Augmented Perception with Cooperative Roadside Vision Systems for Autonomous Driving in Complex Scenarios

Stefano Masi, Sio-Song Ieng, Philippe Xu, Philippe Bonnifait

Augmented Perception with Cooperative Roadside Vision Systems for Autonomous Driving in Complex Scenarios

Stefano Masi[1], Sio-Song Ieng[2], Philippe Xu[1] and Philippe Bonnifait[1]

***Abstract*— Performing autonomous driving in urban environments is a challenging task, especially when there is a reduced visibility of traffic participants in complex driving scenarios. For this reason, we investigate the advantages of cooperative perception systems to enhance on-board perception capabilities. In this paper, we present a cooperative roadside vision system for augmenting the embedded perception of an autonomous vehicle navigating in a complex urban scenario. In particular, we use an HD map to implement a map-aided tracking system that merges the information from both on-board and remote sensors. The road users detected by the on-board LiDAR are represented as bounding polygons that include the localization uncertainty whereas, for the camera, the detected bounding boxes are projected in the map frame using a geometric constrained optimization. We report experimental results using two experimental vehicles and a roadside camera in a real traffic scenario in a roundabout. These results quantify how the cooperative data fusion extends the field of view and how the accuracy of the pose estimation of perceived objects is improved.**

I. Introduction

During the past decades, research on cooperative perception systems has been growing significantly. Individual autonomous vehicles perception is always constrained by the range of on-board sensors. To enhance vehicles field of view perception ability, improving detection accuracy and navigation safety, different vehicles need to exchange information [5], [23], [1]. V2X communication offers an appealing solution to share perception adding connectivity to vehicles by means of ETSI or SAE standards [13], [12].

Shared perception information can be combined via data fusion techniques to improve the accuracy in detected objects [25], [15], [24]. This technique allows both to improve the detection of obstacles in the driving environment via multi-sensor data fusion [11] and to enhance safety in autonomous vehicles navigation [16]. The main advantage is that data collected from different sensors may contain complementary information [6], and data collected from remote sensors can help in filling blind spots [17]. To achieve data fusion, techniques such as the Kalman Filter [14], Extended Kalman Filter [21], Split Covariance Intersection Filter [18] and others have been used.

Furthermore, shared perception can also be broadcast from a remote intelligent roadside infrastructure [26]. In particular, infrastructure can be used to a wide range of use cases, from sensing the driving environment providing additional information about road users [20] to help autonomous vehicles for cooperative driving maneuvers [7].

To test cooperative perception, nowadays the majority of research in this field uses simulated or hybrid data as in [9], or some existing datasets such as [8]. For this reason, there is the necessity of high quality referenced data from fixed, reproducible and static environments to test and compare the performance of cooperative perception systems [8]. However, according to [8], many state-of-the-art datasets such as KITTI [10] are not suited for cooperative perception because they are focused on standalone perception, while other datasets such as [22], [29] are focused on predicting intentions rather than cooperative perception.

In this paper, we present a cooperative system composed of a roadside camera than communicates with an AD vehicle which is itself equipped a LiDAR sensor. The system is validated experimentally in a roundabout scenario. First, we introduce in Section II how a High Definition (HD) map can be used in all the subsystems to help the navigation task. We then introduce in Section III the map-aided on-board LiDAR-based perception system. In Section IV, we detail the image-based vehicle detection from the roadside camera and how to compute the pose of the detected vehicles in the map frame. Next, Section V introduces our map-aided road objects tracking system that can be used for both the camera or LiDAR data but also to do the fusion of the two. Finally, we present experimental results in Section VI using real data acquired with two experimental vehicles driving in a roundabout in the city of Compiègne, France.

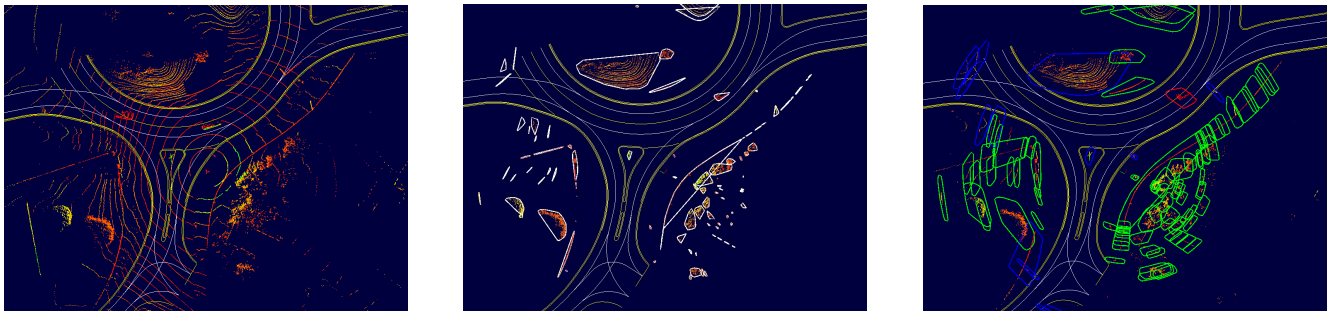
II. HD Maps

Maps play naturally a central role in vehicle navigation [19], but when enriched with detailed and precise information about the environment, as in HD maps, they can serve multiple additional purposes. In this paper, we consider an HD map containing information about driving lanes (represented by a center line and some borders, e.g., lane markings, road edges, pedestrian crossings, etc.) and also the road infrastructure such as traffic signs, poles, traffic lights, etc. The geometric position of these elements coded as points or sequences of points (called polylines) are considered to be centimeter accurate.

