Multi-modal object detection and localization for high integrity driving assistance

Sergio Alberto Rodriguez Florez, Vincent Fremont, Philippe Bonnifait, Véronique Cherfaoui

▶ **To cite this version:**

Sergio Alberto Rodriguez Florez, Vincent Fremont, Philippe Bonnifait, Véronique Cherfaoui. Multi-modal object detection and localization for high integrity driving assistance. Machine Vision and Applications, 2014, 25 (3), pp.583-598. 10.1007/s00138-011-0386-0 . hal-00777387

HAL Id: hal-00777387
https://hal.science/hal-00777387

Submitted on 18 Jan 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Noname manuscript No.
(will be inserted by the editor)

Multi-Modal Object Detection and Localization for High Integrity Driving Assistance

Sergio A. Rodríguez F. · Vincent Frémont · Philippe Bonnifait · Véronique Cherfaoui

Received: date / Accepted: date

Abstract Much work is currently devoted to increasing the reliability, completeness and precision of the data used by driving assistance systems, particularly in urban environments. Urban environments represent a particular challenge for the task of perception, since they are complex, dynamic and completely variable. This article examines a multi-modal perception approach for enhancing vehicle localization and the tracking of dynamic objects in a world-centric map. 3D ego- localization is achieved by merging stereo vision perception data and proprioceptive information from vehicle sensors. Mobile objects are detected using a multi-layer lidar that is simultaneously used to identify a zone of interest in order to reduce the complexity of the perception process. Object localization and tracking is then performed in a fixed frame which simplifies analysis and understanding of the scene. Finally, tracked objects are confirmed by vision using 3D dense reconstruction in focused regions of interest. Only confirmed objects can generate an alarm or an action on the vehicle. This is crucial to reduce false alarms that affect the trust that the driver places in the driving assistance system. Synchronization issues between the sensing modalities are solved using predictive filtering. Real experimental results are reported so that the performance of the multi-modal system may be evaluated.

Keywords Multi-modal perception · Visual odometry · Object tracking · Dynamic map · Intelligent vehicles

S.A. Rodríguez F. · V. Frémont · P. Bonnifait · V. Cherfaoui
Université de Technologie de Compiègne (UTC)
CNRS Heudiasyc UMR 6599, France
Centre de Recheches de Royallieu
BP 20529, 60205 Compiègne cedex, France
E-mail: sergio.rodriguez@hds.utc.fr

V. Frémont
E-mail: vincent.fremont@hds.utc.fr

P. Bonnifait
E-mail: philippe.bonnifait@utc.fr

V.Cherfaoui
E-mail: veronique.cherfaoui@hds.utc.fr

1 Introduction

Advanced Driver Assistance Systems (ADAS, *an acronym list is given at the end of the paper*) can improve road safety through their obstacle detection and avoidance functions. For these functions, the location and the speed of nearby mobile objects are key pieces of information.

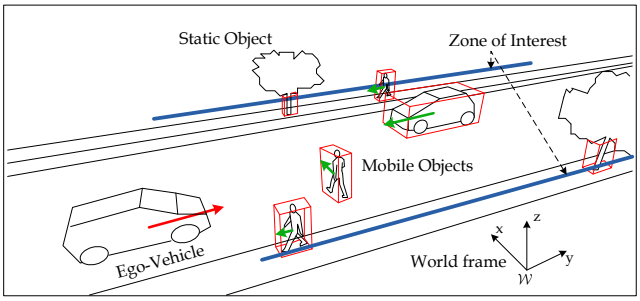


Fig. 1 A Dynamic Map is composed of a list of the states of tracked objects located within a zone of interest together with the changing vehicle dynamics in the 3D scene.

In the literature, a number of different approaches have been applied to problems of object localization and tracking. Robotics approaches have been used to distinguish the static part of the environment [9] and

