Road Side Perception Systems for Safer Intersections : Field Test

1st Sio-Song Ieng
COSYS
Univ Gustave Eiffel
Marne-la-Vallée, France
Sio-Song.Ieng@univ-eiffel.fr

2nd Mathias Paget
COSYS
Univ Gustave Eiffel
Marne-la-Vallée, France
M.Paget@geo-sat.com

3rd Matossouwé Agninoube Tchalim
COSYS
Univ Gustave Eiffel
Marne-la-Vallée, France
AgninoubeTchalim@gmail.com

4th Jean-Philippe Tarel
COSYS
Univ Gustave Eiffel
Marne-la-Vallée, France
Jean-Philippe.Tarel@univ-eiffel.fr

***Abstract*—Intersections have been known as hazardous points of the road networks for over a century. A diversity of solutions have been developed to reduce the number of accidents at intersections, such as traffic lights, pedestrian crossings, as well as roundabouts. In the past decade, many cities have had to deal with a growing number of alternate means of transport such as bicycles, powered two-wheelers and electric personal transporters, which are both more mobile and more vulnerable. As for the announced automated vehicles, they might be less adaptive that human drivers. These changes again raise the question of how accidents can be prevented at intersections of roads, bus lanes, cycling lanes and walkways. In order to improve the awareness of road users, we propose a cooperative information system based on computer vision to monitor moving obstacles and on communication to warn road users facing imminent danger of collision. We have implemented and tested this system on a roundabout with a high traffic volume, and we report the difficulties that we experienced.**

***Index Terms*—Smart Transportation, Smart Cities, Perception System, Intersection**

I. Introduction

Two decades ago, vehicle communication opened opportunities to develop information systems with various objectives [1], [2]. Communication can be inside the vehicle, between vehicles (V2V) and between vehicles and roadside units (V2I). It is at the basis of the so-called Cooperative, Connected and Automated Mobility (CCAM) strategy in Europe [3].

The first and foremost objective of these information systems is to improve driving safety. In general, the principle of these information systems consists in sharing information about current vehicle status, the environment and disrupting events. Such collected information allow vehicles to have a better knowledge of their environment, even beyond than a driver can have, and thus can be used to take safer and more efficient driving decisions. Some examples are the possibility to better perceive vehicles in bad weather and bad visibility conditions, the possibility to warn about a traffic accident ahead or about an incoming emergency vehicle.

The second objective is for traffic control. Sharing information about vehicle positions and speeds can be used to identify traffic congestion in real time on the entire road network. This better knowledge of the traffic can be used to improve its control or to faster send help in case of accident. A typical local use case is to collect queue sizes at an intersection to increase vehicle flow by optimizing red lights time periods.

Depending on the objectives, on the site, on the sensor technologies, several constrains apply to the information systems such as high frequency, low message latency and low message loss rate. A few camera-based systems have been proposed for intersections, including roundabouts [4], [5] for the purpose of traffic monitoring, without the time constrains of information broadcasting.

In this paper, we focus on information systems about the vehicle traffic at intersections, based on one or several cameras linked with a communication unit able to deliver messages to the connected vehicles. These messages contain rich information about detected moving obstacles such as their type, their footprint, their heading and their tracking index. In Sec. II, related works are discussed. In Sec. III, the camera-based information system is described and its characteristics are discussed. Then Sec. IV describes the experiments performed on a roundabout in the City of Rambouillet, France during the project Tornado-Mobility, where estimation errors and system latency were evaluated using several camera models.

II. Related works

Studies over the past two decades have demonstrated the potentials of cooperative driving systems in improving the safety and efficiency of autonomous mobility [6]–[9]. Cooperative perception systems leverage the information collected from multiple sensors and vehicles to achieve a more accurate and comprehensive understanding of the surrounding environment. Different communication protocols have been proposed for Cooperative Perception, including the Dedicated Short-Range Communications (DSRC) and Cellular Vehicle-to-Everything (C-V2X) technologies [10], [11]. Moreover V2X communication offers an attractive solution for sharing perception by adding connectivity to vehicles using ETSI or SAE standards [12], [13]. For example, [14], [15] recently proposed on-board cooperative perception systems using Vehicle-to-Vehicle (V2V) communication and [16] designed a road-

Thanks to FUI project Tornado-Mobility and ADEME project ENA (Expérimentations de Navettes Autonomes) for funding.

side stereo-vision camera-based system with Infrastructure-to-Vehicle (I2V) communication for a roadside cooperative and collaborative perception system. Besides, in the context of camera-based perception systems, several studies have investigated the use of computer vision and deep learning techniques for object detection [17]–[20] and Multiple Object Tracking (MOT), with the DeepSort algorithm for instance [21], [22]. These methods can be used in the monitoring of road and urban mobility as evidenced by the AI city challenges [23], [24]. For instance, [5] proposed a solution to the 2021 Challenge Track 3: Multi-Camera Vehicle Tracking at City-Scale with multi-cameras focused on crossroad zones where re-identification methods are also needed. However, despite the promising results obtained by previous works, several challenges need to be addressed for the widespread deployment of cooperative perception systems. These challenges include the development of robust, efficient and fast algorithms for object detection and tracking, the optimization of the communication protocols, the integration of different sensor modalities, not to mention the need to evaluate and validate the system as a whole (communication, sensors, perception, control algorithms among others) with realistic use-cases in simulation but also in real road situations as it is proposed in the following.

