Automatic labeling of vulnerable road users in multi-sensor data

Martin Dimitrievski*, Ivana Shopovska†, David Van Hamme‡, Peter Veelaert§ and Wilfried Philips¶

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

***Abstract*—A growing interest in technologies for autonomous driving emphasizes the demand for safe and reliable perception systems in various driving conditions. The current state-of-the-art perception solutions rely on data-driven machine learning approaches, and require large amounts of annotated data to train accurate models. In this study we have identified limitations in the existing radar-based traffic datasets, and propose a richer, annotated raw radar dataset. The proposed solution is a semi-automatic data labeling tool, which generates an initial set of candidate annotations using state-of-the-art automatic object recognition algorithms, and requires only minimal manual intervention. In the first qualitative evaluation ever for automotive radar datasets we measure the quality of automatically computed labels under various light conditions, occlusion, behavior and modeling bias based on a multitude of tracking metrics. We determined the specific cases where automatic labeling is sufficient and where a human annotator needs to inspect and manually correct errors made by the algorithms.**

***Index Terms*—automatic labeling, radar, sensor fusion, vulnerable road users**

I. Introduction

Accurate perception of vulnerable road users (VRUs) is central to the deployment of fully autonomous driving. Although the scientific literature offers numerous camera-based systems with excellent performance, there is still a large discrepancy between controlled scientific experiments and real-world situations. Difficult light conditions, adverse weather, clutter and occlusion are rarely identified during evaluation, resulting in unreliable predictions of the model behavior in real-world situations. However, it is in these specific situations where passive optical perception systems often fail, increasing the potential for traffic accidents.

Reinforcing camera-based perception with more robust modalities such as lidar and radar, offers sensor redundancy for fail-safe operation, as well as the potential for increased accuracy by sensor fusion. Due to the abundance of annotated camera and lidar datasets [1], [2], [3], [4], [5], most of the VRU perception research has been focused on camera-lidar fusion. Lidar is an active light sensor, unaffected by ambient light levels. However, it operates in the near-infrared light range and suffers from some of the same weaknesses as cameras. In rain, snow, mist and fog, the accuracy of the lidar data is reduced due to scattering and diffraction of the light beams by particles in the air.

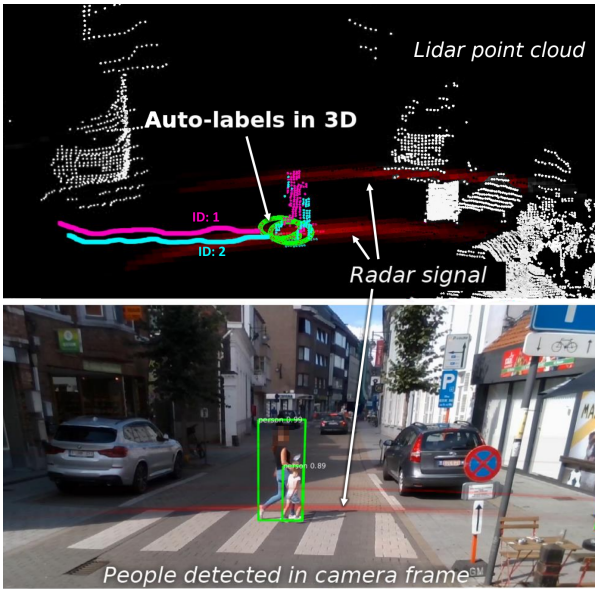

Figure 1. Urban traffic scene with automatically labeled VRUs and their trajectories from the proposed method. Top: 3-D visualization of the data captured by Ouster OS1-128 lidar and TI AWR1443 radar; bottom: 2-D image captured by an Intel Realsense D435 camera (in rgb-only mode) with objects detected by Faster R-CNN [6].

Automotive radar is far less susceptible to weather phenomena than lidar, owing to its much longer carrier wavelength. Its widespread adoption for adaptive cruise control is testament to this robustness. However, automotive radar is rarely relied upon in more complex environments than basic highway driving and niche applications such as parking assistance. One reason for this is the relatively low angular resolution of current off-the-shelf radar compared to lidar and camera systems. This limitation will at least partially be addressed by the advent of "imaging radars" with a higher number of antenna elements [7].

A second limitation of contemporary automotive radars is the poor capability for detecting vulnerable road users (VRUs). VRUs have a much lower radar cross-section (RCS) than vehicles, a measure of a target's reflectivity that depends on factors such as target material, shape and size.

Detecting this faint signal in cluttered urban environments is extremely challenging. However, VRUs can be effectively distinguished from their surroundings by their unique micro-Doppler signatures, information which only radar can perceive. The micro-Doppler signature is a representation of the frequency modulations in the radar return signal, introduced by moving parts of the target. In literature, it has been applied in target classification based on the distinct motion of target's parts [8], [9]. However, none of the existing large automotive datasets created for late sensor fusion offer such raw radar signal, discarding the powerfully discriminative Doppler information.