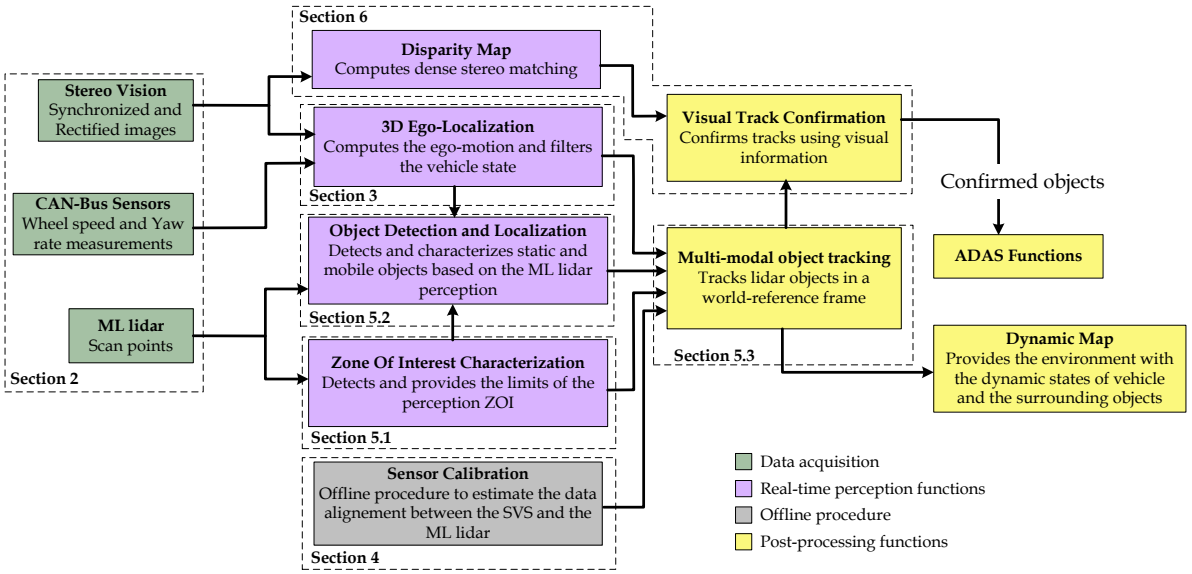


Fig. 2 Multi-Modal System Design

to simultaneously detect moving objects [32]. Leibe et al. have presented in [18], a stereo vision strategy for obtaining a 3D dynamic map using a Structure-from-Motion technique and image object detectors. Using lidar information only, it is possible to estimate the ego-motion and to detect mobile objects through a dense 3D grid-map approach [21]. In contrast, for [4] and [17], real-time sensor-referenced approaches (i.e. ego- localization is not considered) are presented using multi-sensor systems showing the complementarity of lidar and vision systems in automotive applications.

A world-centric approach presents interesting properties once the ego-localization is estimated accurately (up to 1 cm per Km/h). The tracking performance can be improved, since the dynamics of the mobile objects are better modeled. This sort of approach simplifies the understanding of the scene and the ADAS implementation, and is also well adapted to cooperative applications [26] (e.g. vehicle-to-vehicle communication). The present study addresses the problem of ego-localization and object tracking in urban environments.

Ego-localization can be achieved using proprioceptive and exteroceptive sensors [5]. GPS is an affordable system that provides 3D positioning. Unfortunately, GPS performance can decrease significantly in urban environments because of multi-paths and satellites outages. Dead-reckoning is a complementary solution which can be used when GPS information is unreliable. Stereo Vision Systems (SVS), often used for 3D reconstruction, detection and recognition tasks, are also useful for dead-reckoning (also called 3D ego-motion estimation) [6].

Object tracking for ADAS is still an active research domain. Urban environments are characterized by complex conditions: moving and static objects, mobile perception, varied infrastructures. Object representation [23,22], association methods [29], motion model and tracking strategies [19] are key points requiring particular attention.

In this work, we study a tightly coupled multi-modal system able to provide a 3D local perception of the vehicle's surrounding environment in a world-centric frame, as depicted in Fig. 1. The modeled environment is composed of static and moving objects and a zone of interest localized in front of the vehicle. Our contribution lies in the estimation of the dynamics (location and speed) of the surrounding objects in order to build a *dynamic* map, and in ensuring the map's integrity by using different, independent sensing principles. A dynamic map is composed of a list of the states of tracked objects together with the changing vehicle dynamics in the 3D scene. A particular difficulty faced in this study is the use of asynchronous modalities which are sampled independently. To address this issue, we propose a multi-sampled strategy.

The multi-sensor system we have designed makes use of essential information provided by an SVS coupled with a Multi-Layer lidar (denoted ML lidar) and proprioceptive sensors (i.e. wheel speed sensors and a yaw rate gyro).

The overall system design and architecture of the proposed system are illustrated in Fig. 2. First, nearby objects are detected based on ML lidar data. The SVS and the proprioceptive vehicle sensors are used simultaneously to estimate the 3D ego-localization of the vehicle. Subsequently, the detected objects are localized and tracked w.r.t. a world reference frame. Finally, tracked

objects are transferred to a vehicle-centered frame in order to be confirmed by the SVS, by taking into account the different sampling times of the two sensing modalities. Confirmed tracks that are declared as verified then become the input into an ADAS for the detection of possible collisions.

In this article four main topics are addressed: 3D ego-localization, exteroceptive sensor calibration (i.e. data alignment), object localization and tracking, and visual track confirmation. First, a detailed description of the embedded multi-sensor system setup is given in section 2. Next, the perception function is presented and experimental results are discussed. Section 3 is devoted to 3D ego-localization using vision and proprioceptive sensors. The extrinsic calibration procedure is described in detail in section 4. Object localization and tracking are examined in section 5. Visual track confirmation and synchronization issues are discussed in section 6. Finally, the conclusion and perspectives of this work are presented.

