Tracking Road Users by Cooperative Fusion of Radar and Camera Sensors

Martin Dimitrievski*, David Van Hamme†, Lennert Jacobs‡, Peter Veelaert§, Heidi Steendam¶ and Wilfried Philips‖
TELIN-IPI, Ghent University - imec,
St-Pietersnieuwstraat 41, B-9000 Gent, Belgium
Email: *martin.dimitrievski@ugent.be, †david.vanhamme@ugent.be, ‡lennert.jacobs@ugent.be,
§peter.veelaert@ugent.be, ¶heidi.steendam@ugent.be, ‖wilfried.philips@ugent.be

***Abstract*—In this paper, we present the results and the lessons learned during development of an intelligent automotive perception system based on fusion of camera and radar. Our cooperative sensor fusion approach closely couples the detector and tracker, which allows to exploit the benefits of low-level fusion of the rich sensor data without the need for high data bandwidth to the fusion center. To demonstrate the accuracy of our perception system, we captured a dataset consisting of various complex traffic scenarios in a European city center. For multiple targets up to 10 m in front of the ego-vehicle, we measure accuracy improvements of 20% over the camera-only system, with only a fractional load to the network compared to low-level fusion.**

I. Motivation and significance of the topic

Accurate detection and tracking of road users is essential for driver-less cars and many other smart mobility applications. As no single sensor can provide the required accuracy and robustness, the output from several sensors needs to be combined. If we exclude practical limitations such as cost and ease of integration, each sensor technology (e.g., radar, video, LiDAR, ultrasound) still has its own intrinsic limitations. For instance, cameras don't work well at night-time, or in dazzling sunlight. Radar can be confused by reflective metal objects, like rubbish bins or soda cans. LiDAR technology is affected by atmospheric conditions that increase the light scattering of the air such as precipitation, mist, fog or fine dust. Fusing the output of these different sensors is thus very important. While both vision based [1], [2] and ranging based [3] tracking of road users are mature research fields well described in literature, fusion techniques between the two specifically for the automotive context are much less covered in literature.

Currently, sensor fusion happens at a relatively late stage, after each sensor has performed object detection based on its own limited collection of sensor data. Such late fusion has obvious benefits on the system level, both in terms of ease of integration and robustness. However, a lot of sensor fusion potential is lost, especially in circumstances where one sensor underperforms compared to another. Employing so-called early or low-level fusion on the rich data gathered from all sensors could improve the performance, but at the same time significantly increases the required data bandwidth and processing power. For instance, the GoPro camera and TI AWR1443 automotive radar used in our experimental setup have a raw data rate of about 30 Mbps (at 30 fps, using video compression) and 160 Mbps (at 20 fps, uncompressed), respectively. A single contemporary automotive 3D LiDAR produces up to 9.6M points per second with data rates of up to 500 Mbps.[1] Despite these large numbers, the amount of sensors added to smart vehicles is expected to dramatically increase in order to reach higher levels of autonomy. In addition, vehicle-to-everything (V2X) communication, which allows, e.g., to share traffic information or sensor data among vehicles, high-precision navigation systems, and advanced infotainment will also contribute to the anticipated explosion of data bandwidth in future cars. For this reason, car manufacturers are currently adopting high-speed Ethernet-based networks in addition to conventional data buses like CAN and LIN. Whereas 100 Mbps (100BASE-T1) Ethernet over single-pair UTP cables has been applied in cars today, 1 Gbps (1000BASE-T1) Ethernet will be introduced in next-generation car platforms, and new standards for 2.5-10 Gbps automotive Ethernet are being developed [4]. Nevertheless, the massive amount of sensor data requires to offload computation as much as possible from the Fusion Center (FC) towards the edge nodes to save both bandwidth and computation power. In this regard, we have developed a cooperative sensor fusion approach which closely couples the detector and tracker without the need for high data bandwidth. The cooperative approach embodies fusion on two levels. Firstly the processing pipelines of different sensors directly exchange mid-level information such as regions of interest, which allows the sensors to resolve ambiguities during the detection process. These detections then continuously update or spawn stable tracks, providing consistent and accurate object localization. In cases where detection or data association fails, the tracker evaluates the likelihood of multiple hypotheses by communicating the positions of particles to the radar and camera processing modules. This way, the tracker-detector feedback loop bypasses the original detection thresholds and thereby unlocks the potential of low-level data fusion without losing the aforementioned benefits of independent sensor pipelines on the system level.

II. Conclusions

It is well established that detection of road users using radar produces targets with excellent range, but poor angular