To advance the efforts in applying learning-based techniques in radar signals, some authors have recorded and annotated small-scale datasets which contain the micro-Doppler signatures of traffic objects such as vehicles. These datasets, however, do not span the wide range of scenarios necessary for robust evaluation of VRU perception. To speed up the data labeling process, authors in the literature have relied on automatically generated labels, weakly supervised by other, more accurate sensors. Frequently, the automatically generated labels are considered as a perfect representation of the ground truth. This assumption can be misleading to radar-based object detection models, as they will be implicitly taught to imitate the errors of the supervising sensors.

Motivated by the scarcity of rich, varied and annotated radar data, in this paper we propose and evaluate a method for automatic and semi-automatic labeling of raw radar data (Figure 1 as an illustration). The main contribution of this paper is in the evaluation protocol, where we analyze the annotation quality against various factors such as light conditions, occlusion level, VRU sub-categories, VRU height, distance and group forming behavior. Our findings are supported by a fully featured tracking dataset of 316 sequences, containing 173K instances of 1318 unique VRUs, captured in several cities in Belgium. Lastly, we offer insights into the cases where automatically generated ground truth can be reliably used for training, as well as cases where we should revert to manual labeling.

II. Related work

Labeling a radar dataset is challenging because it is difficult for a human to recognize objects in radar data [10], [11], [12]. Depending on the complexity of the scenario, it may take several hours to completely label only a few seconds of radar data. Since manual labeling is prohibitively labor intensive for large datasets, automatic pre-labeling followed by manual refinement is essential. In this section, we present an overview of methods for automatic labeling of radar datasets, for the task of 3-D object perception.

The method explained in [13] computes semantically labeled 3-D data for the following categories: flat space, human, vehicle, construction, nature, pole and unknown. The dataset provides dense detection-level data aggregated from 6 radars, however, does not offer the raw radar signal. The authors labeled this data in a semi-automatic manner, relying on a roof mounted 360-degree camera and lidar array. A CNN is applied to semantically segment the camera and lidar data. Subsequently, the proposed semantic radar labels are manually inspected to correct erroneous labels. Finally, this method has been evaluated over three reference test tracks: urban area, small village and an industrial park, but a quantitative evaluation of the annotation quality is lacking. We find this method conceptually similar to our work in that it also employs camera-lidar fused information for automated pre-labeling of radar data. However, this method does not apply tracking, thus the annotations lack the temporal dimension needed for evaluation of tracking performance.

In [10], the authors equip vulnerable road users with hand-held global navigation satellite system (GNSS) modules that communicate to another GNSS module mounted on a vehicle. Reference camera frames, GNSS and radar data are presented to a human annotator on a detection level, i.e., pre-filtered by a constant false alarm rate (CFAR) detector. By comparing the radar image with the camera view, objects are manually identified and individual radar points are assigned a label. A series of experiments were conducted to evaluate the auto-labeling system. However, the experiments are limited to a single pedestrian and a single cyclist, moving along a figure-eight trajectory.

The authors of [11] provide a semi-automatically generated and manually refined 3-D ground truth data for radar object detection. They use an Astyx 6455 HiRes radar, which produces 10x denser data than the radar used in the nuScenes [2] dataset. Yet, it only provides sparse target data rather then the raw radar return signal. The authors of this paper use active learning approach that proposes candidate frames for manual correction, and at the same time improves the proposal model based on the manual input. The resulting dataset is limited to only 500 frames containing 3000 objects of the classes: bus, car, cyclist, person, trailer, truck, with most of the samples falling into the "car" category. This paper, as well as the work in [10], proposes a labeling method based on a similar principle to our work where camera and lidar information is used to label objects in the radar data.

In [14] the authors claim to provide the first high resolution radar dataset of complex urban driving scenes. The annotation was performed by unsupervised dynamic object detection, aided by a Synthetic Aperture Radar (SAR) image of the static environment created using high quality inertial sensors on the vehicle. An initial radar object detection outputs a sparse list of points, each belonging to either a moving object or a stationary object. By projecting these radar points into the static scene (i.e., SAR image), they can distinguish a moving object from a stationary one. All the potential moving object points are projected back into the raw radar data and segmentation is performed around each point. Segmented clusters are subsequently labeled. This pre-training is then supplemented with supervised fine-tuning. The provided dataset consists of over 11,000 moving cars in 27 diverse driving scenes with over 400,000 automatically generated labels of moving cars. Similar to our own work, it is one of the few approaches in the literature which exploit the temporal aspects of captured data to reduce the data