In cooperative ITS, when a common map is shared among the users, it also serves easily as a common working frame to exchange information. In the rest of the paper, we assume that all the agents are able to provide information within a local ENU (East-North-Up) frame.

[1]The authors are with Université de Technologie de Compiègne, CNRS, Heudiasyc, UMR 7253, Compiègne France.

[2] The author is with Université Gustave Eiffel, COSYS PICS-L Champs sur Marne, France



(a) (b) (c)

Fig. 1: Road users detection pipeline. (a): Raw data LiDAR point cloud. (b): Ground filtering and bounding polygons computation. (c): Bounding polygons with localization uncertainty injection and map filtering to filter objects that do not belong to the road surface. The colors green, red and blue indicate respectively out of the road, on the road and on the boundary.

In the following, the HD map will be used for different tasks. First, the geometry of the lanes is used within the embedded perception system in order to filter out the obstacles that are not lying on the road surface. This enables to focus the perception only on road users. Second, the map is used for the extrinsic calibration of the infrastructure camera. This calibration also serves to estimate the homography of the road surface from the HD map onto the image frame. And inversely, this enables to project the detected object in the image frame back to the HD map frame. For the object detection task itself, the HD map also helps to estimate the heading of the detected object by assuming that they generally follow the curve of the driving lanes. Finally, the center lines of the driving lanes, which are representative of the vehicles trajectories, are used to assist the tracking system. Intuitively, the HD map improves the evolution model of the tracking as it provides additional constraints between the tracked trajectories and the shapes of the lanes.

III. On-board Perception with 3D Lidar

In our system, we exploit a Velodyne VLP23 LiDAR sensor to scan the driving environment. The detection pipeline is composed of the following main blocks. Further detail about this detection pipeline can be found in [3]:

a) Ground segmentation and object clustering: To detect road users from, we first separate the LiDAR point cloud belonging to the ground from the rest. Then, from the remaining points, we exploit a clustering algorithm to individuate and group LiDAR points that belong to a same object. To achieve this task, the algorithm proposed by Zermas *et al.* [28] is used.

b) Bounding polygon: This step takes as input the clusters obtained in the previous step and computes, for each cluster, a 2D convex bounding polygon that bounds the projections of the LiDAR points on the ground plane. To perform this computation, the monotone chain algorithm [30] has been used.

c) Map filter: Next, we use an HD map to filter out bounding polygons that are not located within the driving space, *i.e.*, polygons that have an empty intersection w.r.t. the driving space. To do so, the bounding polygons need to be transformed from the LiDAR sensor frame to the map frame via the AD vehicle localization. The uncertainty of the localization information is taken into account by extending the bounding polygon as in [3].

Figure 1 illustrates the subsequent steps of the raw point cloud data treatment. Note that steps *a)* and *b)* can be replaced by other LiDAR-based object detection methods of the literature.

IV. Perception from the infrastructure

The contributions of a cooperative perception system installed in the infrastructure are a wider and complementary field of view to the on-board sensors. This makes it possible to handle occlusion problems in complex areas, such as roundabouts, where multiple road users interact with each other. In this work, we assume that information from infrastructure is trustworthy and no infrastructure fail can happen. Integrity and trustworthiness of received data is out of the scope of this work. To be useful for the AD vehicle, the perception information from the infrastructure needs to be expressed in a common frame. In our study, our goal is to compute 2D Euclidean bounding boxes in the HD map frame from bounding boxes expressed in the image frame.

A. Obstacle detection and size estimation

Image-based obstacle detection can be solved efficiently with state-of-the-art deep learning approaches such as YOLO [4] or R-CNN family detectors [27]. These methods detect and classify objects by returning bounding boxes (BBox) in the image frame. From a monocular camera setup, it is difficult to retrieve 3D information such as the size of a vehicle (length and width). However, as the detection network is capable of differentiating different types of vehicle, it is possible to approximate their sizes. Indeed, the same kind of vehicle has approximately the same dimension in Europe. The dimensions of passenger cars are quite homogeneous and the length is in the majority approximately 5.0 m and the width is approximately 1.8 m. The truck dimensions show more disparity in length and in width. It is difficult to distinguish the smallest trucks from the biggest personal cars. Buses have well-defined dimensions in France: the classical

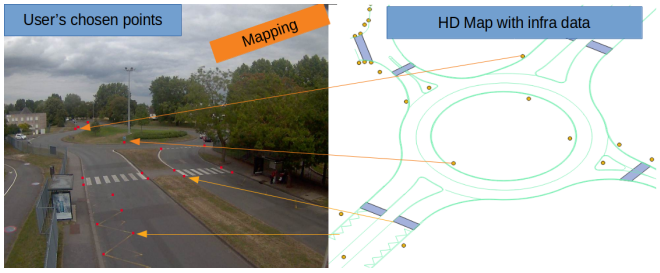


Fig. 2: Some road information which visible and georeferenced in the HD map is mapped into the image to perform the camera calibration.

model has a length of 12 m and the width is 2.50 m. The articulated version is longer with 18 m.