[1] https://velodynelidar.com/vls-128.html

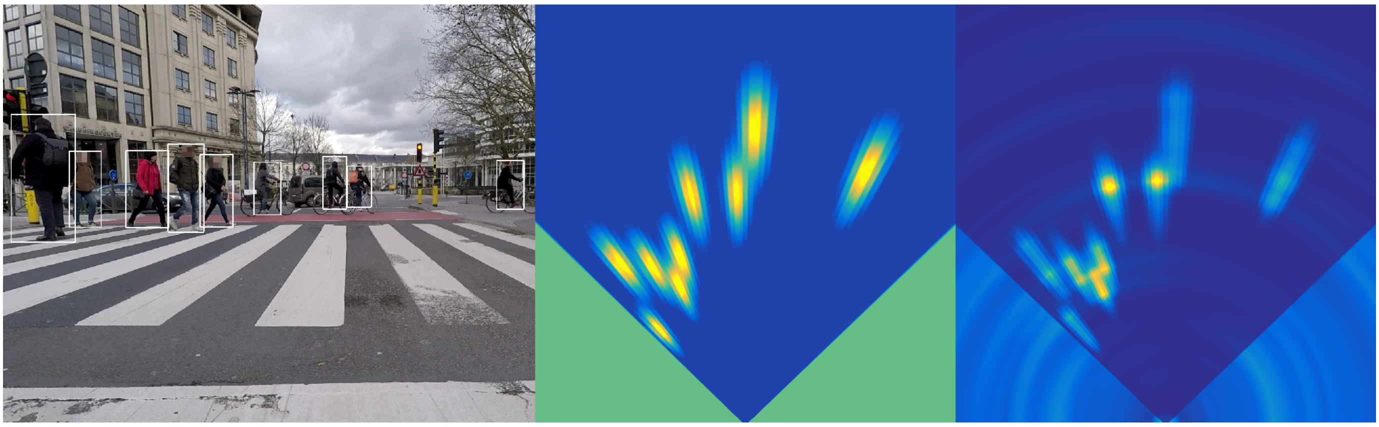

Figure 1. Sample from the data captured in the city of Ghent showcasing data interaction at the sensor level. Left: camera frame with objects detected by Faster R-CNN; Center: uncertainty regions on the ground plane generated by back-projecting CV objects; Right: uncertainty regions on the ground plane fused with radar (averaged over all Doppler velocities).

resolution. On the other hand, by using Computer Vision (CV) and projective geometry, objects detected by a camera sensor have poor range, but excellent angular resolution. Naïvely, one can expect that combining such complementary information will lead to dramatic performance increase over any singular sensor. However, during our study we found out that tracking performance is greatly hindered by ambiguities that exist even in the fused data. For example, in figure 1 we see a typical situation where multiple pedestrians and cyclists are detected in the camera frame. These regions of interest (Bounding Boxes) are communicated to the radar processing module and projected on the ground plane (e.g. using expected person height) taking into account the uncertainty on these projected locations. While augmenting this ground plane map with the radar signal greatly reduces the range uncertainty in spatially separated objects such as the cyclists in the background, it is clear that ambiguities still exist.

Candidate targets are obtained at the edge nodes by employing Constant False Alarm Rate (CFAR) detection on the fused data, thereby avoiding transmission of raw data to the FC. The algorithm extracts detections with high signal-to-noise ratios which are then fed into our multiple object tracker. A particle filter maximizes the belief in the state of each road user in an iterative prediction/update cycle. We have shown that in the context of autonomous driving, particle filters [5] are well suited for tracking unpredictable pedestrian motion by keeping multiple hypotheses. State updates are done using a joint-likelihood observation model. On the other hand, when tracks fail to associate with detections, or there are no detections above a required threshold, most trackers in the literature rely only on the motion model to predict the next state. This tracking scenario occurs often in ambiguous situations, e.g. a person is unobserved for multiple frames and can appear at multiple locations, thus the probability density map of their location is multi-modal. Our tracker is different because it can intelligently update unassociated tracks by evaluating each hypothesis (particle) over the low-level camera-radar data. We call these sub-threshold or soft-associations where the weight of each particle is inversely-proportional to the uncertainty in the camera-radar data, figure 1 right.

For each uncertain association, the particle filter communicates 1000 predicted particles to the sensor processing modules, with each particle consisting of 4 parameters (2D position and velocity). For a single track, we observe a peak bit rate of 2.4 Mbps streaming from the FC to the sensors and 0.6 Mbps worth of likelihoods streamed in the other direction. This is rare occurrence though, as most of the targets can be continuously updated with observations and very rarely does the tracker need to evaluate likelihoods of the full particle cloud. Experimental results on tracking vulnerable road users (pedestrians and cyclists) show that the benefits of fusing camera with radar increase with the decrease of scene complexity. We observed that the most significant improvement in accuracy occurs when there is a single person in front of the ego-vehicle with localization errors improving by 50% over the camera-only system. On the other hand, when the scene becomes very complex, smaller performance gains are observed. Nonetheless, we measured improvements of 20% for multiple targets up to 10 m in front of the ego-vehicle. These results are remarkable since they show that although the radar cross section of a pedestrian is relatively small compared to clutter such as fences, light poles etc. our smart system can extract more information from the two sensors exerting only a fractional load to the network compared to low-level fusion.

References

[1] S. Sivaraman and M. M. Trivedi, "Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, pp. 1773–1795, Dec 2013.

[2] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: An experimental survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, pp. 1442–1468, July 2014.

[3] S. M. Patole, M. Torlak, D. Wang, and M. Ali, "Automotive radars: A review of signal processing techniques," *IEEE Signal Processing Magazine*, vol. 34, pp. 22–35, March 2017.

[4] G. W. d. Besten, "30.1 single-pair automotive phy solutions from 10mb/s to 10gb/s and beyond," in *2019 IEEE International Solid- State Circuits Conference - (ISSCC)*, pp. 474–476, Feb 2019.

[5] 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.