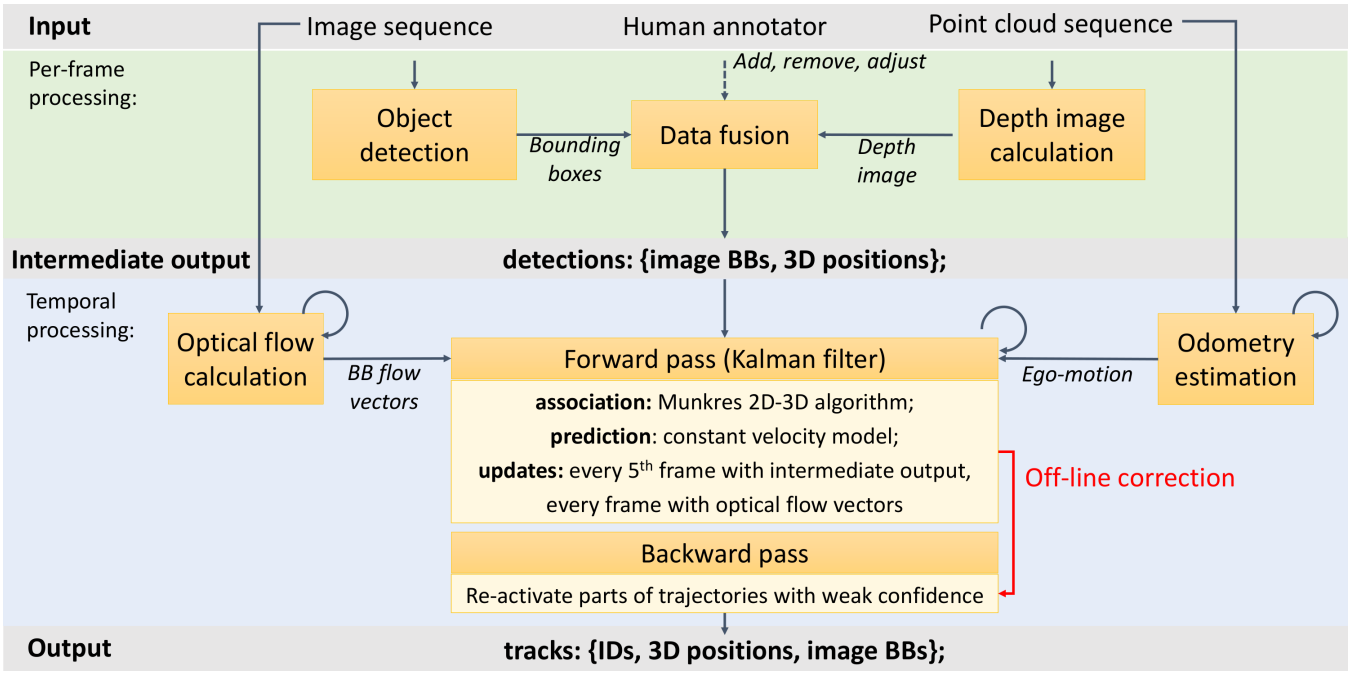

Figure 2. System diagram of the proposed cooperative fusion method: Camera observations are back-projected on the ground plane and fused with RADAR data. Depending on target-track association outcome the tracker uses raw data or joint-likelihood to update the tracks.

annotation effort. Unfortunately, this dataset has no labeled VRUs and no evaluation of the quality of the labels is mentioned.

Finally, the authors in [12] provide a dataset and labeling method that resulted in the most comprehensive radar dataset to date. It includes a large number of labeled road actors in the following categories: car, van, truck, bus, motorbike, bicycle, pedestrian and groups of people, and more importantly, features a variety of weather conditions (including dense fog and heavy snowfall) and driving context (e.g. motorway, suburban). Radar data is provided by a Navtech cts 350-x radar which offers high resolution but does not provide Doppler information. Data labeling is facilitated by an annotation tool which automatically correlates and visualizes multiple sensors through sensor calibration. Similar to our approach, the authors employ a tracker (CamShift) to reduce labeling effort. Unfortunately, pedestrians are underrepresented in this dataset with only 16.5K VRU labels.

III. Proposed method

In this paper we propose a semi-automatic labeling method for generating annotations of the ground plane positions and identities of objects in the radar field of view. The goal is to create ground truth for training and evaluating not just object detection, but also tracking algorithms. The focus on tracking imposes additional requirements on the labels, namely the consistency of object positions and identities over the entire sequence, which complicates the annotation task. Furthermore, any practically feasible automatic annotator should work effectively on data captured over a range of weather and traffic conditions. This objective has been only partially achieved by the state of the art, whereas a complete solution and performance analysis is, thus far, lacking in the literature.