B. Camera calibration using HD Map

The perception system observes a roundabout with a very large area (approximately 4200 m^2). The perception system is calibrated by estimating the homography from the HD map to the image plane, using the information available on the HD map: ENU location of traffic signals, street lampposts and road markings. The extrinsic calibration corresponds to the estimation of the parameters of the rotation matrix and of the translation vector which relates the 3D world coordinates frame to the image coordinates system. The estimation of the parameter is illustrated in Figure 2.

C. Position and heading estimation of detected vehicles

In the rest of this section, we consider that the detectors provide bounding boxes along with the estimated dimensions (length ℓ and width w) and that the camera is well calibrated. The goal is now to compute the 2D top-down bounding boxes in the HD map coordinates frame. By projecting a BBox of a detected vehicle in the ground plane, we obtain a trapezoid-like quadrilateral ($(ABCD)$ in Fig. 3) within which the algorithm extracts the position of the vehicle and its heading (represented as the rectangle $(EFGH)$ in Fig. 3).

When the vehicle is seen from the front/back or from the side, the heading is easy to derive from the HD map. We will now be interested in the case where the vehicle is seen in an intermediate position between these two positions (front and side) as depicted in Figure 3. In this situation the goal is to find the geometric configuration such that $E \in [AB]$, $F \in [BC]$ and $H \in [AD]$.

To solve this problem, we constrain $E \in [AB]$ and $H \in [AD]$ and parameterize the problem with $\lambda = |AE| \in (0, w)$. Thanks to the calibrated camera, we can compute the ENU coordinates of $A = (x_A, y_A)$, the angles α and β as well as the equation $y = ax + b$ of the line (BC). For a given value of $\lambda \in (0, w)$, the coordinates of the points E and F are given as follows:

$$(x_E, y_E) = (x_A + \lambda \cos\beta, y_A + \lambda \sin\beta) \quad (1)$$

$$(x_F, y_F) = (x_E + \ell \cos\varphi, y_E + \ell \sin\varphi) \quad (2)$$

The heading angle φ can also be written as $\varphi = \alpha + \beta + \gamma - \pi/2$. Furthermore, when applying the law of sines in the triangle (AEH), we have $\frac{\sin\alpha}{w} = \frac{\sin\gamma}{\lambda}$, therefore the coordinates of F can be rewritten as

$$\begin{cases} x_F = x_A + \lambda\cos\beta + \ell\sin\left(\alpha + \beta + \arcsin\left(\frac{\lambda\sin\alpha}{w}\right)\right) \\ y_F = y_A + \lambda\sin\beta - \ell\cos\left(\alpha + \beta + \arcsin\left(\frac{\lambda\sin\alpha}{w}\right)\right) \end{cases} \quad (3)$$

To solve the problem, we need to find the value of λ such that the point F belongs to the line (BC) which is given by the equation $y = ax + b$. This is equivalent to finding the root of the function

$$f(\lambda) = y_F - ax_F - b \quad \text{with } 0 < \lambda < w. \quad (4)$$

The function f is monotonic w.r.t. $\lambda \in (0, w)$, it is increasing if the slope of the line (BC) is positive (as in Fig. 3) and decreasing otherwise. If $0 \in \{f(\lambda) | 0 < \lambda < w\}$, then the solution is unique and can be computed numerically using Newton-Raphson's method. Otherwise, it means that the size of the vehicle is largely overestimated or underestimated. The configuration pictured in Figure 3 corresponds to a vehicle heading in the right direction w.r.t. the camera, the other direction is solved similarly by symmetry. Once the root of f is found, the coordinates of the rectangle $(EFGH)$ are computed and broadcast to the AD vehicle.

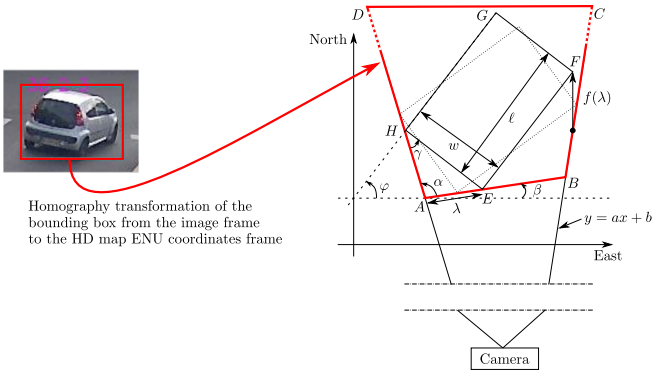


Fig. 3: The BBox is transformed from the image to the HD map frame as the quadrilateral $(ABCD)$. The top-down 2D bounding rectangle $(EFGH)$ is computed by constraining $E \in [AB]$ and $H \in [AD]$ and finding the parameter λ such that $f(\lambda) = 0$ which leads to the optimal configuration represented by the dotted rectangle.

