People Tracking by Cooperative Fusion of RADAR and Camera Sensors

Martin Dimitrievski*, Lennert Jacobs†, Peter Veelaert‡ and Wilfried Philips§

TELIN-IPI, Ghent University - imec,
St-Pietersnieuwstraat 41, B-9000 Gent, Belgium
Email: *martin.dimitrievski@ugent.be, †lennert.jacobs@ugent.be, ‡peter.veelaert@ugent.be, §wilfried.philips@ugent.be

***Abstract*—Accurate 3D tracking of objects from monocular camera poses challenges due to the loss of depth during projection. Although ranging by RADAR has proven effective in highway environments, people tracking remains beyond the capability of single sensor systems. In this paper, we propose a cooperative RADAR-camera fusion method for people tracking on the ground plane. Using average person height, joint detection likelihood is calculated by back-projecting detections from the camera onto the RADAR Range-Azimuth data. Peaks in the joint likelihood, representing candidate targets, are fed into a Particle Filter tracker. Depending on the association outcome, particles are updated using the associated detections (Tracking by Detection), or by sampling the raw likelihood itself (Tracking Before Detection). Utilizing the raw likelihood data has the advantage that lost targets are continuously tracked even if the camera or RADAR signal is below the detection threshold. We show that in single target, uncluttered environments, the proposed method entirely outperforms camera-only tracking. Experiments in a real-world urban environment also confirm that the cooperative fusion tracker produces significantly better estimates, even in difficult and ambiguous situations.**

***Index Terms*—radar, sensor fusion, pedestrian tracking, autonomous vehicles**

I. Introduction

Environmental perception requirements for reaching Level 4 and 5 autonomous driving, [1], demand for complete integration of hardware and software as well as the development of smarter object detection and tracking algorithms. All of these systems must be able to cope with low Signal-to-Noise Ratio (SNR) data, dynamic occlusion, unpredictable motion as well as sensor failure. To reach these requirements, contemporary prototypes are usually equipped with an array of complementary and redundant sensors. On the other hand, at the low cost, high Technology Readiness Level (TRL) segment, traditional systems have limited the possibilities. A system of independent sensors, detectors and trackers can be easily analyzed and standardized, but sharing the rich sensor information between the sensors can lead to larger gains in robustness and performance. This essentially explains the archetypal difference between low level and high level data fusion. In the context of Advanced Driver-Assistance Systems (ADAS) it is pivotal to have an accurate situational awareness image of the environment. The position of potential collision threats must be estimated on the ground plane relative to the ego-vehicle. To that end, the most discriminating information can be captured using a high-resolution visible light camera. However, back-projecting objects from the camera image to the ground plane cannot be done accurately, so, often ranging is performed by additional sensors. Technologies such as stereo cameras, rotating and flash LiDAR, ultrasound and RADAR have the advantage that they can measure distance to objects directly. When measuring distances to soft targets such as pedestrians and cyclists it can be seen, from the leader boards of the KITTI 3D object benchmark [2], [3], that LiDAR has a clear advantage over all other sensors. However, LiDAR has limited practical application due to its high cost. More so, LiDAR alone does not have the necessary data density to perform robust object classification, so it is often used in conjunction with a visible light camera.

As an alternative to expensive LiDAR technology, car manufacturers commonly rely on automotive Frequency Modulated Continuous Wave (FMCW) RADAR sensors. Not only are these compact, low-cost, and low-power sensors providing range, (radial) velocity, and angular information, RADAR technology is also largely insensitive to environmental conditions like rain, snow, fog, dust, dirt, darkness, or glaring sun. 77GHz FMCW RADAR has become the de-facto standard in automotive applications, as the large available bandwidth and small wavelengths allow to resolve objects in range and velocity with high resolution. However, as compared to LiDAR, automotive RADAR provides much lower angular resolution. Hence, although ranging of vehicles has been successfully performed in highway environments, RADAR struggles to track the large variety of road users that are commonly present in urban environments. Since pedestrians have a much lower Radar Cross Section (RCS) than other road users and infrastructure, a single pedestrian in a highly cluttered scene is very hard to identify in the received signal spectrum. This is especially true when the pedestrian is static, occluded or close to RADAR reflecting objects such as cars, light poles, traffic lights, signs and various other corner reflectors. Therefore, current environmental perception systems mainly rely on a combination of cameras and RADARs. In this paper we propose a hybrid low and high level cooperative fusion architecture for object detection and tracking. Cooperative sensor fusion is accomplished by continuously tuning the low level operating characteristics of