In Figure 2 we present a general system diagram of the proposed annotation tool. The tool relies on multi-modal data, more specifically: RGB video sequences for reliable object recognition, and lidar point clouds for the corresponding depth information needed to reconstruct the 3D structure of the scene. A state-of-the-art object detector is used to find bounding boxes in the RGB image. Simultaneously, 3D lidar information is projected into a corresponding depth image, augmenting the bounding boxes with range data. These 3D object positions are then converted into accurate object tracks by an off-line two-pass tracking algorithm. In the forward pass, the positions of detected objects are associated and stabilized over time,both in the video sequence and on the ground plane. In the backward pass, the accuracy of the positions is further improved and the track's start time is moved forward to the earliest occasion the object was visible. In order to reduce the computational burden of the off-line detection and tracking steps, object detection is performed on every 5-th frame, while the annotations between these frames are interpolated based on predictions from motion models and and optical flow measurements in the image.

The proposed annotation method can operate in a fully unsupervised mode and provides accurate fully-automatic candidate annotations. The association across time and the two-pass temporal stabilization greatly facilitate human correction of the resulting labels. However, as mentioned throughout the text, the quality of the unsupervised annotations can be affected by environmental and behavioral factors. This effect has been greatly overseen by the state of the art and will be a special subject of our evaluation in Section IV. For the

cases where this pre-labeling produces less than ideal output, we created a graphic user interface (GUI) that allows labor-efficient fine-tuning of the labels by adjusting single-frame positions. Each manual adjustment is followed by re-applying the two-pass tracker in order to propagate the manual changes through the complete tracking solution.

Using this methodology, we produce a dataset that includes raw radar data, time-consistent positions and identities in both image plane and 2.5-D ground plane coordinates, targeted specifically at VRU detection and tracking. At the time of writing, this combination of attributes is unique as other available datasets are either limited to target-level radar point clouds, focus on single-frame annotations rather than tracking labels, or contain an amount of VRU labels not suitable for a broad analysis. In the remainder of this section we will define the annotation problem, explain the design of the proposed labeling algorithm in detail, and discuss the choice of sensors involved in the labeling supervision.

Definitions

In this paper, *Vulnerable Road Users* (VRUs) are all traffic participants that are directly exposed during a potential collision. Generally, the class VRU comprises pedestrians, cyclists, people sitting as well as people in wheelchairs, etc.

We define the *raw radar signal* as the received return signal strength across the full range, velocity, azimuth and elevation dimensions offered by the antenna layout. Each individual radar reflection can be described by a simple mathematical model parameterized by: range (the distance of the reflecting object from the radar); Doppler velocity (the rate of change of the range); and direction of arrival (DOA; the direction vector from the radar to the object), often further decomposed into azimuth (horizontal) and elevation angles. The signal reflected from the world is then modeled as a linear superposition of the reflections from all individual objects.

Finally, a unique object in the dataset is defined as any VRU visible for at least a minimum period of time (e.g. $T > 10s$), and is assigned a unique object identifier (ID). Using this information, we can effectively evaluate multi-object tracking with respect to qualitative and quantitative tracking metrics [15], [16].

Proposed Semi-supervised Labeling Method

The proposed annotation tool processes calibrated and synchronized data, captured by other modalities than radar, and outputs ground truth labels intended for training and evaluating radar analysis algorithms. In our case, we use a full HD RGB camera module (Intel Realsense D435 in rgb mode) and Ouster OS1-128 scanning lidar as the supervising sensors. The approach can be easily extended towards other modalities such as infrared imaging or time of flight cameras in future work.

In the camera frames, we run a state of the art object detector [6], with a high object confidence threshold ($p > 0.9$). The output from Faster R-CNN is thus biased towards high precision at the cost of reduced recall. The high confidence threshold is motivated by the fact that the vast majority VRUs relevant for safety decision making will be well detected during at least part of the time they are visible. This configuration therefore ensures that the initial tracks will not be polluted by many false positive detections at frame level, while still providing high recall at object level.

Further, an intermediate fused detection output is formed by augmenting the camera detections with depth information. The lidar provides this depth information, but the recorded point clouds are sparse and do not match image resolution. Therefore we use the very effective depth completion method described by [17], to create a depth image which matches the RGB camera resolution. Data fusion then amounts to associating each VRU bounding box from the RGB detector with a depth value corresponding to the object center. To that end, we estimate the object depth as the median of the depth map values that fall within the central 70% of the detection bounding box.

The intermediate outputs are then tracked over time using a multi-object Kalman Filter (KF). Because object motion is temporally consistent in the real world but not necessarily relative to the sensor (e.g. due to vibration or sudden changes in vehicle attitude), the tracker operates in 2.5-D or ground plane coordinates. Image bounding box coordinates and depth values are therefore projected into an ego-motion compensated coordinate system using odometry calculated from the lidar data as described in [18]. However, it is necessary to also maintain state vectors in the image plane because the image labels cannot be constructed from the ground plane positions alone. Since the relationship between the ground plane and image plane attributes is not easily described by linear models, we chose to decouple the Kalman Filter into two parts. Each object track therefore consists of a dual state vector, one for the state on the image plane and one for state on the ground plane. The former comprises the bounding box shape, position and motion vector in the image, while the latter comprises the position, size and physical velocity on the ground plane. This choice makes the annotated labels to be accurate in both the image and ground plane, increasing the robustness to vibrations caused by the road surface as well as to small sensor calibration errors. For simplicity, and due to the sufficiently high frame-rate of the data, we use a constant velocity model for both image and ground plane state vector prediction.