V. Map-aided Road Objects Tracking

To estimate the ENU localization of the perceived road users we use a multi-sensor data fusion approach that considers both on-board sensors and remote infrastructure data. The filter performs prediction and update steps of an Extended Kalman Filter. To handle the problem of out-of-sequence data due to communication and processing latency, we use backward updates [2]. As each piece of information is time-stamped to GPS time in a common reference time frame, there is no clock synchronization issue and the backward updates are done exactly at the right time.

For each road object, we track the position (x, y), the heading angle φ and the longitudinal speed v with a constant yaw rate and linear velocity. Indeed, one of the objectives of the tracking is to estimate the speed of objects. In this work, we have implemented a map-aided tracker because

it is our opinion that this solution behaves as a good compromise between a Cartesian tracker and a curvilinear one and better encompasses the true pose of detected objects at lane-level. This means that, for both the LiDAR and camera observations, we map-match the observations at lane-level. Note that to localize LiDAR detected objects in ENU coordinates, the AD vehicle localization has to be used. The map-matched Cartesian coordinates of an object are therefore given as observations to the tracking system. As the state vector $\hat{X}_{k|k}$ of the tracked objects is expressed in Cartesian coordinates, the resulting estimate is free to move in the 2D space and not constrained only to HD map polylines. This makes it possible to manage behaviors that are not close to the polylines such as lane change maneuvers.

In order to properly feed the tracking system with meaningful observations, we have to select a point on the object and we need a good representation of the uncertainty of the map observations.

Let us consider the LiDAR polygons first. As shown in [3], it is possible to propagate the AD vehicle localization uncertainty directly into the cluster bounding polygons (Figure 1 (c)). The tracker needs a point from the observation to carry on its estimation process. In our case, we choose the barycenter Z_k of the bounding polygon. This measurement is then map-matched to give the observation $Y_k = [x_k,\ y_k]$. To compute the uncertainty of this observation, we first compute a covariance matrix (denoted ${}^{L}R_l$) along the polyline of the HD map. To do so, we manipulate confidence domains. For a given risk α, a $1-\alpha$ confidence interval is classically given by:

$$Pr(\hat{X} - \gamma\sigma \leq X \leq \hat{X} + \gamma\sigma) = 1 - \alpha. \tag{5}$$

In this work, we use $\gamma = \Phi^{-1}(1-\alpha/2) \simeq 1.96$ for $\alpha = 0.05$. Where the function Φ represents a normal distribution. We can then propagate the localization uncertainty of the AD vehicle to point Z_k accordingly to [3]. The polygon obtained by the uncertainty injection into Z_k is then expressed in the HD map polyline frame in the along and cross directions (see L_s and L_n in Fig. 4). For instance, the variance along the polyline is given by ${}^{L}\sigma_s = |Ls|\,/4$.

The same reasoning is done for the transverse component which leads to a covariance matrix ${}^{L}R_l = \text{diag}({}^{L}\sigma_s^2,\ {}^{L}\sigma_n^2)$ in the frame L of the polyline. In order to obtain the covariance matrix R_l in the world reference frame to be used by the tracker, we consider the angle ψ of the polyline and we do a rotation $\text{Rot}(\psi)$ (under the hypothesis that the angle ψ has a negligible uncertainty):

$$R = \text{Rot}(\psi).{}^{L}R_l.\,(\text{Rot}(\psi))^T \tag{6}$$

Regarding the camera information, they are always received from the infrastructure in the form of an oriented rectangle, the orientation being calculated as described in section IV. The point measurement Z_k is chosen to be the center of the box which is projected onto the map polyline in the same way as the LiDAR observation. Regarding the uncertainty of this observation, there is no need to manage the uncertainty of the localization of the camera of the remote intelligent infrastructure. The uncertainty comes from the detection errors of the objects in the image and their back-projection in ENU. We have observed on experimental sequences that this uncertainty is globally constant in the areas of interest. For this reason, the covariance matrix (denoted ${}^{L}R_c$) of the camera observation is chosen to be constant. However, one improvement could be to use an uncertainty that is inversely proportional to the distance between the camera and the detected objects. This is useful to model the fact that objects closer to the camera are detected more precisely w.r.t. far objects.

