surprising, as the height of a vehicle has no trivial effects on the radar signal. This shift may be caused by the length of the vehicles: vans tend to be longer than cars, which can influence the propagation below the vehicle.

In general, including radar (*Naive/OAF fusion*) helped to detect the pedestrian earlier in both cases, i.e. any chosen threshold of probability is reached sooner by the *OAF fusion* than any other method. E.g., on car scenarios, the threshold of 0.8 is reached by *OAF fusion* 0.3 seconds earlier than the *camera* and 0.12 seconds earlier than the *naive fusion*.

Advantages of occlusion awareness are seen by comparing *naive* and *OAF fusions*. *OAF fusion* reports a higher probability of a present pedestrian than the *naive fusion* at all timestamps at which the pedestrian is occluded ($T < 0$). The reason for this is two fold. Firstly, *OAF fusion* 'acknowledges' that parts of the scenes are occluded and cannot be measured, causing uncertainty. That is, lack of detections from these areas are not considered as evidence for the lack of a pedestrian in contrast to *naive fusion*. Instead, particles behind occlusion gets higher weights compared to the unoccluded particles to represent this uncertainty. This results in higher a priori awareness even before any detections occur. Secondly, detections received from these regions are valued more than in the naive one, as the number of received detections fits the expectations better in Eq. (14). As a consequence, the likelihoods are higher for the same detections than in the *naive fusion* (19). In contrast, the *naive fusion* reports the lowest probability at the beginning of the tracks. This is expected as it receives more evidence for a non-present pedestrian than the two sensors individually, and ignores the uncertainty of occluded regions. In the unoccluded area both fusion yields very similar probabilities as expected, caused by (13).