III. System

The traffic is sometimes heavy at a suburban roundabout and vehicles sometimes enter the roundabout at high speeds. This leads to difficulties for the slower road users such as agricultural vehicles, automated vehicles, trucks, cyclists and pedestrians to safely cross the roundabout, in particular when the visibility is reduced by bad weather conditions or by occultation. We thus designed a roadside information system able to track road users on every lanes entering the roundabout and able to broadcast information to all users with a receiver. The goal of this information system is to broadcast the following features about the moving obstacles: accurate location on the road, coarse footprint, speed vector and object class, at the time of observation.

A. Camera-based roadside information system

As illustrated in Fig. 1, the roadside information system consists in a RGB camera with HD resolution (1280×960) in a waterproof case attached at 5 meters high on a pole, a computer connected to the camera for video acquisition and processing, a roadside unit (RSU) for sending messages about the moving obstacles observed by the camera to the road users with on-board units (OBU), i.e the receivers. The V2X systems, RSU and OBU use the DSRC protocol, including GPS for synchronization. The sooner the information about obstacles is broadcasted, the lower the location uncertainty due to motion. The time of observation is thus an important information that must be broadcasted along with other mobile features. The latency between the time of observation and the time information is delivered to users is one of the critical parameters of the system on which its usefulness depends. Another critical parameter is the period of time at which the information about moving obstacles is refreshed. This period of time can be smaller than the previous latency when the systems uses parallelism between the different tasks.

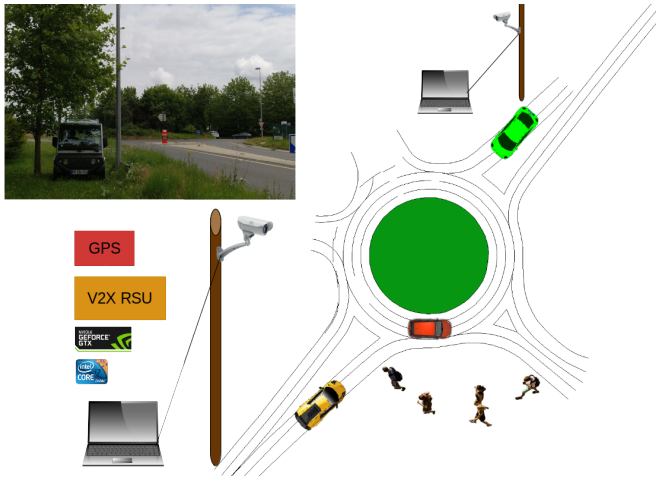

Fig. 1. The information system with a camera attached to a public lighting pole close to a roundabout in the City of Rambouillet, France. The camera is connected to a computer for image processing. The computer is also connected for communication to a roadside unit (inside the service vehicle in the topleft picture).

B. Video acquisition

A camera is installed on each road entering the roundabout. The camera should be installed between 50 and 100 meters from the roundabout along the roadside, at a minimum height of 5 meters to avoid occultation by trucks or busses. In addition, the camera should be tilted downwards to avoid direct sunlight which has a negative effect on image quality. Moreover, the camera should be looking at the roundabout allowing to better see incoming vehicles while they are still far from the roundabout. The camera lens should be of fixed focal length, and its field of view should allow full view of the road going through the roundabout. The camera should have automatic iris and gain tuning to deal with various lighting conditions. The camera can be connected to the computer by an IP cable up to 100 meters long (possibly with a PoE injector) or by a USB cable up to 5 meters long. Each image must be time-stamped by the camera at the time of acquisition, unless the transmission is fast enough (within a few milliseconds) to allow a time-stamp by the connected computer. The time between camera and computer must be synchronized, using NTP or PTP for more accuracy. For communication purposes, the computer time is synchronized with the GPS time.