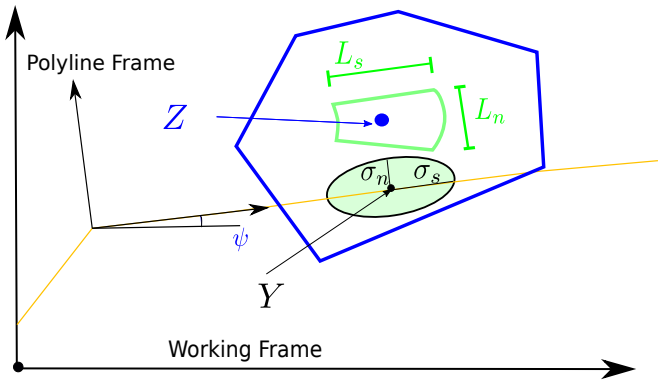


Fig. 4: Map-aided observation Y and its uncertainty. The blue polygon is the bounding LiDAR cluster. The LiDAR measurement Z is the barycenter of this polygon. The map-aided observation Y is the projection of point Z on the polyline. The green polygon represents the direct propagation of the localization uncertainty on point Z and the green ellipsoid represents the resulting uncertainty on Y with the standard deviations σ_s and σ_n in the polyline frame.

The data association of the observations with the tracks is based on the Hungarian algorithm (with Mahalanobis distances) which makes it possible to associate only one track with an observation. This method guarantees that there is always a match between detections. However, to reject unlikely matches, we introduced a thresholding after the assignment based on Mahalanobis distance. As a consequence, each unassociated observation creates a new track so that none of the detections are ignored. This entails that several tracks can correspond to the same object, which makes the tracker's workload a little heavier, but increases the safety of the decision. Tracks that are not associated with observations during 3 update steps are eliminated. This allows the system to be robust against little occlusions and small observation losses. However, for a multisensor system, it is opinion of the authors that a threshold based on a time window instead of one based on update steps will lead to better performances.

VI. Experimental Results

The evaluation of our tracking system has been carried out on real data collected in a roundabout scenario in the city of Compiègne, France. A camera was set up to observe the roundabout and two experimental vehicles were used, one played the role of the AD vehicle while the other was used as a target to evaluate the tracking performance.

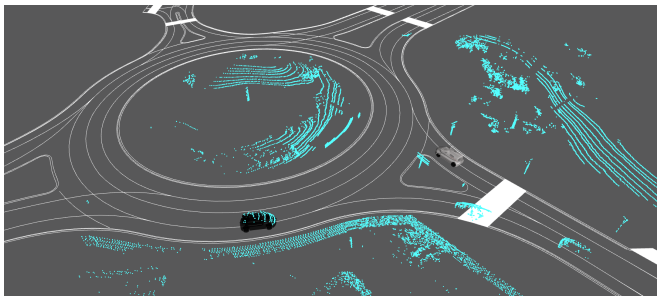


Fig. 5: The AD vehicle used for the data acquisition (white) and the target vehicle of the platform used for ground truth localization (black).

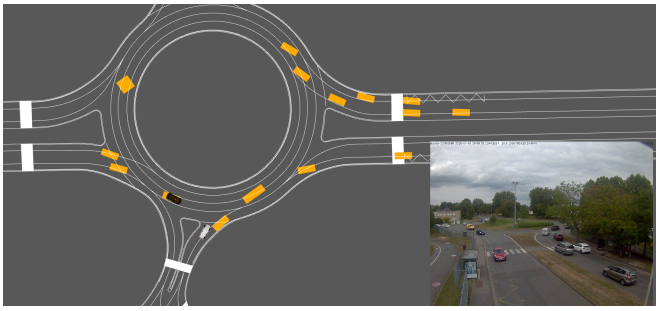


Fig. 6: Projected bounding boxes in the HD map frame and the target experimental vehicle with ground truth in black.

A. Experimental Setup Infrastructure

Regarding the infrastructure system, we have installed a fixed camera on a bridge in the southern direction of the roundabout, in order to observe the traffic on that side of the roundabout. The camera was synchronized with the clocks of the on-board systems of the two cars (through GNSS) via an NTP server and it recorded images of the road users in the roundabout. After the recording session, the camera data have been treated offline to detect vehicles in the recorded images, using YOLOv3, and to provide the road users ENU position estimates in a CPM like format. The camera total rate is fixed to 10 Hz. Figure 6 illustrates one frame of the dataset with the raw-data image and its corresponding bounding boxes. The perception computer was a desktop with 2X Intel Xeon Silver 4110 (8 cores) and Nvidia Quadro P5000 GPU.

B. On-board Experimental Setup and Driving Scenario

For this experiment, the dataset contains recorded accurate positions of both the AD vehicle and the target vehicle used as a target obstacle. For both cars, the accurate positions were obtained with a NovAtel Span CPT IMU/GNSS receiver, with Post Processed Kinematics (PPK) corrections, which gives a centimetric-level of accuracy. Moreover, the AD vehicle was also equipped with a Velodyne VLP32 LiDAR sensor used to provide the observations for the LiDAR-based objects tracking system.. The result of the detection process is provided with a rate of 10 Hz.On our experimental car, we used a standard laptop to process sensors data and to perform data fusion

In the experiment, the AD vehicle navigation went through a roundabout, also stopping it at the roundabout entrance to acquire some data about the incoming traffic flow. At the same time, the target car was navigating into the roundabout among other vehicles, providing extra obstacles for the AD car. The main advantage of this approach is that when the AD car detects and tracks the target, it is possible to compare the resulting estimated states of the track (e.g. position and velocity) w.r.t. the ones obtained with the ground truth localization system. To do so, we select, for every time instant and among all the detected objects, the one that corresponds to the track of the target vehicle and we compare its estimated state with the ground truth. Figure 5 illustrates a scenario of the dataset with the two cars and the corresponding LiDAR point cloud. Figure 7 depicts the navigation scenario for this validation tests with the corresponding trajectories of the AD vehicle.