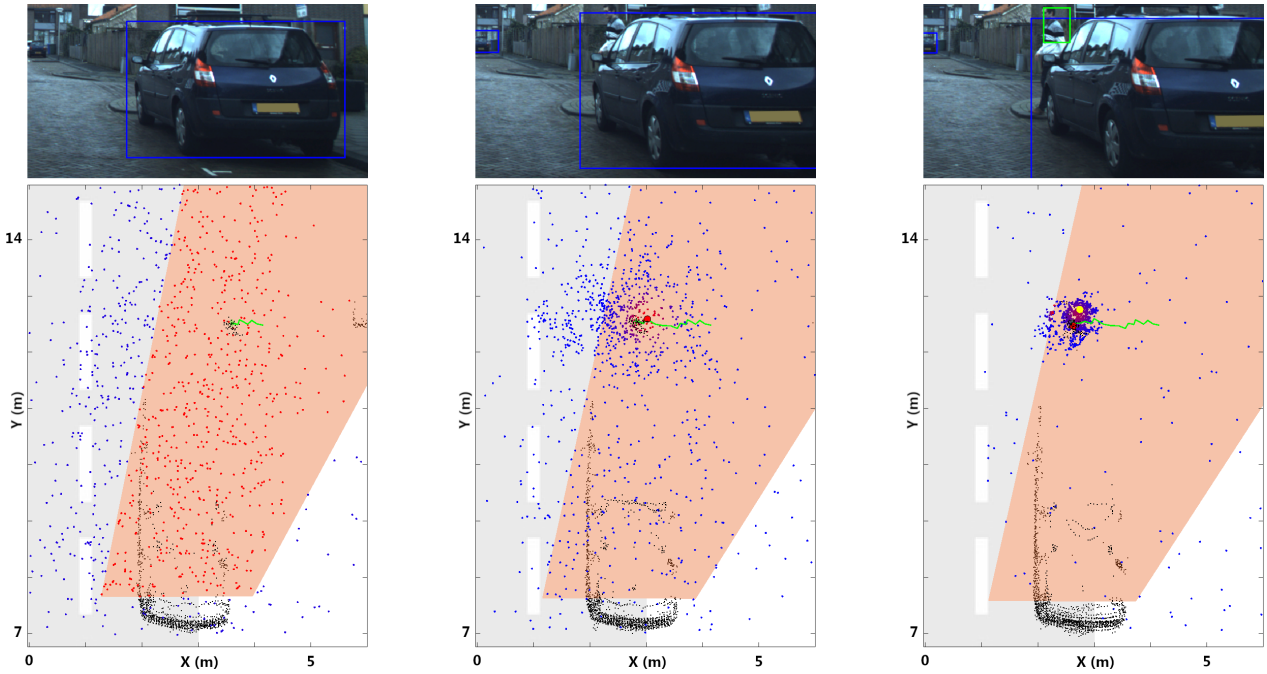
B. Spatial accuracy

Spatial error is calculated as the Euclidean distance of the ground truth and the estimated position. Subfigures 4c and 4d present the results from the fully and partially occluded scenes. Errors are high at the beginning as no detections occured yet, and particles are distributed uniformly, see Subfigure 3a.

Including radar into the system improves the accuracy while the pedestrian is still occluded. The average error of the *radar* converges faster than the *camera*. *Naive fusion* yields smaller errors than *radar*, but not drastically. This is reasonable, as most detections in the occluded area come from the radar.

OAF fusion (see Figure 3) outperforms all other methods and reach its minimal error the fastest. The reason for this is two-fold. Firstly, the occlusion awareness means the filter acknowledges that occluded areas are not possible or hard to observe. However, it can measure that the pedestrian is not in the unoccluded area, see Subfigure 3a. This results in particles in the occluded region having higher weights, and thus, being resampled more often than particles in the open area. The estimated position (the weighted average of the particles) is drifted in the direction of the occlusion. We argue that this is a valid prior, as the area where the particles converge is the expected entrance point of the pedestrian. Secondly, it values detections received from occluded regions more than the *naive fusion* (explained in Subsection VI-A).

All methods converge to a low ($\sim 30\ cm$) spatial error after

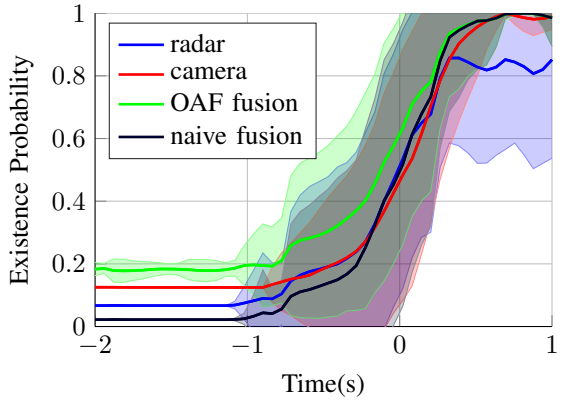


(a) No detection. Occluded particles have larger weights.

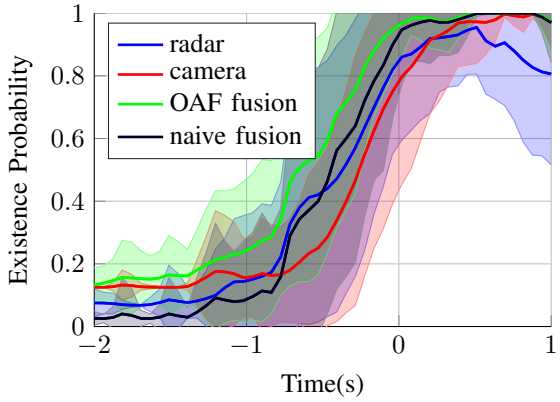
(b) Radar (red filled circle) detects the pedestrian.

(c) Both radar and camera (yellow filled circle) detect the pedestrian.

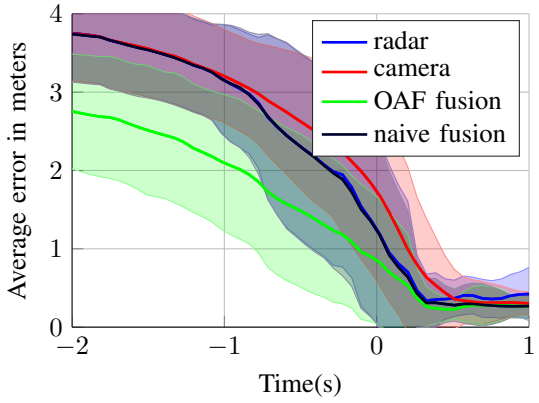
Fig. 3: Camera and top views of the scene at consecutive timestamps using *OAF fusion*. Occlusion (orange) is calculated as the 'shadow' of the blue bounding box from SSD. Initially, particles (blue to red for small to high relative weights) are uniformly distributed (3a), but occluded ones have larger weights. They converge on the location of the pedestrian (ground truth track with green) after first radar (3b) then camera (3c) detects him. Lidar reflections are shown as black for reference.