2 Multi-Modal Perception System

2.1 Experimental Set-up

An experimental vehicle belonging to the Heudiasyc Laboratory was used for the implementation and validation of the global perception system in real-life conditions. As illustrated in Fig. 3, the vehicle is equipped with a 47cm-baseline Videre SVS. This SVS is composed of two CMOS cameras, with 4.5mm lenses configured to acquire 320x240 gray-scale images at 30 fps. The vision system provides essential information for ego-motion estimation and visual track confirmation. This system covers 45-degrees with an effective range up to 50 m in front of the vehicle.

Fig. 3 The experimental vehicle with the stereo vision system and the IBEO Alasca XT. (Velodyne is not used in this study)

A large surrounding region is covered by an IBEO Alasca XT lidar which transmits a sparse perception of the 3D environment at 15 Hz. Installed at the front of the vehicle, this sensor emits 4 crossed-scan-planes covering a 3.2° field of view in the vertical direction and 140° in the horizontal direction with an advertised 200m range. The ML lidar technology is particularly suitable for automotive applications since detected objects are not easily lost under pitch vehicle movements, in contrast to single-raw range-finders. Additionally, the 4-layer configuration allows the extraction of 3D scene structure attributes such as camber and curbs, since two layers are angled downwards.

The exteroceptive sensors (i.e. SVS and ML lidar here) were set up to ensure a complete Field-Of-View (FOV) overlap, as depicted in Fig. 4. The redundant sensing coverage is intended to achieve reliable object detection and tracking.

A CAN-bus gateway provides the speed of the rear-wheels (WSS) and the yaw rate of the vehicle (from the ESP). These sensors deliver complementary information for the vehicle localization task.

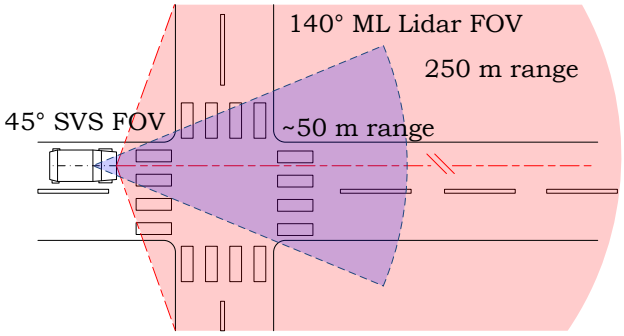


Fig. 4 Layout representation of the perception coverage provided by the exteroceptive sensors SVS and the ML lidar

2.2 Frames

The geometrical set-up of the multi-sensor system is formalized through four local sensor frames as illustrated in Fig. 5 and 6. The stereo vision camera system is represented by two classical pinhole camera models (i.e. with focal length f and the principal point coordinates $[u_0\ v_0]^T$ in pixels, assuming no distortion and zero skew [16]). The two cameras are rigidly linked and horizontally aligned at a baseline distance, b. As illustrated in Fig. 5, the reference frame of the SVS, denoted $\mathcal{S}$, is located midway between the two cameras. Information referenced w.r.t. the left camera frame, $\mathcal{G}$, can then be expressed in the $\mathcal{S}$ frame (X-Right, Y-Down and Z-Front) by a translation ${}^{\mathcal{G}}\mathbf{t}_{\mathcal{S}} = [-b/2\ \ 0\ \ 0]^T$. The SVS is fully calibrated and the delivered image pairs are rectified.

The ML-lidar measurements (i.e. a 3D points cloud) are reported in a Cartesian frame, denoted $\mathcal{L}$ (X-Front, Y-Left and Z-Up).

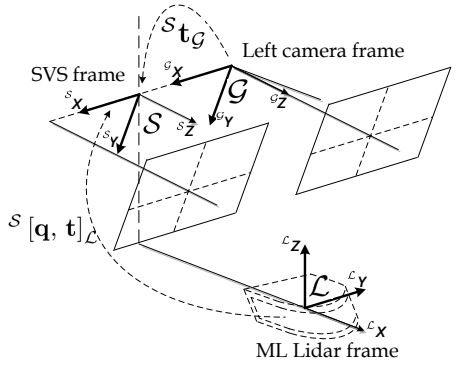


Fig. 5 The multi-sensor frames are located at the optical center of the left camera, $\mathcal{G}$, the midpoint of the SVS, $\mathcal{S}$, and the ML-lidar sensor, $\mathcal{L}$.

In order to sense information in a common perception space, the relative pose of the sensors frames (i.e. SVS and ML-lidar frames) has to be estimated through an extrinsic calibration procedure (Sensor Calibration module in Fig. 2). Extrinsic parameters can be obtained using the left camera images and the ML-lidar measurements, given that the frame transformations between the cameras composing the SVS are known (see Fig. 5). This process is presented in detail in section 4. The complete frame transformation from the lidar frame $\mathcal{L}$ into the vision frame $\mathcal{S}$ is denoted ${}^{\mathcal{L}}[\mathbf{q},\,\mathbf{t}]_{\mathcal{S}}$ and composed of a unit quaternion and a translation vector.