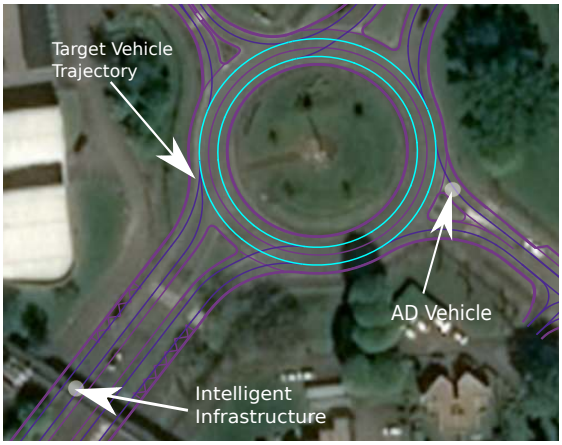


Fig. 7: An overview of the experimental setup used in this experiment. The target vehicle trajectory is represented by the cyan polylines and the positions of the infrastructure and the AD vehicle are indicated by the white dots pointed by the arrows.

C. Results

The main objective of this section is to analyze and discuss the performance of the tracking system regarding the state estimation and the enhancement of the AD vehicle field of view. We evaluate the performance of the tracking system when it uses the on-board and infrastructure data separately and then combined together. Figure 8 illustrates the results of the tracking process for a data sequence of the dataset. This figure illustrates the tracking result for all the vehicles present in the scenario when using only the on-board LiDAR data (a), only the data from the intelligent infrastructure (b) and the combination of them (c). As one can see, the two sensors have complementary fields of view and, when they are combined together, the final result enlarges the autonomous vehicle field of view, providing a more complete, robust and reliable perception of the ongoing driving situation.

To illustrate the accuracy of the tracking system, Figure 9 depicts the tracking result for only the target vehicle. It shows the tracking result using only the LiDAR data, the camera data and both compared with the ground truth of the vehicle. From this figure, one can notice that even if the target car is free to move in the 2D space of the map, the

TABLE I: Root mean square error and the duration of the sequence (in seconds) for the tracking methods obtained using only the LiDAR, the camera and both sensors.

Sensors	Whole (m)	Overlapping (m)
LiDAR	0.70	0.99
Camera	1.88	0.78
Both		0.49

tracking result is more constrained to be close to the HD map polylines. This is because of the map-aided tracking method. However, one can notice some cases where the tracking result deviates from the polylines. This can happen for example in the case of a lane change maneuver. To quantify the gain of the cooperative system w.r.t. the two single-sensor ones, we compute the percentage of time where the target vehicle has been tracked by each of them. This percentage is computed as the ratio between the number of samples that correspond to the target vehicle and the total number of samples of the ground truth in the time interval represented in Figure 9. For the single-sensor systems, the LiDAR-only system tracks the target vehicle for 68% of the time, while the camera-only system tracks it for 37%. Regarding the multi-sensor tracker, it tracks the target vehicle for 92% of the total time, outperforming the two others. Notice that the sum of the single sensor system percentages does not correspond to the multi-sensor system percentage because of the presence of overlapping zones in sensors fields of view.

We have computed, for the target vehicle, the root mean squared error for the three cases described before. To do so, we have projected both the estimated state and the ground truth position of the vehicle on the roundabout polylines and compared the estimation error in terms of along track error in curvilinear coordinates. This is useful because, many autonomous vehicles navigation and motion planning strategies rely on curvilinear coordinates [19]. For this reason, it is interesting to investigate the accuracy in estimating objects in a curvilinear framework. Table I illustrates the computation of the RMSE for the three cases. When using the whole sequence, the LiDAR-only tracking performance is better than the camera-only one. This is mainly due to the different accuracy of the sensors. When focusing the analysis on the parts of the sequence where both the AD vehicle and infrastructure can track the target vehicle, the combination of both sensors performs better than the camera-only and the LiDAR-only ones. In this particular situations, which correspond to the zones where we have a transition from the camera-only setup to the LiDAR-only one, the target vehicle is almost out of the field of view of the sensor and perceived from a frontal orientation w.r.t. the LiDAR, making the estimation of its size less accurate.

VII. Conclusion

In this paper, we have presented a collaborative map-aided tracking system to estimate road users position and speed along the HD map polylines. The tracking process relies on a cooperative perception system composed of an on-board LiDAR sensor and a road side vision system that broadcast the objects it detects. We have presented the main key steps of the processing. In the first experimental part, the performance of this system has been studied in terms of augmentation of the AD vehicle field of view, highlighting the gain of the cooperative perception system over the standalone AD on-board perception. Clearly, a cooperative system improves of the ability to track vehicles in almost all the parts of the roundabout ring that has been considered in this work.