C. Detection and recognition in images

After video acquisition, images captured from the video stream are sent on the fly to the computer to perform the first step of the processing. Moving objects have to be detected in each image. The recognition is used to approximately estimate the width and length of the footprint of each moving object. So rather than relying on moving object detection in image

Fig. 2. Detection results obtained using refined YOLOv3 for different cameras at different positions.

sequences, we tested detection and recognition using the well-known state-of-the-art convolutional neural network (CNN) named YOLO (You Only Look Once) which has the advantage of speed [17]. YOLO predicts a probabilty map of the presence of the objects of interest, several possible bounding boxes and possible classes. The learning consists in the optimization of the sum of three loss functions: the class recognition loss, the boundary box loss and the classification loss.

We experimented with two different versions: YOLOv3 [18] from darknet and YOLOv5. The later is easier to interface within the processing chain. We reduced the number of object classes from the original 80 to six: person, car, bicycle, motorbike, truck/tractor, and motor bus. With images such as in Fig. 2, we performed a refinement of the learning on our own labeled dataset to improve detection and recognition performances on different camera locations. Our labeled dataset is built from images extracted from a few videos captured with cameras on different roundabouts. The images were selected carefully so as to have various moving objects in various classes at different positions on the road and in the roundabout. These images were annotated using YOLO_mark tool.

D. Detection tracking along images

The detection being performed on single images, speed estimation required a fast tracking of detected objects between frames. We have used Deep SORT [21] because it is fast and reliable after fine tuning of the parameters to the camera view. Deep SORT consists in two steps: a prediction step using a linear Kalman filter, followed by an association step between new detections and existing tracks using the Hungarian algorithm. The association step takes into account the geometric distance, the speed and the visual similarity of the YOLO features. Each newly detected object is assigned with a new index and keeps that index until it leaves the camera field of view. However, objects sometimes lose their index when they are masked for too long.

E. Camera calibration

To be able to backproject detected object onto the road surface, the camera transformation between the road and the image must be estimated which calls for intrinsic and extrinsic camera calibration. The intrinsic camera calibration consists in estimating the parameters of the camera model such as pixel size, image center and distortions. We used the pinhole camera model with three terms for radial distortions. The method for the intrinsic calibration is based on [25]. This intrinsic calibration needs only be performed once before the camera is fixed in its waterproof case.

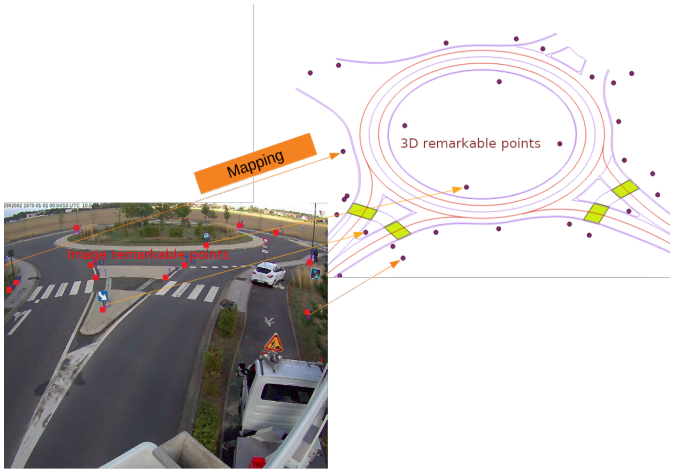

Fig. 3. The image coordinates of remarkable points are associated with their 3D coordinates to estimate the camera position and orientation.

The extrinsic camera calibration estimates the position and the orientation of the camera with respect to a reference coordinate system on the ground. Thus the extrinsic camera calibration must be performed in situ each time the camera is moved.

To be able to provide information about the whereabouts of moving obstacles, a shared reference coordinate system should be used for positioning. We used the Earth Centred, Earth Fixed (ECEF) coordinate system which is used in GPS, and the local tangent plane coordinates East-North-Up reference (ENU). The extrinsic camera calibration consists in selecting remarkable points in the imaged scene such as corners of roadsigns, streetlamps and corners of lane markings. As shown in Fig. 3, the image coordinates of these remarkable points are associated to their 3D coordinates. Knowing the intrinsic camera parameters, the position and orientation of the camera can be estimated with in the ENU coordinate system by minimizing the error between the 2D image coordinates and the 3D coordinates projected into the image. The Nelder-Mead method (a.k.a the downhill simplex method) [26] is used for optimization.

F. Detected object localization

From the intrinsic and extrinsic camera calibration, it is possible to backproject, for each detection, the corners of the detected bounding box on the ground, which is assumed

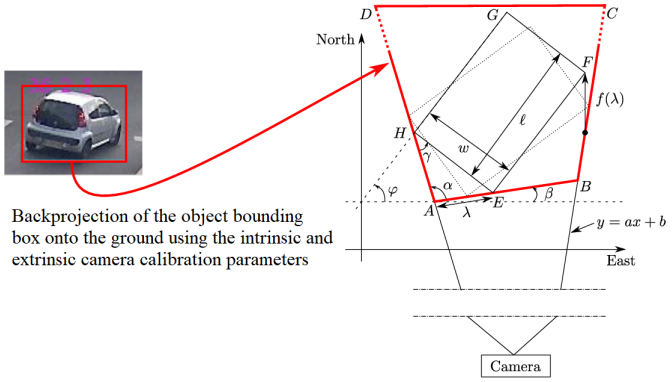

Fig. 4. From the length and width of the rectangular footprint, the footprint angle can be estimated by keeping E into $[A, B]$, H into $[A, D]$ and enforcing F into $[B, C]$. Figure from [27].