Additionally, the gyro and wheel speed sensor measurements are referenced with respect to the midpoint of the vehicle's rear-axis.

The ego-frame $\mathcal{E}$ (X-Right, Y-Down and Z-Front) is rigidly linked to the vehicle (i.e. body-frame). As this frame is chosen tangential to the road plane, it simplifies the geometrical information analysis for the visual lidar track confirmation (see Fig. 6).

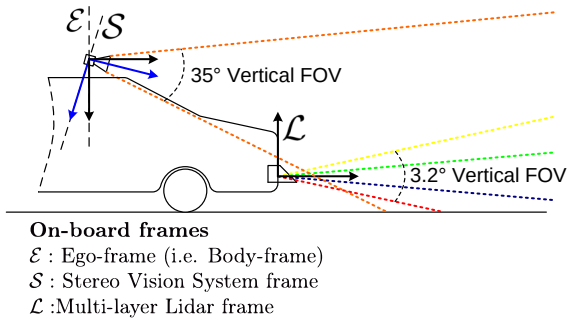


Fig. 6 On-board frames and exteroceptive vertical sensor coverage. The depicted lidar layer divergence and the camera FOV have been exaggerated for clarity.

3 3D Ego-Localization

3D ego-localization consists in estimating the 3D pose of the vehicle as a function of time with respect to a static frame lying, for instance, in a plane tangential to the earth's surface. Odometry methods using stereo vision systems can provide very precise 3D pose estimations based on multiple-view geometrical relations (i.e. 4-views or quadrifocal constraints) [6]. However, visual odometry may require considerable computation time. Here, 3D vehicle ego-localization is estimated using sparse visual odometry for rapid processing, aided by the embedded proprioceptive sensors of the vehicle [25].

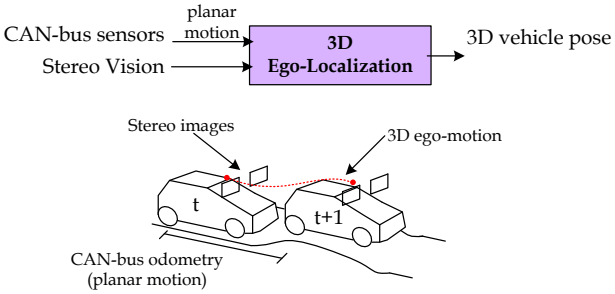


Fig. 7 Multi-modal 3D ego-localization scheme

3.1 Visual odometry aided by CAN-bus sensors

The ego-motion of the vehicle is defined by an elementary transformation (rotation-translation composition, 6 degrees-of-freedom) performed in an interval of time Δt. This 3D motion is represented by an axis-angle rotation and a translation vector, ${}^{\mathcal{S}(t-1)}[\Delta\boldsymbol{\omega},\,\Delta\mathbf{v}]^{T}_{\mathcal{S}(t)}$. First, an initial planar motion guess is computed using the proprioceptive sensors in Δt. Secondly, a 3D visual motion estimation algorithm is initialized with this motion guess and is then iteratively refined (see Fig. 7).

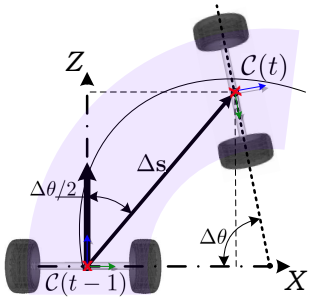


Fig. 8 Yaw rate-WSS dead-reckoning for planar odometry estimation

Let $\mathcal{C}(t)$ be the center of the body frame defined at time $t - \Delta t$. If the sampling frequency of the gyro and the WSS is high enough (about 40 Hz), the wheel speed is almost constant and the planar ego-motion can be approximated by the arc of a circle. As illustrated in Fig. 8, the planar ego-motion of the vehicle is modeled as follows [7]:

$$\Delta\boldsymbol{\omega}_0 = \begin{bmatrix} 0 \\ \Delta\theta \\ 0 \end{bmatrix} \quad \Delta\mathbf{v}_0 = \begin{bmatrix} \Delta s \cdot sin(\Delta\theta/2) \\ 0 \\ \Delta s \cdot cos(\Delta\theta/2) \end{bmatrix}$$

where $\Delta\theta$ is the angle obtained by integrating the yaw rate, and Δs is the integrated rear-wheel odometry in meters. $\Delta\boldsymbol{\omega}_0$ is a vector representing the axis-angle rotation of the vehicle motion and $\Delta\mathbf{v}_0$ is a vector representing the estimated displacement of the rear wheel axis center.