Then; we have presented an evaluation of the tracking performance thanks to the use of a target vehicle accurately localized which allows computing errors on real data. As observed, the three trackers have different accuracy, depending on the sensors capabilities. This means that, in this particular configuration and with the chosen sensors, enhancing the field of view of the AD vehicle by adding an extra source of information implies an accuracy loss in the overall estimation of the position of the tracked objects by the standalone AD vehicle perception. Nonetheless, we have shown that this multi-sensor data fusion leads to benefits in the zones where the field of view of the two sources of perception overlap, providing a more accurate estimation of the state of the perceived objects.

A perspective is to improve the quality of the infrastructure information and to integrate it in the state estimation process of the detected objects in terms of pose and occupancy. Furthermore, an other approach would be to consider a map-aided tracker that tracks the bounds of the objects instead of considering the center of mass of perceived clusters. This would help to provide a more stable and consistent estimation of the state because, as we observed in our experiments, the shape of the detected objects can change a lot from one instant to another, with the consequence that the center of mass moves a lot too.

Acknowledgments

This work has been carried out in the framework of Equipex ROBOTEX (ANR-10- EQPX-44-01) and Labex MS2T (ANR-11-IDEX-0004-02). It was also carried out within SIVALab, a shared laboratory between Renault and Heudiasyc CNRS/UTC, through the TORNADO project.

References

[1] S. Aoki, T. Higuchi, and O. Altintas. Cooperative perception with deep reinforcement learning for connected vehicles. In *Intelligent Vehicles Symposium*, pages 328–334, 2020.

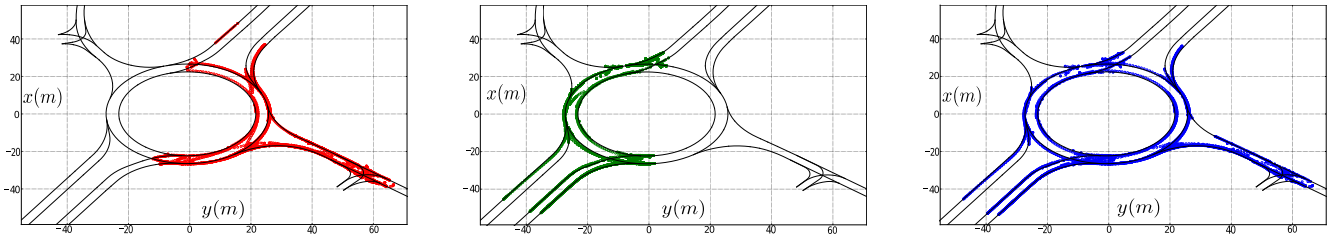
[2] Y. Bar-Shalom. Update with out-of-sequence measurements in tracking: exact solution. *IEEE Transactions on Aerospace and Electronic Systems*, 38(3):769–777, 2002.

[3] E. Bernardi, S. Masi, P. Xu, and P. Bonnifait. High integrity efficient lane-level occupancy estimation of road obstacles through lidar and hd map data fusion. In *Intelligent Vehicles Symposium*, page in proceedings, 06 2020.

[4] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. 2020.

[5] M. Boehme, M. Stang, F. Muetsch, and E. Sax. Talkycars: A distributed software platform for cooperative perception. In *Intelligent Vehicles Symposium*, pages 701–707, 2020.

[6] E. R. Corral-Soto and L. Bingbing. Understanding strengths and weaknesses of complementary sensor modalities in early fusion for object detection. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1785–1792, 2020.



(a) LiDAR only (b) Camera only (c) Camera & LiDAR

Fig. 8: Representation of all the tracked objects in the roundabout with the different systems. Notice that the field of view augments in the multi-sensor case. The black dashed line represents the HD map.

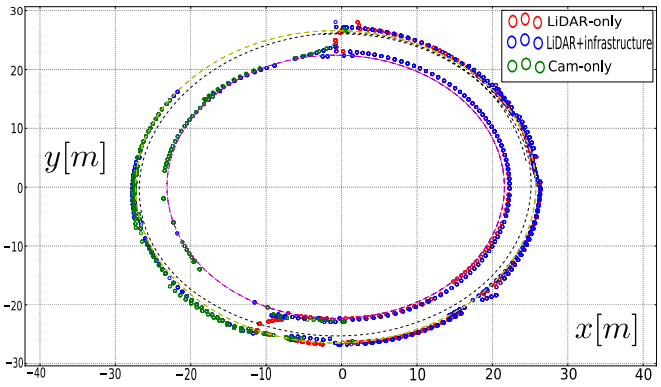


Fig. 9: Comparison of the performance of the three tracking systems (red, blue and green dots) in terms of field of view extension w.r.t. the ground truth data (dashed black line). Notice that, due to the multi-hypothesis criterion of our map-based tracking system, sometimes an instance of the tracked vehicle is present on both lanes.