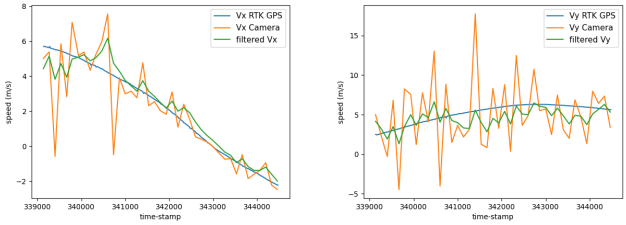

Fig. 5. The two estimated speed components along the South-North and East-West axes of the ENU reference system are filtered and compared with the RTK GPS measurements.

locally flat. The size of the bounding box being related to the height of the detected object, we decided to use the backprojection of the mid point of the bottom segment of the bounding box as the object location on the ground. This location changes with the yaw angle of the detected object on the road, but we found this choice to be relatively robust in our experiments.

G. Detected object geometrical model

Besides the position of each detected objects, there are far more information we would like to extract, like the size, height, orientation and so on. So a geometrical model has to be set up for each object, as it was proposed in [27]. Indeed, a detector such as YOLO provides bounding boxes from which only approximations of the geometrical information about the object can be obtained, as opposed to information provided by an actual sensor such as a LIDAR. The bounding box backprojected on the ground is a too large because the object is not flat. A better geometrical representation can be predicted when the class of the detected object is recognized. Indeed, the six classes selected in III-C can be represented by rectangles with length and width that are directly related to the typical length and width of the corresponding vehicles. Pedestrians will be represented by a square. For instance, for cars in France, the geometrical model is a rectangle with average width 1.8 m and with average length 4.3 m. For motor buses in France, the average width is 3 m and the average length is 9 m. For vehicles, when an accurate map of the road is available, the rectangular footprint may also be assumed oriented along the road axis. With all these assumptions, we improved the geometrical representation, or footprint, of the detected objects, which helped us to improve the estimation of the space they occupy on the roadway and their headings.

When an accurate map of the road is not available, and for not too distant objects, we found a way to estimate the vehicle footprint orientation (heading) assuming that its size is known. When the vehicle is seen from the front/back or from the side, the heading is easy to derive knowing the camera orientation. In the case where the vehicle is seen at an intermediate angle as depicted in Fig. 4, an approximate footprint orientation can be estimated. The backprojection of the bounding box within the image is a trapezoidal quadrilateral $(ABCD)$. We assume that the rectangular footprint $(EFGH)$ of the vehicle is of length l and width w. For a vehicle going right, the lower corner E of the footprint should be in the segment $[A, B]$, the left corner H should be in the segment $[A, D]$ and the right corner F should be in the segment $[B, C]$. To achieve this, the idea is to slide the corner E of the vehicle along $[A, B]$ by varying the parameter λ until $f(\lambda) = 0$. These three constraints lead to a unique possible orientation for correct values of l and w, as detailed in [27].

H. Detected object speed

A Kalman filter can be used to smooth the position of tracked objects and to estimate their speed. The better the dynamic model, the more accurate the speed estimation. We consider a dynamic model with constant speed as in (1):

$$\begin{cases} x_{t+dt} &= x_t + \dot{x}_t dt \\ y_{t+dt} &= y_t + \dot{y}_t dt \\ \dot{x}_{t+dt} &= \dot{x}_t \\ \dot{y}_{t+dt} &= \dot{y}_t \end{cases} \quad (1)$$

where (x_t, y_t) is the ground position at time t and $(\dot{x}_t, \dot{y}_t)$ is the ground speed vector. Compared to the difference between two consecutive values, the Kalman filter provides smoother results with less noise. In Fig. 5, the raw speed components are compared with the ones obtained by Kalman filtering and the reference measurement obtained with an RTK GPS onboard the observed vehicle.

I. Message broadcast

Once the moving obstacles are detected and the required features are estimated, a message is broadcasted about the traffic information using the roadside unit (RSU). In order to manage different kinds of information to be shared at a high frequency (up to 10 Hz), the ETSI standard provides two kinds of message format: the Cooperative Awareness Message (CAM) and the Collective Perception Message (CPM) [12]. The CAM is broadcasted by a vehicle for sharing its own information [28] and the CPM is broadcasted from the surrounding environment. In the experimented system, we have implemented a roadside unit from LACROIX City which

broadcasts information about up to 255 perceived objects using CPM, at a maximum frequency of 10 Hz.