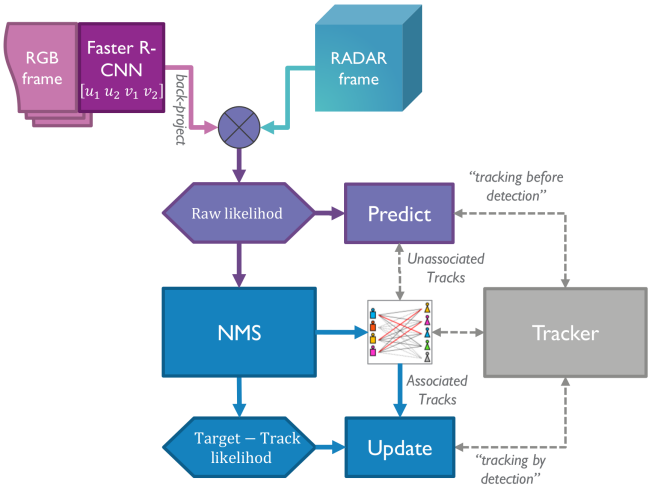


Figure 1: System diagram of the proposed cooperative fusion method: Camera observations are back-projected on the ground plane and fused with RADAR data. Depending on target-track association outcome the tracker uses raw data or joint-likelihood to update the tracks.

one sensor by means of feedback data from another sensor or higher level information. The goal is to adapt the current data of one sensor using the strengths of another sensor, while at the same time exploiting high level temporal and historical knowledge in reaching a common goal. In our case the goal is accurate tracking of vulnerable road users on the ground plane using 77GHz automotive RADAR and wide angle camera. The state of each person is represented by a set of random samples, particles, whose weights and positions allow flexibility in exploiting low-level sensor information. Initial ground plane detections are obtained by back-projecting camera detections onto the the raw RADAR range/azimuth signal, violet color on figure 1 and figure 2c. Next, extracted local peaks in this likelihood, representing expected candidate target centers, are fed into a multi-target tracker. Candidate targets, which can be associated with existing tracks, update the respective state via a Target-track likelihood, blue color on figure 1, while unassociated tracks are updated using the sensor observations before detection. Therefore, in situations with well associated data, the tracker gets less ambiguous likelihood information from the peaks in the data, while in ambiguous situations, the tracker updates sampling the entire energy field where multiple weak targets can update multiple unassociated tracks. Our hybrid method is essentially a Tracking by Detection design for good data associations, and Tracking Before Detection (TBD) when no association is possible.

The following sections are organized as follows, in section §II we give a brief overview of relevant tracking approaches from the literature, in section §III we provide the theoretical foundation of our approach while in section §IV we show experimentally that the proposed cooperative fusion improves upon the baseline. Finally, in section §V we conclude this paper with some remarks on the applicability and potential improvements to the system.

II. Related work

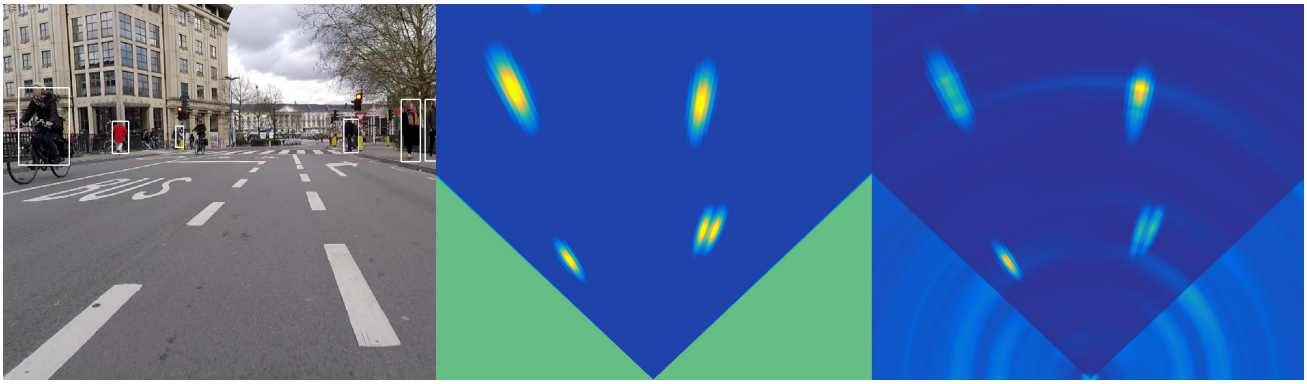
There already exists a vast amount of literature dealing with sensor fusion, the review of which is outside of the scope of this paper. We hereby provide an overview of several relevant papers that conceptually intersect with our work.