Using successive stereo image pairs, we obtain a set of n tracked stereo feature points denoted, $\mathsf{p}=\{p_1, ..., p_n\}$ and $\mathsf{p}' = \{p'_1, ..., p'_n\}$, (i.e. the corresponding left-hand and right-hand features respectively) and their corresponding optical flow constituting the image motion. For this purpose, a set of stereo feature points denoted, $\mathsf{p}^*=\{p^*_1, \ldots, p^*_n\}$ and $\mathsf{p}'^* = \{p^{'*}_1, \ldots, p^{'*}_n\}$, is extracted using Harris features [15] with a ZNCC (Zero-mean Normal Cross Correlation) criterion and image constraints (disparity and epipolar constraints) [28]. The set of stereo feature points, p^* and p'^*, is then tracked over time using Lucas-Kanade method [3].

Alternatively, a stereo feature can be "predicted" after a 3D motion of the vision system by using a warping function [6] based on geometrical constraints. These constraints are induced by the SVS configuration and the static scene assumption. The idea is to predict the sets $\hat{\mathsf{p}}$ and $\hat{\mathsf{p}}'$ as a function of the sets p^* and $\mathsf{p}^{'*}$ of stereo features at time $t-1$ (i.e. $t-\Delta t$) and the vehicle motion encapsulated in the trifocal tensors ${}^{l}\mathcal{T}_i^{jk}$ and ${}^{r}\mathcal{T}_i^{jk}$ [30].

Consider a stereo feature pair, $\{p^*, p'^*\} \forall\ p^* \in \mathsf{p}^*$, $p'^* \in \mathsf{p}'^*$. Thus, their predicted image position is given by:

$$\begin{bmatrix} \hat{p} \\ \hat{p}' \end{bmatrix} = \begin{bmatrix} p^* l'_j {}^{l}\mathcal{T}_i^{jk} \\ p'^* l_j {}^{r}\mathcal{T}_i^{jk} \end{bmatrix} \tag{1}$$

where l_j and l'_j are respectively the left and right image lines passing through the image points p^* and p'^* , and perpendicular to the epipolar line. ${}^{l}\mathcal{T}_i^{jk}$, is the trifocal tensor composed by the stereo image pair at time $t-1$ and the left image at time t. The second tensor, ${}^{r}\mathcal{T}_i^{jk}$, is composed by the stereo image pair at time $t-1$ and the right image at time t. It is worth recalling that the tensors ${}^{r}\mathcal{T}_i^{jk}$ and ${}^{l}\mathcal{T}_i^{jk}$ are nonlinear functions of the SVS parameters (i.e. intrinsic and extrinsic) and of the vehicle's motion ${}^{\mathcal{S}(t-1)}[\Delta\boldsymbol{\omega},\ \Delta\mathbf{v}]^T_{\mathcal{S}(t)}$.

However, since urban scenes are not composed exclusively of static objects, the static scene assumption is not respected. To address this issue, a robust iterative nonlinear minimization is performed according to the following criterion:

$$\min_{{}^{\mathcal{S}(t-1)}[\Delta\boldsymbol{\omega},\ \Delta\mathbf{v}]^T_{\mathcal{S}(t)}} (\epsilon) = \sum_{i=1}^{k} \mathbf{W} \left[||p_i - \hat{p}_i|| + ||p'_i - \hat{p'}_i|| \right] \tag{2}$$

where k, is the number of tracked stereo feature pairs and p_i and p'_i are the left- and right-tracked stereo features at time t. $\hat{p}_i$ and $\hat{p'}_i$ are the left and right stereo features at time t warped by the estimated motion (i.e. ${}^{\mathcal{S}(t-1)}[\Delta\boldsymbol{\omega},\ \Delta\mathbf{v}]^T_{\mathcal{S}(t)}$) and the warping function stated in Eq. 1. $\mathbf{W}$ is the weighting matrix, estimated by an M-estimator function [31] updated using the Iterative Re-weighted Least Squares algorithm (IRLS).

This robust minimization converges into a solution by rejecting the feature points that are mainly generated by mobile objects. The convergence is guaranteed if at least 50% of the stereo feature points correspond to static objects (i.e. the environment). The criterion of Eq. 2 is minimized by using the Levenberg-Marquardt Algorithm (LM) in an IRLS loop [31]. The convergence speed of the LM algorithm is increased using the planar ego-motion ${}^{\mathcal{S}(t-1)}[\Delta\boldsymbol{\omega}_0,\ \Delta\mathbf{v}_0]^T_{\mathcal{S}(t)}$, from the CAN-bus sensors. This information provides a close initialization guess, and thus helps to reduce the number of iteration cycles.

Once the ego-motion parameters are estimated, the vehicle position fix can be obtained by integrating all the successive estimates. With the aim of reducing noise before integrating the ego motion ${}^{\mathcal{S}(t-1)}[\Delta\boldsymbol{\omega},\ \Delta\mathbf{v}]^T_{\mathcal{S}(t)}$, a filter is implemented for the estimated parameters. To this end a constant accelerated model is considered, since the ego-vehicle can experience significant speed changes in braking situations. Regarding the axis-angle formalism for the attitude changes, $\Delta\boldsymbol{\omega}$, it will be remarked that for a non-holonomic moving vehicle, the variations of the attitude parameters are smooth, except for the yaw angle. The latter assumption does not hold for extreme driving situations (e.g rollover). Under these assumptions, the linear speed and the attitude changes are filtered using a single first-order model that reconstructs the derivatives of the ego-motion parameters. This approach is useful for predictive filtering.