J. Latency and refreshing period

TABLE I
AVERAGE LATENCY (IN SECOND) BETWEEN ACQUISITION AND MESSAGE CREATION FOR DIFFERENT YOLO VERSIONS AND CAMERA MODELS.

Camera	YOLOv3	YOLOv5
Basler BIP2 1300C (10 Hz)	0.543 s	0.543 s
Basler BIP2 1300C (30 Hz)	0.206 s	0.207 s
Allied Vision Prosilica GT 1930 C	Not tested	0.111 s
Allied Vision Mako G192C	Not tested	0.095 s

Due to the maximum frequency of the roadside unit, the refreshing period can not be lower than 0.1 s. Ideally, the latency between acquisition and message creation should be lower than 0.1 s. As shown in Tab. I, this is not always the case. By using a Dell T7920 workstation with Nvidia GPU Quadro RTX 5000, image processing time is around 0.008 s. The latency is thus mainly due to the camera acquisition and image transmission and thus depends on the implemented camera and on its tuning. We have experimented with three camera models with different tuning and with two YOLO versions. The ideal latency of 0.1 s is not achieved, even if it is achieved in average with the Allied Vision Mako G192C. Indeed, the latency is subject to variations, usually with a standard deviation of 10 % of the average. The smallest latency is obtained with the Allied Vision Mako G192C and the Allied Vision Prosilica GT 1920 C cameras. With both cameras, maximum frequency can be achieved parallelizing the camera acquisition and the video processing on the computer. With the Basler BIP2 1300C set at 30 Hz, we only managed a refreshing frequency of 8 Hz.

IV. EXPERIMENTS

Onsite experiments of the described information system were performed within the Tornado-Mobility FUI project. The demonstration site is close to the Bel-air shopping center near Rambouillet and has four roundabouts. The road infrastructure in this area is too basic and the environment is too complex to allow Connected and Autonomous Vehicles (CAV) to move efficiently. One roundabout is large with a central island diameter of 26.7 m with poor visibility due to dense vegetation, as shown in Fig. 1. To cover this large roundabout, we tested with two position of the information systems: the first position is in the avenue entering from the North, 54 m from the center of the roundabout, and the second is in the avenue entering from the South at 55 m. After intrinsic and extrinsic camera calibration, the average backprojection error was 2.5 pixels for the North camera and 3.2 pixels for the South camera. These values are relatively high due to the small number of used remarkable points. Nevertheless, it leads to an accuracy of a few decimeter on the ground for closer objects, which was deemed enough for our application. The perception system validations consisted in detecting and tracking a Target Instrumented Vehicle (TIV) among other vehicles in a real road traffic situation. This target vehicle, a Renault ZOE car, developed by Université Technologique de Compiègne (UTC), is equipped with a NovAtel Span CPT IMU/GNSS receiver, with Post Processed Kinematics (PPK) corrections, for centimetric accuracy [27]. The GNSS receiver was installed near the rear axle. The vehicle and the cameras were synchronized with the same reference: the GPS time. This instrumented vehicle provided accurate pose measurements in the same ENU reference system and groundtruth for the system evaluation. The grountruth consisted in: the timestamp t, the TIV position (x, y) in the ENU reference, the TIV speed (v_x, v_y), the TIV heading φ. The perception system was evaluated in terms of position and orientation accuracy, with Multiple Objet Tracking (MOT) tests and real-time tests.