In [4] authors give a comprehensive evaluation of tracking performance for various single and multi-sensor system setups in urban and highway scenarios. Their analysis is based on the Cramer-Rao Lower Bound (CRLB) which provides a lower bound on the variance of an estimator. The metric can be used as a design tool in order to estimate the tracking performance limits. Although their analysis is thorough, it is based on loose sensor model assumptions that limit real-world applicability. Nevertheless, this work reaches interesting conclusions that, in highway environments, best performance can be achieved by sensor data fusion of radar and LiDAR while in urban environments, any two sensor combination provides satisfactory performance.

An object tracker based on heterogeneous sensors (LiDAR, RADARs and cameras) using the Extended Kalman Filter (EKF) is presented in [5]. These authors employ the sequential-sensor method [6] that treats observations from individual sensors independently and sequentially feeds them to the EKF's estimation process. Tracking of pedestrians, cyclists and vehicles is performed by applying a class-specific motion model. Data association is performed by back-projecting camera object detections on the ground plane and searching for the nearest LiDAR and RADAR detection. Thus, the rich RADAR information is greatly discarded in this process. This paper provides an extensive experimental evaluation from a tracking perspective, however the target localization accuracy is not fully evaluated.

In [7] authors use a 24GHz RADAR and a Track-Before-Detect approach, to consider measurements which are usually being discarded. Their approach is based on a Particle Filter whose weights are influenced by Doppler signatures caused by human walking. This system is limited to tracking only a single target where events such as track creation and termination are not considered. Finally, this paper offers only a small scale experimental evidence which is not entirely relevant to real traffic situations.

In [8] authors describe an algorithm for a multi-Bernoulli filter for TBD by eliminating targets from the original observation of an automotive Fast Chirp Modulation (FCM) RADAR. With sequential Monte Carlo (SMC) implementation of the proposed algorithm, their approach is validated through a small scale experiment using a simulation of an urban road scene. A weakness of this approach is the process of eliminating the other targets in the likelihood calculation, where elimination regions are defined by the other target's intensity and the actual radar parameters, such as the number of array elements and the effective aperture. A vast improvement in the performance can be achieved



(a) RGB camera frame with detected objects. (b) Uncertainty regions from back-projected camera detections on the Ground plane. (c) Uncertainty regions from combined camera and RADAR.

Figure 2: Data sample and detections from an urban environment containing multiple vulnerable road users.

by introducing information from a camera sensor that can separate multiple targets in the image domain.

III. Tracking by Cooperative Fusion

Our proposed system extends on the TBD idea of these papers to tracking multiple soft targets from a moving vehicle. We propose a Particle Filter based multi-object tracker which uses a neural network [9] to perform initial object detection. These detections are then fused with the RADAR data to form a joint RADAR-Vision likelihood for potential targets. During tracking prediction, particle weights are updated using the joint-likelihood function, while detected targets are used to innovate existing or create new tracks. Thus, during tracking the rich RADAR information is fully exploited. Whereas authors in [7] use a constant velocity model for pedestrian motion, we adopt our behavioral motion model from [10]. We solve the data association problem using the Hungarian algorithm [11] while track management is done using a Markov Decision Process approach similar to [12].

Formally, the goal of the system is to estimate the state and cardinality of the set of unknown targets $X = \{\mathbf{x}_j\}, j \geq 0$ by maximizing the belief in the state using past and current sensor information. We model a target as a random variable on the ground plane, with four parameters (position and velocity) in the space $\mathbf{x} \in \mathbb{R}^4$; $\mathbf{x} = [x, y, \dot{x}, \dot{y}]^{\mathrm{T}}$. The system relies on a set of sensors $Y = \{Y_C, Y_R, ...\}$ to scan the environment around the vehicle for potential collision threats. The RADAR sensor generates observations in the form of a 3D data cube: $Y_R : \mathbf{y}_R \in \mathbb{R}^3$ in the range, (radial) velocity, and azimuth space. The camera detector, on the other hand, generates a set of observations, y_i, represented by rectangular bounding boxes in image plane coordinates $Y_C : \mathbf{y}_{C,i} = [u_1, u_2, v_1, v_2, s]_i^{\mathrm{T}}$ with u, v representing image rows and columns. Since this camera observation vector is incompatible with the target state space, we transform each bounding box to an expected location on the ground plane. To that end we employ the back-projection:

$$\mathbf{y}'_C = [x', y', s]^{\mathrm{T}}; \; y' = \frac{h f_y}{u_2 - u_1} \text{ and } x' = \frac{(v - v_0)\, y'}{f_x}, \quad (1)$$

where h is assumed an average person height, and f_x, f_y and v_0 are the intrinsic camera parameters. For better readability we will drop the prime symbol in all further equations, assuming that the back-projection is already applied to all camera observations. In the following sub-section we will use Bayesian tracking theory under the Markovian assumption to provide an analytical solution for tracking a single target using cooperating sensors. In our previous work [10] we give for more details on how we solve the association and track management tasks.

A. Bayesian tracking

Under the Markov process assumption, the state of the target is conditioned on the previous state, i.e. the target motion model, and on current sensor observations. Since the target of interest can be observed by multiple sensors, we model the probability for a target to be in the state $\mathbf{x}$, at time t as a function of the previous state and process noise: $p(\mathbf{x}_t | \mathbf{x}_{t-1}) = f(\mathbf{x}_{t-1}, \xi_{t-1})$. The sensors measure the true state of each target providing an observation $\mathbf{y}_t$ which suffers from measurement noise, $\mathbf{y}_t = h(\mathbf{x}_t, \eta_t)$. In our multi sensor system, the measurement model represents the conditional dependence between the target state and the various sensor observations:

$$p(\mathbf{y}_t | \mathbf{x}_t) = p(\mathbf{y}_{C,t}, \mathbf{y}_{R,t} | \mathbf{x}_t). \quad (2)$$

In order to maximize the belief in the state, given the state transition model $p(\mathbf{x}_t | \mathbf{x}_{t-1})$, past and current observations $\{\mathbf{y}_{C,1:t}, \mathbf{y}_{R,1:t}\}$, we employ Bayesian tracking recursion. This enables us to estimate the posterior density function in two steps, i.e. using the motion model to make a prediction, and the observation likelihood function to update with observations. During the prediction step the past state is propagated to the current time by using the state transition probability $p(\mathbf{x}_t | \mathbf{x}_{t-1})$, where by dropping the observation from the state transition term, which holds under the Markov assumption, we get:

$$p\left(\mathbf{x}_t \left| \mathbf{y}_{C,1:t-1}, \mathbf{y}_{R,1:t-1} \right.\right) = \int p\left(\mathbf{x}_t \left| \mathbf{x}_{t-1} \right.\right) p\left(\mathbf{x}_{t-1} \left| \mathbf{y}_{C,1:t-1}, \mathbf{y}_{R,1:t-1} \right.\right) d\mathbf{x}_{t-1}. \quad (3)$$

During the update step, new observations $\mathbf{y}_{C,t}, \mathbf{y}_{R,t}$ become available and innovate the state variable of the respectively associated target $\mathbf{x}_{j,t}$ using the Bayes' rule. Assuming that the track is a first order Markov process we re-write the posterior as:

$$p\left(\mathbf{x}_t \left| \mathbf{y}_{C,1:t}, \mathbf{y}_{R,1:t} \right.\right) = \frac{p\left(\mathbf{x}_t \left| \mathbf{y}_{C,1:t-1}, \mathbf{y}_{R,1:t-1} \right.\right) p\left(\mathbf{y}_{C,t}, \mathbf{y}_{R,t} \left| \mathbf{x}_t \right.\right)}{p\left(\mathbf{y}_{C,t}, \mathbf{y}_{R,t} \left| \mathbf{y}_{C,1:t-1}, \mathbf{y}_{R,1:t-1} \right.\right)}. \quad (4)$$