Let $\mathbf{x}(t|t)$, be the state vector and $\mathbf{A}(t)$ the state transition matrix defined as follows:

$$\mathbf{x}(t|t) = \begin{bmatrix} \boldsymbol{\omega}(t) \\ \mathbf{v}(t) \\ \boldsymbol{\Omega}(t) \\ \mathbf{a}(t) \end{bmatrix} \quad \mathbf{A}(t) = \begin{bmatrix} I_{6\times6} & \Delta t \cdot I_{6\times6} \\ \mathbf{0}_{6\times6} & I_{6\times6} \end{bmatrix} \tag{3}$$

where $\mathbf{v}(t)$ and $\mathbf{a}(t)$ are the linear speed and acceleration. $\boldsymbol{\omega}(t)$ represents the derivatives of the axis-angle parameters which have been considered linear to

a drift $\boldsymbol{\Omega}(t)$, assumed to be constant over time. In this model, the sampling period, Δt, is not constant. The covariance of the model, denoted $\mathbf{Q}$, is chosen taking into account the errors due to the model approximation and, the covariance noise $\mathbf{R}$ is estimated considering a zero-mean Gaussian white noise.

Through the discrete Kalman filter equations [10], the predicted state (*a priori*), $\mathbf{x}(t|t-1)$, and its covariance, $\mathbf{P}(t|t-1)$, are estimated.

Once a new ego-motion estimation (i.e. state observation), denoted $\tilde{\mathbf{x}}(t)$, is obtained, the predicted state and its covariance are corrected providing the *a posteriori* state, $\mathbf{x}(t|t)$, and covariance, $\mathbf{P}(t|t)$.

The filtered ego-motion which is reconstructed from the state and the observation model, $\mathbf{H}$, is given by:

$$ {}^{\mathcal{S}(t-1)}[\Delta\boldsymbol{\omega},\ \Delta\mathbf{v}]^T_{\mathcal{S}(t)} = \mathbf{H}\cdot\mathbf{x}(t|t-1) \quad (4) $$

$$ with \quad \mathbf{H} = \left[\ \Delta t.I_{6\times6}\ \mathbf{0}_{6\times6}\ \right] $$

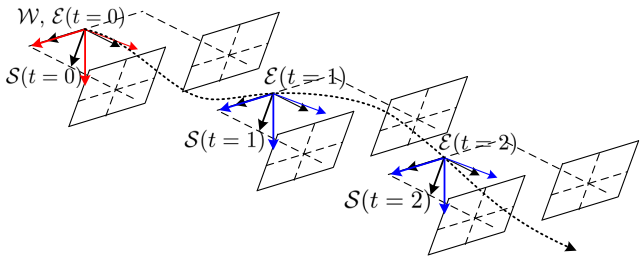


Fig. 9 World ($\mathcal{W}$) and Ego ($\mathcal{E}(t)$) and SVS ($\mathcal{S}(t)$) frames at 3 sampling times

Making use of the smooth estimates provided by the predictive filter using Eq. 4, the vehicle localization is estimated in the world frame. For this, let $\mathcal{W}$ be the world reference frame and $\mathcal{E}(t)$, the ego frame which is linked to the vehicle as illustrated in Fig. 6 and 9. The vehicle pose at time t is denoted ${}^{\mathcal{W}}\mathcal{E}(t) = {}^{\mathcal{W}}[\mathbf{q}(t),\ \mathbf{p}(t)]^T$ and is represented in the world frame by its attitude - ${}^{\mathcal{W}}\mathbf{q}(t)$ a unit quaternion - and its position - ${}^{\mathcal{W}}\mathbf{p}(t)$ - a vector in meters. It is obtained as follows:

$$ {}^{\mathcal{W}}\mathbf{q}(t) = {}^{\mathcal{E}}\mathbf{q}_{\mathcal{S}} \star {}^{\mathcal{S}}\mathbf{q}_{\mathcal{S}(t)} \quad \text{and} \quad {}^{\mathcal{W}}\underline{\mathbf{p}}(t) = {}^{\mathcal{E}}\mathbf{q}_{\mathcal{S}} \star {}^{\mathcal{S}}\underline{\mathbf{p}}_{\mathcal{S}(t)} \star {}^{\mathcal{E}}\bar{\mathbf{q}}_{\mathcal{S}} \quad (5) $$