A. Position and orientation accuracy

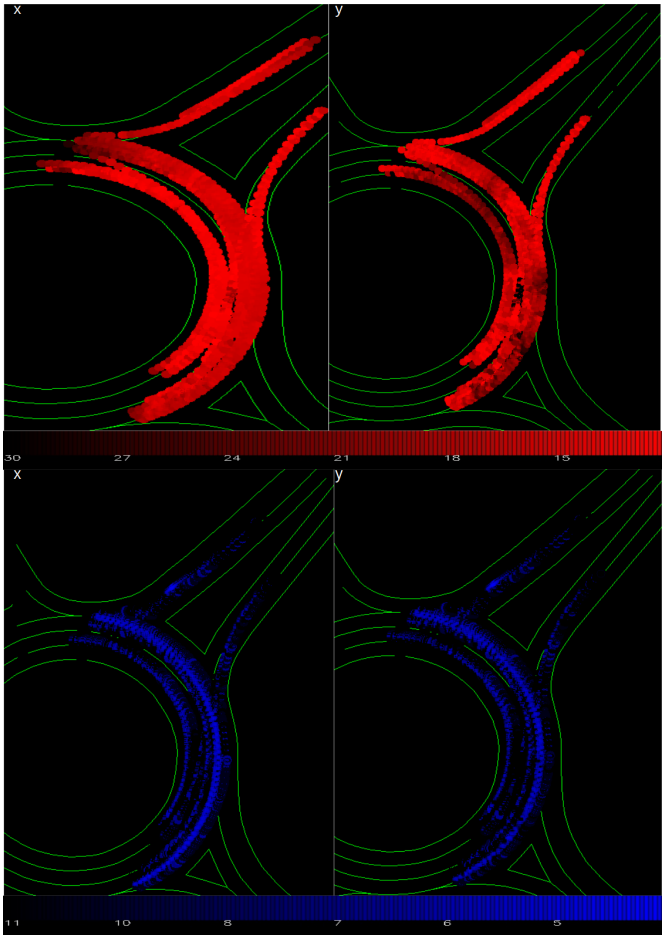

Fig. 6. The average South-North and East-West errors between the observed and RTX GPS locations are shown in the first line where the scale is in centimeter (cm). In the second line, the two components South-North and East-West of the standard deviation of the observed location are shown, where the scale is in decimeter (dm).

The tracks of the instrumented vehicle, detected from the camera, were selected manually. There are several tracks due to the different round trips. By comparing RTK GPS measurements with the results from processing the images of the camera, the location error is estimated. For these tests,

only the position of the system on the avenue entering from the North was used. The instrumented vehicle made several round trips in the roundabout and in the avenue which enters the roundabout from the North.

The RTK GPS receiver is installed inside the instrumented vehicle and thus it is not visible from the camera. It is necessary to compensate for the positional difference between the receiver location and the point used for vehicle location. Along the trajectory, we computed local average errors and standard deviation for every position (x, y), by averaging in a centered squared windows of size $2d$ and thus with corner coordinates $(x-d, x+d)$ and $(y-d, y+d)$. Fig. 6 shows the average South-North and East-West errors between observed and RTK GPS vehicle locations, in the first line. One may notice that the average error is higher for the South-North component, which is along the camera axis. The East-West component of the location error is around 15 cm but can reach up to 1 m in the worst cases due to the occultation by tree foliage. For the South-North component, the location error is closer to 30 cm, but can be larger in case of occultation of the instrumented car by another vehicle. The two components South-North and East-West of the standard deviation of the observed location are shown on the second line of Fig. 6. Again, the error along the axis of the camera is higher than along the perpendicular direction due to the perspective effect.

Fig. 7 shows the result of the comparison between yaw angle estimated from the camera and estimated from the RTK GPS successive positions of the instrumented vehicle. The error on the yaw angle is around 0.4 rad in most cases except in a few localized spots near the trees or near the roundabout exit where other vehicles are often slowing down.

B. Multiple Object Tracking test

The Multiple Object Tracking (MOT), described in the section III-D, was tested in real road traffic conditions with the TIV as the only reference vehicle. To measure the performance of the tracker, we used the two MOT metrics proposed in [29]: the Mutliple Object Tracking Precision (MOTP) $MOTP = \frac{\sum_{i,t} d_t^i}{\sum_t C_t}$ and the Mutliple Object Tracking Accuracy (MOTA) $MOTA = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t}$, where d_t^i is the distance between the ground truth and the estimated position at time t for each track i, C_t is the number of matches found for times t, fp_t is the false positive at time t, m_t is the missed target at time t, mme_t is the mismatch at time t ang g_t is the number of objects at time t. As the TIV position is known, the MOTP is only computed for the TIV. The MOTA can be computed for every visible vehicle in the images. To measure the MOTA of our tracker we used 200 images of our database (images and groundtruth). The results provided in Table II are quite honorable.

C. Real-time tests

During the Tornado-Mobility FUI project, a real-time test was organized at the beginning of 2021 in the area of the Bel-air shopping center near Rambouillet. The TIV was the prototype of the Université Technologique de Compiègne (UTC) which is able to receive the information broadcast by the roadside camera-based system and able to use the received information in the driving control algorithm [27]. During this test, it was demonstrated that the camera-based system was able to broadcast useful information to help the CAV to enter smoothly into the roundabout. We observed that the processing time was not constant but depended on traffic density. Indeed, the processing time from detection to CPM message was 0.077 s in average (i.e 13 Hz) when only a few vehicles were seen by the camera, but when the traffic was heavy, the processing time increased up to 0.09 s (i.e 11 Hz). The processing time is 10 times higher compared to the one obtained with the Dell T7920 workstation due to the use of Dell G5 laptop with a Nvidia GPU GeForce RTX 2060. Even with the use of a more powerfull computer and GPU, the objective of a global processing time most of the time lower than 0.1 s implies that