The camera and RADAR sensors operate using different principles and in different EM spectra which makes their measurements conditionally independent given the state. Thus, the first factor in the numerator is the result of the prediction step: $p\left(\mathbf{x}_t \left| \mathbf{y}_{C,1:t-1}, \mathbf{y}_{R,1:t-1} \right.\right) = p\left(\mathbf{x}_t \left| \mathbf{x}_{t-1} \right.\right)$ and the second factor is the likelihood function of two conditionally independent variables: $p\left(\mathbf{y}_{C,t}, \mathbf{y}_{R,t} \left| \mathbf{x}_t \right.\right) = p\left(\mathbf{y}_{C,t} \left| \mathbf{x}_t \right.\right) p\left(\mathbf{y}_{R,t} \left| \mathbf{x}_t \right.\right)$. The denominator term $p\left(\mathbf{y}_t \left| \mathbf{y}_{1:t-1} \right.\right)$ can be computed using the likelihood function and the previous state. We find the solution to this Bayesian recursion using non-parametric distributions and Monte Carlo simulations. We model the posterior as a weighted sum of N discrete samples:

$$p\left(\mathbf{x}_t \left| \mathbf{y}_{C,1:t}, \mathbf{y}_{R,1:t} \right.\right) \approx \sum_{i=1}^{N} w_t^i \delta\left(\mathbf{x}_t - \mathbf{x}_t^i\right), \quad (5)$$

where $\mathbf{x}_t^i$ is a random sample from this distribution: $\mathbf{x}_t^i \sim p\left(\mathbf{x}_t \left| \mathbf{y}_{C,1:t}, \mathbf{y}_{R,1:t} \right.\right)$, δ is the Dirac delta function and w_t^i are sample weights, initially $w_t^i = \frac{1}{N}$. An approximation of this distribution can be obtained by means of importance sampling. The importance of each particle can be computed recursively, [13], using the proposal distribution $q()$:

$$w_t^i = w_{t-1}^i \frac{p\left(\mathbf{y}_{C,t} \left| \mathbf{x}_t^i \right.\right) p\left(\mathbf{y}_{R,t} \left| \mathbf{x}_t^i \right.\right) p\left(\mathbf{x}_t^i \left| \mathbf{x}_{t-1}^i \right.\right)}{q\left(\mathbf{x}_t^i \left| \mathbf{x}_{1:t-1}^i, \mathbf{y}_{C,1:t}, \mathbf{y}_{R,1:t} \right.\right)}, \quad (6)$$

where the numerator is the product of the observation model $p\left(\mathbf{y}_{C,t} \left| \mathbf{x}_t^i \right.\right) p\left(\mathbf{y}_{R,t} \left| \mathbf{x}_t^i \right.\right)$ and the motion model $p\left(\mathbf{x}_t^i \left| \mathbf{x}_{t-1}^i \right.\right)$ and the denominator is the proposal distribution. For simplicity, we use a Bootstrap PF (6) weight update assuming that the proposal distribution is the state transition prior:

$$w_t^i = w_{t-1}^i p\left(\mathbf{y}_{C,t} \left| \mathbf{x}_t^i \right.\right) p\left(\mathbf{y}_{R,t} \left| \mathbf{x}_t^i \right.\right). \quad (7)$$

This way we greatly simplify the computation of the particle's weight updates which enables us to perform Tracking by Detection or Tracking Before Detection whether a specific peak in the joint likelihood can be found and associated or not. Specifically, if the track is predicted and not updated, we update particle weights using the raw likelihood. figure 1, before running NMS/CFAR [14]. Even if, at time step t, all observations fall below the tracker sensitivity threshold and no detections $\mathbf{y}_t$ are available for innovation, we have already used the RADAR data for updating the particle weights. In cases where detection and association can be made, we use the a target-track joint likelihood function which we present in the following sub-section.

B. Joint-likelihood function

Since likelihoods govern the PF weights which are a component of the posterior (5), accurately modeling this function is essential to the performance of the tracker. For this reason, we learn the observation likelihood model parameters from data using the actual physical sensors in controlled and uncluttered environments. For the camera sensor, we model the measurement positional uncertainty, that a person will be detected at position $\mathbf{y}_C = [x', y']^{\mathrm{T}}$ if this person is standing at position $\mathbf{x} = [x, y]^{\mathrm{T}}$ as the likelihood function $p\left(\mathbf{y}_C \left| \mathbf{x} \right.\right)$. Multiple factors, such as bounding box errors, pose variability, occlusion, etc. influence the position of the back-projected bounding box on the ground plane. These appearance factors are difficult to model which makes the detection uncertainty function unknown. We therefore approximate this function with a two-dimensional Gaussian in polar coordinates, where the radial and angular variances are a function of the range. Incorrect person height in the model creates ground plane errors that scale linearly with distance which can be well captured by our model. Thus, for $p\left(\mathbf{y}_C \left| \mathbf{x} \right.\right)$ we have:

$$p\left(\mathbf{y}_C \left| \mathbf{x} \right.\right) = \exp\left(-\frac{\left(\rho_{\mathbf{y}} - \rho_{\mathbf{x}}\right)^2}{\sigma_\rho^2} - \frac{\left(\theta_{\mathbf{y}} - \theta_{\mathbf{x}}\right)^2}{\sigma_\theta^2}\right), \quad (8)$$

where the range variance is a function of the track range $\sigma_\rho = a\rho_{\mathbf{x}}$ and the azimuth variance σ_θ is constant.