The initial state space of the Kalman Filtering algorithm is represented by the state estimation error covariance matrices P_1 and P_2, the process noise covariance matrices Q_1 and Q_2, and the measurement noise variances r_1 and r_2 for the image and ground plane filters:

$$P_1 = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \quad Q_1 = \begin{bmatrix} 20 & 0 \\ 0 & 20 \end{bmatrix}, \quad r_1 = 10, \tag{1}$$

$$P_2 = \begin{bmatrix} 2 & 0 \\ 0 & 10 \end{bmatrix}, \quad Q_2 = \begin{bmatrix} 0.05 & 0 \\ 0 & 0.05 \end{bmatrix}, \quad r_2 = 200, \tag{2}$$

expressed in pixel values and meters respectively.

Detection-to-track association is done by optimizing the likelihood between predicted track states and the depth-coupled detections using the Hungarian algorithm [19]. Due to the high redundancy in the camera and lidar data captured at 30FPS, we find it sufficient to update KF using detections at every 5-th frame, while updates in the intermediate frames are performed based on optical flow vectors. The optical flow method we rely on is the CNN-flow method described in [20]. This approach reduces the computational complexity by performing expensive, high-quality object detection at every 5th frame, and relying on more efficient optical flow method in-between.

Building correct tracks from the raw detections is not a straightforward task. Spurious detections, missing detections and poorly placed bounding boxes complicate the association. A considerable time window is necessary to distinguish stable, accurate tracks from transient detection artefacts. The tracking algorithm therefore continually estimates the confidence in each candidate track based on the number of supporting detections and their consistency over time, and only converts such a candidate track into a unique VRU object when the confidence exceeds a threshold. This highly accurate forward pass ensures that only limited to no manual correction of the unique VRU list is needed. However, it also results in an initial delay in identifying unique VRUs when they first enter the sensor viewpoint. Since we are applying the tracking algorithm for off-line generation of ground truth, it is necessary to eliminate this delay in VRU recognition and avoid missing labels. An off-line, backward, pass of the tracking algorithm corrects missing labels caused by the aforementioned delay. More specifically, the backward pass focuses on tracks which have been consistently tracked with a high confidence value, finds the early corresponding track fragments of low confidence, and re-introduces them back in the final output. This approach ensures that only high-confidence tracks will be included in the tracking output, resulting in high-quality annotations.

For each frame, this fully automated annotator outputs is a list of VRUs defined by their unique object identifier, their position on the ground plane and their bounding box in the image. The quality of these initial automatic labels strongly depends on the accuracy of the detection algorithm and the quality of the depth data. As we discussed in the introduction, both camera and lidar have overlapping failure conditions in which automated label quality can be compromised. In such cases, a human annotator can often resolve ambiguities based on detailed analysis of the scene context, as the ability of a human to take broader context into account still surpasses that of state of the art detectors. Our framework is therefore integrated with a GUI to verify and correct object labels in individual frames, and to re-compute a new automatic tracking output including these manual corrections.

The proposed GUI (implemented in MATLAB) visualizes the sensor data, displays the proposed automatic annotation solution, and offers the user to inspect and modify the labels. When deemed appropriate, the human expert can add, remove or adjust detections in the image or on the ground plane. All manual interventions in the object bounding boxes maximize their confidence scores, since human input is considered as confirmation for the existence and the location of an object. The tool then re-runs the two-pass tracker to update the track labels and displays the new optimal solution taking into account the human input.

IV. Experiments and evaluation

The main focus of the work described in this section is to evaluate the impact of various real-world criteria on the quality of automatically labeled VRUs. We aspire to provide an answer to the question: "Under which conditions is the fully-automated labeling data sufficient to be used directly for training new algorithms, and in which cases must a human annotator be employed?". To that end, we will compare the annotations from the method described in Section III running in fully automatic mode, to the annotations obtained when a human annotator is included in the loop. In our evaluation we test the impact on annotation quality of the following environment and scene conditions: ambient light level, level of occlusion, VRU sub-category, ego-motion, object height and group forming behavior. Quality is expressed in terms of detection and tracking metrics, outlined in the MOT16 benchmark [21]. This methodology combines both quantitative CLEAR metric [15], and qualitative Track Quality Measure [16]. Additionally, we measure tracking latency in terms of Track Initialization Duration (TID) and report the detection Average Precision as well as the F_2 score. For detailed discussion of these metrics, we refer to [21].