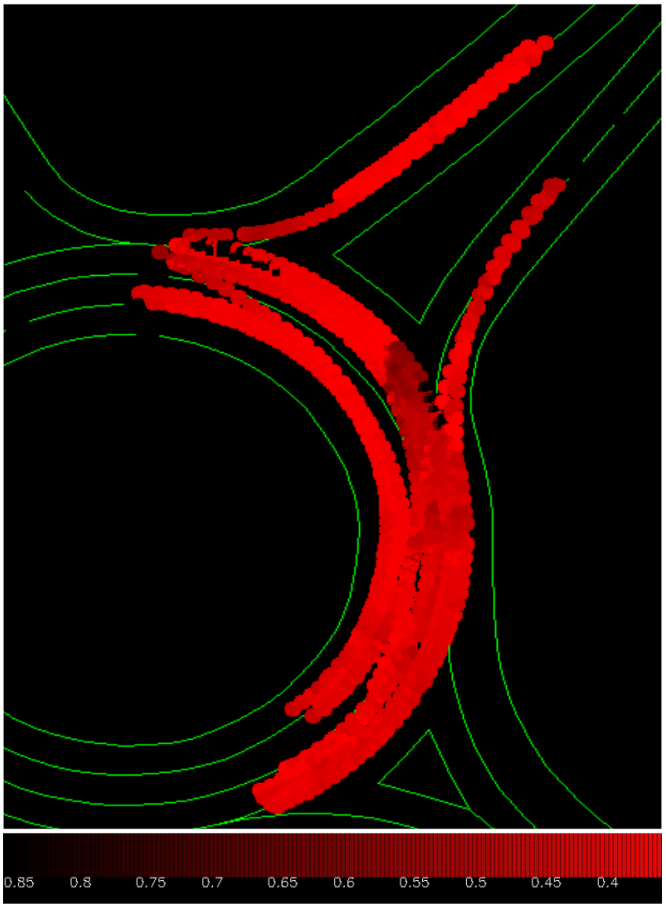

Fig. 7. Errors between the yaw angle estimated from RTK GPS positions and from the camera, in radian.

TABLE II
RESULTS OF THE MOT TEST.

MOTP	m rate	fp rate	mme rate	MOTA
375 mm	1.5%	0.2%	1.3%	90.1%

the image acquisition and transmission time should be much lower than 0.1 s. From Tab. I, it appears that even the use of IP cameras does not allow to achieve such reduced values of acquisition and transmission delay.

This kind of real-time tests were also performed in a very large roundabout close to the UTC, in the city of Compiègne, where traffic is heavy, with similar results.

V. Conclusion

We have described a camera-based system able to broadcast traffic information to connected vehicles that may help crossing road intersection or entering roundabout when visibility is reduced or when traffic is heavy. The proposed system consists in a video camera, a computer for the video processing and a roadside unit able to broadcast messages about the detected traffic. The different steps of the image and video processing are described and discussed, in particular the difficulty to achieve low latency and low refreshing time. The proposed system was evaluated in terms of accuracy and was tested on suburban sites with success to help a CAV enter two large roundabouts in the cities of Rambouillet and Compiègne in France.

References

[1] M. L. Sichitiu and M. Kihl, "Inter-vehicle communication systems: a survey," *IEEE Communications Surveys & Tutorials*, vol. 10, no. 2, pp. 88–105, 2008.

[2] J. Zhang, E. Vinkhuyzen, and M. Cefkin, "Evaluation of an autonomous vehicle external communication system concept: a survey study," in *Proceedings of the AHFE International Conference on Human Factors in Transportation*. Springer, 2018, pp. 650–661.

[3] M. A. Raposo, M. Grosso, A. Mourtzouchou, J. Krause, A. Duboz, and B. Ciuffo, "Economic implications of a connected and automated mobility in europe," *Research in transportation economics*, vol. 92, p. 101072, 2022.

[4] T. Fleck, S. Ochs, M. R. Zofka, and J. M. Zollner, "Robust tracking of reference trajectories for autonomous driving in intelligent roadside infrastructure," in *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2020, pp. 1337–1342.

[5] C. Liu, Y. Zhang, H. Luo, J. Tang, W. Chen, X. Xu, F. Wang, H. Li, and Y.-D. Shen, "City-scale multi-camera vehicle tracking guided by crossroad zones," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4129–4137.

[6] J. Yin, T. A. ElBatt, G. Yeung, B. Ryu, S. Habermas, H. Krishnan, and T. Talty, "Performance evaluation of safety applications over dsrc vehicular ad hoc networks." in *Vehicular Ad Hoc Networks*, K. P. Laberteaux, R. Sengupta, C.-N. Chuah, and D. Jiang, Eds. ACM, 2004, pp. 1–9.