Detections from the RADAR data cube are affected by various types of noise which create errors on the position $\mathbf{y}_R = [\rho, \theta, \dot{\rho}]^{\mathrm{T}}$. Since people have a radar cross-section compared to other road users and infrastructure, the useful signal that we seek to extract from the data cube is strongly corrupted. Firstly, objects moving with velocity larger than the maximum unambiguous velocity suffer from aliasing and appear as ghost targets. Second, in a real world traffic environment the RADAR signal suffers from the effects of multipath propagation which is difficult to model without knowledge of the scene structure. Lastly, there is a strong signal mixing when two people stand close to each other and without prior knowledge of the number of targets it is very difficult to resolve such targets. In order to reduce the computational load of the likelihood function, we model the RADAR observation likelihood with the two-dimensional Gaussian function in polar coordinates similar to (8). We note that $p\left(\mathbf{y}_R \left| \mathbf{x} \right.\right)$ has variances learned from control experiments which are completely different from the ones of $p\left(\mathbf{y}_C \left| \mathbf{x} \right.\right)$. The camera likelihood has a distinctive range dependent uncertainty ($\sigma_{C,\rho} = 0.039\rho_{\mathbf{x}}\,[m]\,, \sigma_{C,\theta} = 0.014\,[rad]$), while the RADAR likelihood has a "banana" shape with constant range variance and much larger angular variance ($\sigma_{R,\rho} = 0.17\,[m]\,, \sigma_{R,\theta} = 0.344\,[rad]$).

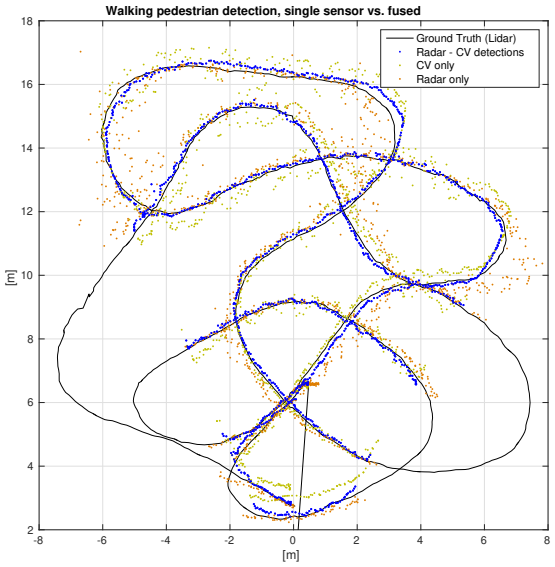


Figure 3: Detections of a single pedestrian walking on a known trajectory. Black line: ground truth trajectory computed by segmenting the LiDAR data; Red: detections from RADAR; Green: back-projected detections from Faster R-CNN; Blue: fused detections.

IV. Experimental results

We conducted two sets of experiments and measured performance of the cooperative system against the control, camera-only tracker. For capturing the data, we equipped an electric cargo tricycle with a sensor array consisting of a single GoPro Hero 4 black RGB camera, Texas Instruments AWR1443 77GHz automotive FMCW RADAR and a Velodyne VLP-16 LiDAR. Data was captured asynchronously at 30Hz, 20Hz and 10Hz respectively, while synchronization is achieved using timestamps. In the first experiment, the ego-vehicle was parked in an open environment (empty parking lot) while the pedestrian was walking in a predictable and known trajectory in front of the sensor array. The goal of this first experiment is to compute the improvement of raw positional accuracy by controlling for data ambiguities such as multiple target association, occlusions and ego-motion. Camera object detection was performed by running the Faster R-CNN [9] object detector trained on the MS-COCO dataset [15]. Performance is measured by means of Root Mean Squared Error of the estimated person position on the ground plane position with respect to the ground truth which comes from a calibrated VLP-16 LiDAR. On figure 3, we present one typical trajectory and the corresponding results obtained from the control and fusion-based system. It is apparent that the camera error pattern, shown by green dots, is more pronounced in the longitudinal direction, while the RADAR error, shown in red dots, is stronger in the lateral direction. Applying the cooperative fusion system in this experiment brought significant improvements in the localization accuracy. We measured a decrease in RMSE from $0.357m$ for the camera-only, and $0.503m$ for the RADAR-only, to $0.188m$ for the fusion method, which is an improvement of close to 47%.