Our dataset consists of a total of 316 sequences of duration between $10s$ and $20s$, captured throughout the year 2020 in several cities across Belgium. These sequences are representative of the difficult driving scenarios encountered in typical large cities in Western Europe: parked vehicles obscuring the view, mixing of motorized traffic with cyclists and pedestrians, occasionally narrow streets and an abundance of poorly thought out road infrastructure. The fully automatic labels contain 1936 unique VRUs across 179422 instances, while the manually corrected labels contain 1318 unique VRUs across 173095 instances. Ground truth is evaluated within the range of $[0m, 20m]$ and an azimuths of $\pm 35°$ (the field of view of the Realsense D435 camera) while the area beyond these ranges is ignored. We apply a spatial gate of $2m$ for accepting true positives.

Hereby we elaborate the details of the evaluation criteria:

- Light conditions: depending on the time of day of the capture, we split the data into three categories: "daytime" (160 sequences), "twilight" (98 sequences), and "nighttime" (58 sequences).
- Occlusion level: depending on the level of occlusion in the scene, we split the dataset into three categories: "visible" (where less than 10% of VRUs are occluded; 136 sequences), "partly occluded" (where 10% to 33% of VRUs are occluded; 136 sequences), and "occluded" (more than 33% of VRUs are occluded; 44 sequences). The level of occlusion of a VRU instance is an automatically computed categorical value. For each labeled VRU,

Split \ Metric	MOTA↑	MOTP↓ [m]	AP↑	F_2↑	MT↑ [%]	PT↓ [%]	ML↓ [%]	TID↓ [ms]	IDS↓ [%]	FRAG↓ [%]
Dataset average	**0.647**	**0.277**	**0.833**	**0.836**	**69.8%**	**18.2%**	**12.0%**	**232**	**0.45**	**0.81**
Daytime	0.702	0.284	0.868	0.867	76.5%	16.0%	7.5%	197	0.58	0.91
Twilight	0.655	0.264	0.837	0.839	67.9%	19.9%	12.2%	213	0.36	0.74
Nighttime	0.471	0.294	0.722	0.737	53.6%	20.5%	25.8%	400	0.34	0.75
Visible	0.705	0.257	0.871	0.867	77.0%	15.2%	7.8%	221	0.41	0.73
Partly occluded	0.622	0.287	0.818	0.823	68.4%	19.7%	11.9%	229	0.48	0.90
Occluded	0.465	0.342	0.705	0.728	54.8%	20.2%	25.0%	277	0.46	0.73
Mostly pedestrians	0.685	0.274	0.839	0.844	71.8%	19.3%	9.0%	196	0.47	0.85
Mixed	0.623	0.274	0.848	0.840	72.9%	14.6%	12.6%	252	0.56	0.78
Mostly cyclists	0.628	0.280	0.823	0.827	66.3%	19.3%	14.4%	253	0.38	0.80
Mostly short	0.601	0.306	0.756	0.778	60.3%	20.3%	19.3%	323	0.44	0.86
Mixed	0.661	0.270	0.853	0.850	73.3%	18.0%	8.7%	199	0.48	0.84
Mostly tall	0.661	0.251	0.924	0.891	76.7%	7.0%	16.3%	160	0.03	0.25
Static platform	0.657	0.225	0.856	0.851	69.2%	23.1%	7.7%	282	0.40	0.69
Moving platform	0.643	0.299	0.825	0.829	69.9%	17.3%	12.8%	222	0.47	0.87
No groups	0.636	0.276	0.836	0.836	69.4%	18.0%	12.7%	226	0.39	0.71
Contain groups	0.697	0.282	0.821	0.836	72.8%	19.7%	7.5%	266	0.72	1.30

Table I

TRACKING PERFORMANCE OF AUTOMATICALLY GENERATED LABELS ACROSS DIFFERENT CONTROL VARIABLES. COLORS INDICATE THAT THE SPECIFIC CONDITION IS SIGNIFICANTLY BETTER THAN DATASET AVERAGE OR WORSE THAN DATASET AVERAGE.

we compare their ground truth distance to the median distance of the corresponding depth image patch. If the discrepancy of the labeled depth and the visible depth is more than 5m, then the VRU instance is marked as occluded.

- VRU subcategory: depending of the object's velocity, we split the data into three categories: "mostly pedestrians" (112 sequences), "mixed" (60 sequences) and "mostly cyclists" (149 sequences). We consider VRUs moving faster than $6Km/h$ as cyclists. Manual inspection of the resulting split on a small random sample confirms that speed is a good indicator for VRU subcategory.
- VRU height: depending on the object's height, we split the data into three categories: "mostly short" (97 sequences), "mixed" (200 sequences) and "mostly tall" (19 sequences). A sequence consists of mostly short or mostly tall people if more than 50% of the instances have height bellow $1.5m$ or above $1.75m$, respectively.
- Ego-motion: depending on the ego-velocity, we split the data into two categories: "static" (54 sequences) and "moving" (262 sequences). A static scene is one where the average ego-velocity is below $10Km/h$.
- Number of groups: depending whether there is group forming behavior in the sequences, we split the dataset into two categories: "no groups" (279 sequences) and "contains groups" (37 sequences). A group is formed when two VRUs are within $0.745m$ distance of each other [22]. We consider a sequence to contain significant group behavior if groups have been identified in at least 30% of the frames.