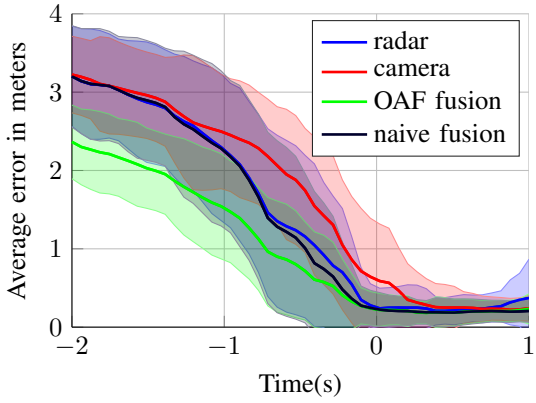
(a) Probability of a present pedestrian (Full occlusion)



(b) Probability of a present pedestrian (Partial occlusion)



(c) Spatial error of estimated state (Full occlusion)



(d) Spatial error of estimated state (Partial occlusion)

Fig. 4: First/second row: Existence probability/Spatial error averaged on all sequences, with variance around the mean. Adding radar resulted in earlier detection than using only camera. *OAF fusion* performed with the highest confidence/accuracy.

the pedestrian left the occlusion, since the occlusion aware methods work identically to the naive ones in unoccluded positions (13).

VII. CONCLUSIONS

In this paper we proposed a generic occlusion aware multi-sensor Bayesian filter to address the challenge of detecting pedestrians. We applied the filter on a real life dataset of darting-out pedestrians recorded with camera and radar sensors.

Our results show that the inclusion of radar sensor and occlusion information is beneficial as pedestrians are detected earlier and more accurately. E.g., on car scenarios, the existence probability threshold of 0.8 is reached 0.3 seconds earlier by our occlusion aware fusion than by a naive camera based detector and 0.12 seconds earlier than the *naive fusion* method.

In this paper, some simplifications were applied. E.g., our method to build the occlusion model was suitable for our experiments, however, more precise 3D estimation of the vehicles could provide a better a priori knowledge to the framework. We assumed uniformly distributed background noise, although their occurrence might not be spatially independent. Including further sensors (e.g. LIDAR or additional radars) into our modular framework is straightforward and is also an interesting research topic.

ACKNOWLEDGEMENT

This work received support from the Dutch Science Foundation NWO-TTW, within the SafeVRU project (nr. 14667).

REFERENCES

[1] World Health Organization, "Road traffic injuries," 2018.

[2] R. Sherony and C. Zhang, "Pedestrian and Bicyclist Crash Scenarios in the U.S," *IEEE Conference on Intelligent Transportation Systems, Proceedings (ITSC)*, vol. 2015-Octob, pp. 1533–1538, 2015.

[3] C. G. Keller and D. M. Gavrila, "Will the pedestrian cross? A study on pedestrian path prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 2, pp. 494–506, 2014.

[4] S. Aly, L. Hassan, A. Sagheer, and H. Murase, "Partially occluded pedestrian classification using part-based classifiers and Restricted Boltzmann Machine model," in *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, 2013, pp. 1065–1070.

[5] S. Kim and M. Kim, "Occluded pedestrian classification using gradient patch and convolutional neural networks," in *Lecture Notes in Electrical Engineering*, 2017, vol. 421, pp. 198–204.

[6] K. Granström, S. Reuter, M. Fatemi, and L. Svensson, "Pedestrian tracking using Velodyne data stochastic optimization for extended object tracking," *IEEE Intelligent Vehicles Symposium (IV)*, no. Iv, 2017.