In the second experiment we tested the real-world impact of tracking accuracy using our cooperative fusion system over a single sensor baseline. To that end, we conducted a large scale data capture and annotation in an urban environment (city center). For this paper, we selected 16 sequences with difficult traffic situations, where ego-motion ranges from $0Km/h$ to $25Km/h$. All sequences were hand annotated by computer vision researchers, in such a way that annotators were asked to draw ground plane bounding boxes around each vulnerable road user visible both in the LiDAR, RADAR and Camera data. The annotator could also advance the time in the past or future in order to accurate label ambiguous cases. In total, the dataset contains 1922 frames captured at 10Hz, with 6734 labeled ground truth objects.

In all experiments, accurate odometry was obtained by applying the LiDAR odometry algorithm of [16]. The tracker [10] solves the association problem, and decides whether to update an existing track, spawn a new track or merge tracks. We output tracks which reach a confidence score of $\chi_t > 0.7$, which we then compare against the ground truth annotations using gated nearest neighbor association. A gating of $2m$ is applied, meaning results are matched only to ground truth annotations within $2m$. Within these gates, we compute the RMSE of the cooperative fusion tracker against the control, camera-only tracker. On table I we provide a full breakdown of the results per sequence and per range bracket. We observed that localization accuracy decreases with the increase of range in both camera-only and fusion experiments. However, we show that the cooperative fusion system can better localize pedestrians and cyclists in almost all sequences and all range brackets. In this realistic experiment, fusion achieves an average RMSE of $0.826m$ for all tracked targets compared to the camera-only RMSE of $0.947m$. This improvement of around 15% is significant and very important because it stems from uncontrolled real-world data. It shows that our cooperative fusion method is able to extract useful RADAR information in the highly cluttered urban environment. As previously discussed, various factors such as multi-target ambiguity and multipath diminish the improvement as compared to the open space single target experiment. Demo videos and additional material can be found on our project page [1] where we show the input and output of the tracker for all test sequences.

V. Conclusion

In this paper we proposed a novel, vulnerable road user tracker based on cooperative fusion of RADAR and camera information. Targets are detected and tracked accurately on the ground plane, with the possibility of using tracking predictions in a path planning algorithm. Particle sample weights are updated using the raw RADAR-camera joint observation likelihood before candidate targets are selected by the CFAR algorithm. In the case where no candidates

[1] https://telin.ugent.be/~mdimitri/tracking

Target localization accuracy, RMSE [m]↓																		
		Sequence																
Range	Method	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	Mean
$[0, 10)\ m$	Camera	1.05	0.89	0.74	1.02	0.82	0.71	0.81	0.67	0.66	0.54	0.76	0.79	0.75	0.79	0.75	**0.70**	0.765
	Fusion	**0.85**	**0.83**	**0.60**	**0.79**	**0.74**	**0.53**	**0.60**	**0.56**	**0.53**	**0.37**	**0.65**	**0.60**	**0.61**	**0.65**	**0.58**	0.74	**0.601**
$[10, 20)\ m$	Camera	1.33	0.87	1.02	1.11	**1.00**	1.02	1.17	0.83	1.32	1.08	0.84	0.77	**0.99**	1.17	1.04	0.87	1.058
	Fusion	**1.21**	**0.84**	**0.94**	**0.80**	1.01	**0.86**	**1.00**	**0.63**	**0.84**	**0.86**	**0.46**	**0.64**	1.11	**1.06**	**0.90**	**0.78**	**0.938**
$[20, 30)\ m$	Camera	n/a	n/a	1.48	1.32	1.22	1.29	**1.30**	n/a	n/a	1.64	n/a	1.39	n/a	1.33	1.36	1.21	1.247
	Fusion	n/a	0.97	**1.28**	**1.24**	1.22	**1.19**	1.33	n/a	n/a	**1.53**	n/a	**0.61**	n/a	**1.28**	**1.28**	**0.95**	**1.218**

Table I: Tracking localization accuracy results, bold indicates better results. (↓-lower is better). n/a fields indicate sequences without targets in the respective range.

can be found, the tracker is able to continuously update the particle weights using the rich likelihood information. This principle of operation is conceptually different to the classical Tracking by Detection and is able to better take advantage of the low level sensor information.