In Table I we present the results of our quantitative evaluation. Overall, the automatically generated labels are of high quality, achieving an average positional deviation of just 28cm from the human corrected labels and an AP of 83.3%. These results indicate that even for this difficult urban dataset, in the majority of cases the human annotator made only small corrections to the position of the proposed labels, and the vast majority of objects were already well detected and tracked. We compare these dataset average values of detection and tracking metrics to the ones computed under various evaluation conditions. Whenever the metric of the specific condition is significantly worse than the dataset average, this indicates that parts of the automatically labeled data are of poorer quality and the corresponding field in the table is marked with red color. On the other hand, green color represents conditions under which automatic labels have higher quality than the dataset average.

All tracking and detection metrics in "daytime" and "twilight" are consistently at or above the dataset average. This indicates that automated labeling is more effective when the camera object detector is operating in good light conditions, and the fully automatic annotations are reliable. However, the quality of the labels for "nighttime" are significantly lower than the average. This is especially evident for the tracking

Method	Scene complexity	Time per frame	Label quality (AP)
Automatic pre-labeling	Easy	1.2s	87%
	Difficult	1.2s	72%
Manual correction w/ pre-labeling	Easy	1.8s	100%
	Difficult	4.0s	100%
Fully manual	Easy	20s	baseline
	Difficult	40s	baseline

Table II
COMPARISON OF THE ANNOTATION EFFORT NEEDED TO LABEL DIFFERENT SEQUENCES IN OUR DATASET. EASY: DAYTIME SEQUENCES WITH FEW UNOCCLUDED VRUS; DIFFICULT: NIGHTTIME SEQUENCES WITH FEW OCCLUDED VRUS.

accuracy (MOTA), fraction of tracked trajectories (MT) and average precision (AP), which show marked decrease, while the fraction of lost tracks (ML) and the tracking delay (TID) are significantly higher. We conclude that in night time sequences, a manual intervention is often warranted.

Another major factor in the quality of automatic annotation is the level of occlusion. All tracking and detection metrics for the "visible" category are significantly above the dataset average, while in the case of "partly occluded" we observe below average quality across all metrics. The automatically computed labels in the "occluded" category suffer the highest loss of quality among all evaluated criteria and it is in these cases that human intervention is most needed.

In the case of labeling quality of different VRU sub-categories, there is a slight advantage for the category of pedestrians versus cyclists. We deem that the annotation quality is almost equally good in both cases and these labels can be used for supervised training.

In the case of VRU height, our automatic labeling method gives an advantage for the category "mostly tall". This is especially prominent in the detection metric AP, with more than 10% above the dataset average. Unfortunately, the automatically computed labels for the category "mostly short" suffer from a significant loss of quality as measured by all tracking and detection metrics. This finding is not surprising, since shorter objects cover a smaller image area, affecting the accuracy of detection and classification. We note that the split based on height is also dependent on the demographic distribution.

With regard to ego-motion, there is a slight drop in annotation quality when the vehicle is moving, as opposed to when it is static. However, tracking and detection performance are sufficient for both cases and no systematic additional human intervention is warranted for faster ego-motion scenarios.

In the analysis of group forming behavior we observed mixed results. The quality of the automatic labels measured by MOTA is slightly higher when the sequence contains groups, while the detection performance measured by AP is slightly lower, which is counter intuitive. We observe the same inconclusive pattern when analyzing the proportion of mostly tracked, partly tracked and mostly lost trajectories. Possibly the parameters which define a group, which were adopted from [22], need to be further tuned to the specific traffic situations, or we need to further differentiate between group layouts e.g. longitudinal queues or pairs of pedestrians walking side by side. The group forming behavior warrants further, more detailed, analysis.

We conclude this section with an estimate of the gains in human annotation effort when using the proposed method. Since measuring human effort is a difficult task depending on highly subjective inputs, we refrain from measuring complex effects such as loss of attention, annotator fatigue or annotator discrepancy. Instead, in Table II we present only the average time needed to label different sequences (expressed as time per frame) as a general metric of annotation effort. Based on these results, we conclude that the proposed pre-labeling followed by manual correction speeds up the annotation by a factor of 6-7x as compared to fully manual labeling. Moreover, during manual labeling, we encountered significant annotator fatigue setting in much faster than when using a pre-labelled solution. This limiting factor reduces the total amount of sequences a person can label in a given work day and requires further investigation.