[7] S. Heuel and H. Rohling, "Pedestrian classification in automotive radar systems," *Proceedings International Radar Symposium*, pp. 39–44, 2012.

[8] ——, "Pedestrian recognition in automotive radar sensors," *International Radar Symposium (IRS)*, vol. 2, pp. 732–739, 2013.

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

[10] R. Streubel and B. Yang, "Fusion of Stereo Camera and MIMO-FMCW Radar for Pedestrian Tracking in Indoor Environments," *Fusion*, 2016.

[11] J. Schlosser, C. K. Chow, and Z. Kira, "Fusing LIDAR and images for pedestrian detection using convolutional neural networks," *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2198–2205, 2016.

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

[13] S. K. Kwon, E. Hyun, J.-H. Lee, J. Lee, and S. H. Son, "Detection scheme for a partially occluded pedestrian based on occluded depth in lidarradar sensor fusion," *Optical Engineering*, vol. 56, no. 11, p. 1, 2017.

[14] R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Ten Years of Pedestrian Detection, What Have We Learned?" in *Computer Vision - ECCV 2014 Workshops*, L. Agapito, M. M. Bronstein, and C. Rother, Eds. Cham: Springer International Publishing, 2015, pp. 613–627.

[15] 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. 0, pp. 1–17, 2018.

[16] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *2009 IEEE 12th International Conference on Computer Vision*, 2009.

[17] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila, "Multi-cue pedestrian classification with partial occlusion handling," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 990–997, 2010.

[18] C. G. Keller, T. Dang, H. Fritz, A. Joos, C. Rabe, and D. M. Gavrila, "Active pedestrian safety by automatic braking and evasive steering," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1292–1304, 2011.

[19] 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," *Sensor Data Fusion: Trends, Solutions, Applications (SDF)*, pp. 1–6, 2017.

[20] M. Heuer, A. Al-Hamadi, A. Rain, and M. M. Meinecke, "Detection and tracking approach using an automotive radar to increase active pedestrian safety," *IEEE Intelligent Vehicles Symposium, Proceedings*, no. Iv, pp. 890–893, 2014.

[21] R. O. Chavez-Garcia and O. Aycard, "Multiple Sensor Fusion and Classification for Moving Object Detection and Tracking," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 2, pp. 525–534, 2016.

[22] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila, "Context-based pedestrian path prediction," *European Conference on Computer Vision (ECCV)*, vol. 8694 LNCS, no. PART 6, pp. 618–633, 2014.

[23] S. Munder, C. Schnörr, and D. Gavrila, "Pedestrian Detection and Tracking Using a Mixture of View-Based Shape-Texture Models," *IEEE Transactions on intelligent transportation systems*, pp. 1–25, 2008.

[24] D. Salmond and H. Birch, "A particle filter for track-before-detect," *Proceedings of the 2001 American Control Conference. (Cat. No.01CH37148)*, pp. 3755–3760 vol.5, 2001.

[25] G. Wang, S. Tan, C. Guan, N. Wang, and Z. Liu, "Multiple model particle filter track-before-detect for range ambiguous radar," *Chinese Journal of Aeronautics*, vol. 26, no. 6, pp. 1477–1487, 2013.

[26] Z. Radosavljević, D. Mušicki, B. Kovačević, W. Kim, and T. Song, "Integrated particle filter for target tracking in clutter," *IET Radar, Sonar and Navigation*, vol. 9, no. 8, pp. 1–2, 2015.

[27] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[28] M. Braun, S. Krebs, F. Flohr, and D. Gavrila, "EuroCity Persons: A Novel Benchmark for Person Detection in Traffic Scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, may 2019.

[29] NuTonomy, "The nuScenes dataset." [Online]. Available: https://www.nuscenes.org/

[30] 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 Systems with Applications*, vol. 41, no. 8, pp. 3944–3954, 2014.

[31] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9905 LNCS, pp. 21–37, 2016.

[32] H. Hirschmüller, "Stereo Processing by Semi-Global Matching and Mutual Information," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 328–341, 2008.