[7] N. M. Rabadi and S. M. Mahmud, "Performance evaluation of ieee 802.11a mac protocol for vehicle intersection collision avoidance system," in *2007 4th IEEE Consumer Communications and Networking Conference*, 2007, pp. 54–58.

[8] F. Eckermann, M. Kahlert, and C. Wietfeld, "Performance analysis of c-v2x mode 4 communication introducing an open-source c-v2x simulator," *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*, pp. 1–5, 2019.

[9] Q. Wu, S. Zhou, C. Pan, G. Tan, Z. Zhang, and J. Zhan, "Performance analysis of cooperative intersection collision avoidance with c-v2x communications," in *2020 IEEE 20th International Conference on Communication Technology (ICCT)*, 2020, pp. 757–762.

[10] J. Choi, V. Marojevic, C. B. Dietrich, J. H. Reed, and S. Ahn, "Survey of spectrum regulation for intelligent transportation systems," *IEEE Access*, vol. 8, pp. 140 145–140 160, 2020.

[11] T. Petrov, P. Pocta, and T. Kovacikova, "Benchmarking 4g and 5g-based cellular-v2x for vehicle-to-infrastructure communication and urban scenarios in cooperative intelligent transportation systems," *Applied Sciences*, vol. 12, no. 19, 2022.

[12] E. T. S. Institute, "Intelligent transport systems (its); vehicular communications; basic set of applications; analysis of the collective perception service (cps)," 2019.

[13] ——, "Etsi tr 103 562 - intelligent transport systems (its); vehicular communications; basic set of applications; analysis of the collective perception service (cps),," 2020.

[14] P. Zhou, P. Kortoi, Y.-P. Yau, B. Finley, X. Wang, T. Braud, L.-H. Lee, S. Tarkoma, J. Kangasharju, and P. Hui, "Aicp: Augmented informative cooperative perception," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 11, pp. 22 505–22 518, 2022.

[15] S. Ingrachen, N. Achir, P. Muhlethaler, T. Djamah, and A. Berqia, "A collaborative environment perception approach for vehicular ad hoc networks," in *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*, 2018, pp. 1–5.

[16] M. Tsukada, M. Kitazawa, T. Oi, H. Ochiai, and H. Esaki, "Cooperative awareness using roadside unit networks in mixed traffic," *2019 IEEE Vehicular Networking Conference (VNC)*, pp. 1–8, 2019.

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[18] A. Farhadi and J. Redmon, "Yolov3: An incremental improvement," in *Computer vision and pattern recognition*, vol. 1804. Springer Berlin/Heidelberg, Germany, 2018, pp. 1–6.

[19] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector." in *ECCV (1)*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9905. Springer, 2016, pp. 21–37.

[20] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[21] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 3645–3649.

[22] Z. Wang, L. Zheng, Y. Liu, and S. Wang, "Towards real-time multi-object tracking," *The European Conference on Computer Vision (ECCV)*, 2020.

[23] M. Naphade, S. Wang, D. C. Anastasiu, Z. Tang, M.-C. Chang, X. Yang, Y. Yao, L. Zheng, P. Chakraborty, C. E. Lopez, A. Sharma, Q. Feng, V. Ablavsky, and S. Sclaroff, "The 5th ai city challenge," 2021.

[24] M. Naphade, S. Wang, D. C. Anastasiu, Z. Tang, M.-C. Chang, Y. Yao, L. Zheng, M. S. Rahman, A. Venkatachalapathy, A. Sharma, Q. Feng, V. Ablavsky, S. Sclaroff, P. Chakraborty, A. Li, S. Li, and R. Chellappa, "The 6th ai city challenge," 2022.

[25] Z. Zhang, "Flexible camera calibration by viewing a plane from unknown orientations," in *Proceedings of the seventh ieee international conference on computer vision*, vol. 1. IEEE, 1999, pp. 666–673.

[26] J. A. Nelder and R. Mead, "A simplex method for function minimization," *Computer Journal*, vol. 7, pp. 308–313, 1965.

[27] S. Masi, P. Xu, P. Bonnifait, and S.-S. Ieng, "Augmented perception with cooperative roadside vision systems for autonomous driving in complex scenarios," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021, pp. 1140–1146.

[28] D. Wesemeyer and J. Trumpold, "Controlling a real-world intersection with connected vehicle information provided by cams (cooperative awareness messages)," in *SUMO User Conference 2019*, ser. EPiC Series in Computing, vol. 62. EasyChair, 2019, pp. 206–212.

[29] K. Bernardin, E. Elbs, and R. Stiefelhagen, "Multiple object tracking performance metrics and evaluation in a smart room environment," in *Proceedings of IEEE International Workshop on Visual Surveillance*, 2006.