where t denotes the temporal frame index and has been conveniently omitted for $t=0$, $\star$ denotes the quaternion multiplication operator, $\bar{\mathbf{q}}$ represents the corresponding quaternion conjugate and the underlined vector (e.g. $\underline{\mathbf{p}}$) denotes an expanded form (i.e. $\underline{\mathbf{p}} = \left[0,\ \mathbf{p}^T\right]^T$) for the use of the quaternion multiplication [14]. As stated in Eq. 5, the rotation ${}^{\mathcal{E}}\mathbf{q}_{\mathcal{S}}$ is used to compute the relative orientation of the world frame, $\mathcal{W}$, w.r.t. the SVS frame, $\mathcal{S}$, since $\mathcal{W}$ has been chosen as the initial position of the ego frame, $\mathcal{E}(t=0)$ ($\mathcal{W}$ is chosen to be coplanar to the road plane). The rigid transformation ${}^{\mathcal{S}}[\mathbf{q},\ \underline{\mathbf{p}}]_{\mathcal{S}(t)}$ corresponds to the visual odometry given by the following equations:

$$ {}^{\mathcal{S}}\mathbf{q}_{\mathcal{S}(t)} = {}^{\mathcal{S}}\mathbf{q}_{\mathcal{S}(t-1)} \star^{\mathcal{S}(t-1)} \mathbf{q}(\Delta\boldsymbol{\omega})_{\mathcal{S}(t)} \quad (6) $$

$$ {}^{\mathcal{S}}\underline{\mathbf{p}}_{\mathcal{S}(t)} = {}^{\mathcal{S}}\mathbf{q}_{\mathcal{S}(t-1)} \star^{\mathcal{S}(t-1)} \underline{\boldsymbol{\Delta\mathbf{v}}}_{\mathcal{S}(t)} \star^{\mathcal{S}} \bar{\mathbf{q}}_{\mathcal{S}(t-1)} +^{\mathcal{S}} \underline{\mathbf{p}}_{\mathcal{S}(t-1)} \quad (7) $$

where ${}^{\mathcal{S}(t-1)}\mathbf{q}(\Delta\boldsymbol{\omega})_{\mathcal{S}(t)}$ is the unit quaternion corresponding to the axis-angle rotation, $\Delta\boldsymbol{\omega}$, composing the ego-motion parameters ${}^{\mathcal{S}(t-1)}[\Delta\boldsymbol{\omega},\ \Delta\mathbf{v}]^T_{\mathcal{S}(t)}$.

3.2 Experimental Real Time 3D Ego-Localization Results

A data set was acquired in an urban environment composed of low-rise buildings, trees and moving objects (i.e. pedestrians and vehicles). During the experiment, the vehicle's speed was around 30 Km/h. The vehicle trajectory is a closed loop featuring pedestrians and other vehicles. Low-textured scenes (e.g. rural environments and parking lots) were not considered in this study.

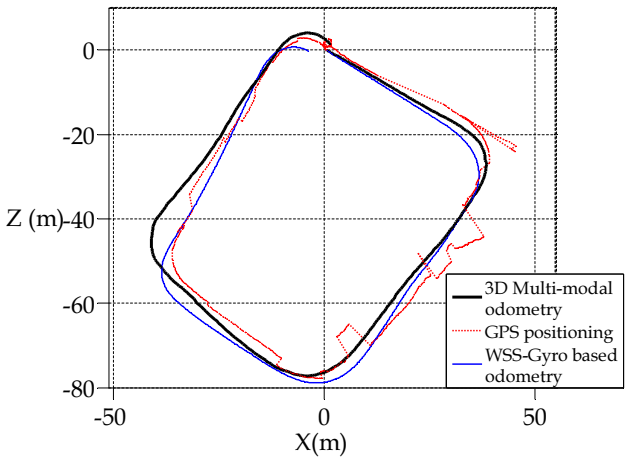


Fig. 10 Bird's eye view of the 3D Reconstructed Trajectory compared to WSS-Gyro odometry and GPS positioning

The 3D ego-localization function was implemented in C/C++. The 3D trajectory is reconstructed in real time and is obtained by integrating the ego-motion estimations. Fig. 10 illustrates one of the tests performed. This test consists of a 227m-clockwise loop (i.e. 90 seconds video sequence duration).

These results show that the cooperative strategy helps to cope with errors from the CAN-sensor-based odometry, errors mainly attributable to wheel slippage. This technique also improved the visual odometry performance in critical situations (e.g. high rotational speed

in 90° turns and roundabouts). These improvements were obtained as a result of the planar motion initialization which avoids any local minima ego-motion solution. It also enhances outlier rejection and reduces the minimization iteration cycles. The 3D ego-localization system performs quite well in situations where GPS cannot provide a precise position (see the GPS jumps in Fig. 10).

4 Exteroceptive Sensor Calibration (ML Lidar - Vision)

The extrinsic calibration of the exteroceptive sensors (i.e. the ML lidar and the stereo vision system) is necessary for sensing information in a common perception space. These extrinsic parameters link the local sensor frames by a rigid transformation estimated from the sensor measurements. Thus, the calibration process aims at estimating the parameters which minimize the transformation error between a common set of measurements.