In a series of real-world experiments, we were able to accurately model the shape of the target-track joint likelihood function where we confirmed that the range and azimuth uncertainties of the camera and RADAR are complementary. Using comprehensive experimental evaluation we showed that target localization performance is dramatically improved in an uncluttered environment which clearly demonstrates the effectiveness of using the raw RADAR signal. In a highly cluttered environment, we also observed gains in localization accuracy of the fusion over the camera-only system. These gains, although significant, are less apparent since there exist a multitude of complex interfering factors that create ambiguities in the likelihood function. Such ambiguities become greater with the increase of distance to the target, since RADAR returns are weaker and camera detections are more uncertain. We conclude that, by using cooperative fusion of RADAR and camera, our system can better detect, track and localize pedestrians and cyclists in an urban environment.

Designing the tracker closely coupled to the sensor data has it's drawbacks. The Particle Filter needs to evaluate the joint likelihood function before detection which means that the RADAR processor must evaluate all particle hypotheses over the range/azimuth/doppler data. This creates memory overhead which scales linearly with the number of targets and the number of particles. To mitigate this issue and still retain real-time operation capability, we implemented all tracking and likelihood computation code in CUDA using the high level Quasar compiler and programming language [17]. In our real-world urban environment experiments we measured an average runtime of $52.9ms$ per frame for the tracking algorithm. This time does not include the CNN object detection and CFAR algorithm which can be offloaded to a separate GPU. A potential weakness in the approach is the assumption of an average person height during the camera back-projection. This introduces large variance in the range estimation from the camera and in cases of people of non-average height can create ambiguities. Our future research will be focused on incorporating the person height in the target state vector, which will be estimated the same way as the person position and velocity.

REFERENCES

[1] "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles; standard j3016 201806," June 2018.

[2] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361, June 2012.

[3] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[4] J. Domhof, R. Happee, and P. P. Jonker, "Multi-sensor object tracking performance limits by the cramer-rao lower bound," *2017 20th International Conference on Information Fusion (Fusion)*, pp. 1–8, 2017.

[5] H. Cho, Y. Seo, B. V. K. V. Kumar, and R. R. Rajkumar, "A multi-sensor fusion system for moving object detection and tracking in urban driving environments," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1836–1843, May 2014.

[6] H. Durrant-Whyte and T. C. Henderson, *Multisensor Data Fusion*, pp. 585–610. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.

[7] M. Heuer, A. Al-Hamadi, A. Rain, and M.-M. Meinecke, "Detection and tracking approach using an automotive radar to increase active pedestrian safety," *2014 IEEE Intelligent Vehicles Symposium Proceedings*, pp. 890–893, 2014.

[8] K. Suzuki, C. Yamano, and N. Ikoma, "Multiple target tracking in automotive fcm radar by multi-bernoulli filter with elimination of other targets," *2018 21st International Conference on Information Fusion (FUSION)*, pp. 527–534, 2018.

[9] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015.

[10] M. Dimitrievski, P. Veelaert, and W. Philips, "Behavioral pedestrian tracking using a camera and lidar sensors on a moving vehicle," *Sensors*, vol. 19, no. 2, 2019.

[11] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, pp. 83–97, 1955.

[12] Y. Xiang, A. Alahi, and S. Savarese, "Learning to track: Online multi-object tracking by decision making," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 4705–4713, Dec 2015.

[13] R. Van Der Merwe, A. Doucet, N. De Freitas, and E. Wan, "The unscented particle filter," vol. 13, 01 2001.

[14] H. Rohling, "Some radar topics: Waveform design, range cfar and target recognition," 2006.

[15] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. DollÃ¡r, "Microsoft coco: Common objects in context," 2014.

[16] M. Dimitrievski, D. V. Hamme, P. Veelaert, and W. Philips, "Robust matching of occupancy maps for odometry in autonomous vehicles," in *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 3: VISAPP, (VISIGRAPP 2016)*, pp. 626–633, INSTICC, SciTePress, 2016.

[17] B. Goossens, "Dataflow management, dynamic load balancing, and concurrent processing for real-time embedded vision applications using quasar," *INTERNATIONAL JOURNAL OF CIRCUIT THEORY AND APPLICATIONS*, vol. 46, no. 9, pp. 1733–1755, 2018.