V. CONCLUSION

In this paper we proposed a semi-automated annotation method for labeling VRUs based on fused RGB-lidar data. The goal of the labeling effort is to provide an accurate ground truth for a large-scale radar dataset designed for VRU detection and tracking. The proposed method can operate in fully automated mode and produces labels that are generally accurate and mostly error-free. However, the quality of annotations is affected by multiple scene and environment factors. By comparing automatically generated labels to ones adjusted by human annotators, we have carried out a thorough evaluation of the labeling quality in various light conditions, occlusion levels, VRU sub-categories, VRU height, ego-motion and group forming behavior.

Based on this experimental evaluation, we found that the automatically computed labels suffer from significantly reduced tracking quality at nighttime and in the presence of occlusion. Furthermore, we found that the VRU height is also a significant factor which correlates with the label quality. Ego-motion did not significantly affect label quality, and performance was roughly equal for pedestrians and cyclists. Furthermore, we observed weak correlation for the factor

of group forming, where more investigation will be carried out in future work. Finally, we presented an analysis of the reduction of labeling effort when using the proposed method opposed to full manual labeling by a human expert.

Our future work will be focused on further extending this dataset in conditions such as fog, rain and glare, currently underrepresented in our data. Moreover, the quality of the labels will be further validated by training deep learning based analysis methods on the raw data in our dataset. We hope that our semi-supervised labeling framework will empower future radar applications by reducing the human effort required to obtain large volumes of labeled data.

ACKNOWLEDGMENTS

This research received funding from the Flemish Government (AI Research Program) and was conducted in collaboration with HiSilicon[1].

REFERENCES

[1] 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.

[2] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," *CoRR*, vol. abs/1903.11027, 2019.

[3] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell, "BDD100K: A diverse driving video database with scalable annotation tooling," *CoRR*, vol. abs/1805.04687, 2018.

[4] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platinsky, W. Jiang, and V. Shet, "Lyft level 5 perception dataset 2020." https://level5.lyft.com/dataset/, 2019.

[5] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2446–2454, 2020.

[6] 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.

[7] S. Brisken, F. Ruf, and F. Höhne, "The recent evolution of automotive imaging radar and its information content," *IET Radar, Sonar and Navigation*, vol. 12, 04 2018.

[8] J. A. Nanzer and R. L. Rogers, "Bayesian classification of humans and vehicles using micro-doppler signals from a scanning-beam radar," *IEEE Microwave and Wireless Components Letters*, vol. 19, no. 5, pp. 338–340, 2009.

[9] I. Bilik and P. Khomchuk, "Minimum divergence approaches for robust classification of ground moving targets," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 48, no. 1, pp. 581–603, 2012.

[10] N. Scheiner, N. Appenrodt, J. Dickmann, and B. Sick, "Automated ground truth estimation of vulnerable road users in automotive radar data using gnss," in *2019 IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM)*, pp. 1–5, 2019.

[11] M. Meyer and G. Kuschk, "Automotive radar dataset for deep learning based 3d object detection," in *2019 16th European Radar Conference (EuRAD)*, pp. 129–132, 2019.

[12] M. Sheeny, E. D. Pellegrin, S. Mukherjee, A. Ahrabian, S. Wang, and A. Wallace, "Radiate: A radar dataset for automotive perception," 2020.

[13] S. T. Isele, M. P. Schilling, F. E. Klein, and M. J. Zoellner, "Annotating automotive radar efficiently: Semantic radar labeling framework (seralf)," in *Conference on Neural Information Processing Systems*, 2020. 47.01.02; LK 01.

[14] M. Mostajabi, C. M. Wang, D. Ranjan, and G. Hsyu, "High resolution radar dataset for semi-supervised learning of dynamic objects," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 450–457, 2020.

[15] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, p. 246309, May 2008.

[16] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors," *International Journal of Computer Vision*, vol. 75, pp. 247–266, Nov 2007.

[17] J. Ku, A. Harakeh, and S. L. Waslander, "In defense of classical image processing: Fast depth completion on the CPU," *CoRR*, vol. abs/1802.00036, 2018.

[18] 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.

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

[20] D. Teney and M. Hebert, "Learning to extract motion from videos in convolutional neural networks," 2016.

[21] A. Milan, L. Leal-Taixe, I. D. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," *CoRR*, vol. abs/1603.00831, 2016.

[22] F. Zanlungo, D. Brscic, and T. Kanda, "Pedestrian group behaviour analysis under different density conditions," *Transportation Research Procedia*, vol. 2, pp. 149–158, 2014. The Conference on Pedestrian and Evacuation Dynamics 2014 (PED 2014), 22-24 October 2014, Delft, The Netherlands.

[1] www.hisilicon.com