[7] A. Correa, R. Alms, J. Gozalvez, M. Sepulcre, M. Rondinone, R. Blokpoel, L. Lucken, and G. Thandavarayan. Infrastructure support for cooperative maneuvers in connected and automated driving. In *Intelligent Vehicles Symposium (IV)*, pages 20–25, 2019.

[8] T. Fleck, S. Ochs, M. R. Zofka, and J. M. Zollner. Robust tracking of reference trajectories for autonomous driving in intelligent roadside infrastructure. In *Intelligent Vehicles Symposium*, pages 1337–1342, 2020.

[9] M. Gabb, H. Digel, T. Muller, and R. Henn. Infrastructure-supported perception and track-level fusion using edge computing. In *Intelligent Vehicles Symposium*, pages 1739–1745, 2019.

[10] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012.

[11] T. N. N. Hossein, S. Mita, and H. Long. Multi-sensor data fusion for autonomous vehicle navigation through adaptive particle filter. In *2010 IEEE Intelligent Vehicles Symposium*, pages 752–759, 2010.

[12] European Telecommunications Standards Institute. Intelligent transport systems (its); vehicular communications; basic set of applications; part 2: Specification of cooperative awareness basic service, 2014.

[13] European Telecommunications Standards Institute. Intelligent transport systems (its); vehicular communications; basic set of applications; analysis of the collective perception service (cps), 2019.

[14] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[15] S. Kim, Z. J. Chong, B. Qin, X. Shen, Z. Cheng, W. Liu, and M. H. Ang. Cooperative perception for autonomous vehicle control on the road: Motivation and experimental results. In *International Conference on Intelligent Robots and Systems*, pages 5059–5066, 2013.

[16] S. Kim, W. Liu, M. H. Ang, E. Frazzoli, and D. Rus. The impact of cooperative perception on decision making and planning of autonomous vehicles. *Intelligent Transportation Systems Magazine*, 7(3):39–50, 2015.

[17] S. Kim, B. Qin, Z. J. Chong, X. Shen, W. Liu, M. H. Ang, E. Frazzoli, and D. Rus. Multivehicle cooperative driving using cooperative perception: Design and experimental validation. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):663–680, 2015.

[18] H. Li, F. Nashashibi, and M. Yang. Split covariance intersection filter: Theory and its application to vehicle localization. *IEEE Transactions on Intelligent Transportation Systems*, 14(4):1860–1871, 2013.

[19] S. Masi, P. Xu, and P. Bonnifait. Roundabout crossing with interval occupancy and virtual instances of road users. *IEEE Transactions on Intelligent Transportation Systems (T-ITS)*, pages 1–13, 2020.

[20] D. Meissner, S. Reuter, E. Strigel, and K. Dietmayer. Intersection-based road user tracking using a classifying multiple-model phd filter. *IEEE Intelligent Transportation Systems Magazine*, 6(2):21–33, 2014.

[21] Corey Montella. The kalman filter and related algorithms: A literature review. 05 2011.

[22] University of Applied Sciences TH Aschaffenburg. Vru trajectory dataset, accessed 2020-02-02.

[23] Q.Chen, S.Tang, Q.Yang, and S. Fu. Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. In *International Conference on Distributed Computing Systems (ICDCS)*, pages 514–524, 2019.

[24] A. Rauch, F. Klanner, and K. Dietmayer. Analysis of v2x communication parameters for the development of a fusion architecture for cooperative perception systems. In *Intelligent Vehicles Symposium (IV)*, pages 685–690, 2011.

[25] A. Rauch, F. Klanner, R. Rasshofer, and K. Dietmayer. Car2x-based perception in a high-level fusion architecture for cooperative perception systems. In *Intelligent Vehicles Symposium (IV)*, pages 270–275, 2012.

[26] B. Rebsamen, T. Bandyopadhyay, T. Wongpiromsarn, S. Kim, Z. J. Chong, B. Qin, M. H. Ang, E. Frazzoli, and D. Rus. Utilizing the infrastructure to assist autonomous vehicles in a mobility on demand context. In *TENCON 2012 IEEE Region 10 Conference*, pages 1–5, 2012.

[27] Shaoqing Ren, Kaiming He, Ross B. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015.

[28] D. Zermas, I. Izzat, and N. Papanikolopoulos. Fast segmentation of 3d point clouds: A paradigm on lidar data for autonomous vehicle applications. In *IEEE International Conference on Robotics and Automation*, pages 5067–5073, 2017.

[29] W. Zhan, L. Sun, D. Wang, H. Shi, A. Clausse, M. Naumann, J. Kummerle, H. Konigshof, C. Stiller, A. de La Fortelle, and M. Tomizuka. Interaction dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps, 2019.

[30] Y. P. Zhang, Z. R. Deng, and R. Q. Zhang. An efficient approach of convex hull triangulation based on monotonic chain. In *Information Technology Applications in Industry*, volume 263, pages 1605–1608, 2 2013.