4.1 Problem Statement

Different calibration methods have been proposed to estimate with accuracy the rigid transformation between a camera and a range sensor. Zhang et al. proposed a target-based calibration method between a camera and a single-row laser range finder [33]. This method was considerably improved by Dupont et al. in [8]. A targetless calibration approach for a 3D laser range finder was later presented by Scaramuzza et al. in [27]. Calibration can also be done by making use of the 4-layer provided by the ML lidar [24].

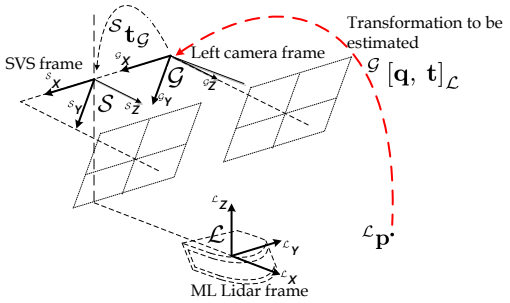


Fig. 11 Extrinsic Camera- Multi-layer Lidar Calibration

Let ${}^{\mathcal{L}}\mathbf{p}$, be the coordinates of a 3D laser point expressed in the lidar frame $\mathcal{L}$. Therefore, ${}^{\mathcal{S}}\mathbf{p}$, the corresponding coordinates of ${}^{\mathcal{L}}\mathbf{p}$ in the stereo vision frame, $\mathcal{S}$, is given by the following expression:

$$ {}^{\mathcal{S}}\underline{\mathbf{p}} = \left({}^{\mathcal{G}}\mathbf{q}_{\mathcal{L}} \star^{\mathcal{L}} \underline{\mathbf{p}} \star^{\mathcal{G}} \bar{\mathbf{q}}_{\mathcal{L}} +^{\mathcal{G}} \underline{\mathbf{t}}_{\mathcal{L}} \right) +^{\mathcal{S}} \underline{\mathbf{t}}_{\mathcal{G}} \tag{8} $$

where ${}^{\mathcal{G}}[\mathbf{q},\, \mathbf{t}]_{\mathcal{L}}$ is the rigid-body transformation from the lidar frame, $\mathcal{L}$, to the left camera frame, $\mathcal{G}$. ${}^{\mathcal{S}}\mathbf{t}_{\mathcal{G}}$ is the rigid-body transformation from the left camera frame, $\mathcal{G}$, to the stereo vision system frame, $\mathcal{S}$, as stated in Eq. 8.

As illustrated in Fig. 11, knowledge of the rigid transformation ${}^{\mathcal{G}}[\mathbf{q},\, \mathbf{t}]_{\mathcal{L}}$ means that SVS and ML lidar data can be merged.

The calibration method relies on a set of simultaneous detections of a circle-based calibration target by the camera and the multi-layer lidar. For every detection, the attitude and the position of the calibration target are estimated with respect to the camera and the lidar frames. The complete frame transformation from the lidar frame $\mathcal{L}$ into the vision frame $\mathcal{S}$, ${}^{\mathcal{S}}[\mathbf{q},\, \mathbf{t}]_{\mathcal{L}}$, is obtained as follows:

$$ {}^{\mathcal{S}}\mathbf{q}_{\mathcal{L}} = \mathbf{q}(I) \star^{\mathcal{G}} \mathbf{q}_{\mathcal{L}} \quad \text{and} \quad {}^{\mathcal{S}}\mathbf{t}_{\mathcal{L}} = {}^{\mathcal{S}}\mathbf{t}_{\mathcal{G}} +^{\mathcal{G}} \mathbf{t}_{\mathcal{L}} \tag{9} $$

where $\mathbf{q}(I)$ is the unit quaternion representation of the identity matrix since the frames $\mathcal{S}$ and $\mathcal{G}$ are aligned.

Once the lidar-vision frame transformation ${}^{\mathcal{S}}[\mathbf{q},\, \mathbf{t}]_{\mathcal{L}}$ is known, all the lidar measurements at time t are carried over to the vision frame $\mathcal{S}$.

4.2 Exteroceptive Sensor Calibration Results

The ML lidar - Vision calibration was obtained using 8 poses. Each pose comprises 20 lidar scans and the corresponding target image of the left camera. The transformation parameters and their corresponding intervals of confidence were computed as an Euler vector in radians and a translation in meters respectively.

Rotation parameters, $[\phi,\, \theta,\, \psi]^T$, are then converted into a quaternion, formalizing the extrinsic parameters to the form ${}^{\mathcal{G}}[\mathbf{q},\, \mathbf{t}]_{\mathcal{L}}$. Fig. 12 illustrates the resulting lidar data re-projection onto the stereo images.

5 Object Localization and Tracking

This stage corresponds to a multi-modal strategy which is able to estimate the planar trajectory of the neighboring objects as they move in the 3D scene. The proposed system operates in a fixed reference frame, since precise ego-localization is available (see section 3). The tracking strategy is presented below as follows: identifying the zone of interest, object detection, track prediction, and object-track association and updating. Finally, the experimental results of this perception function are reported.

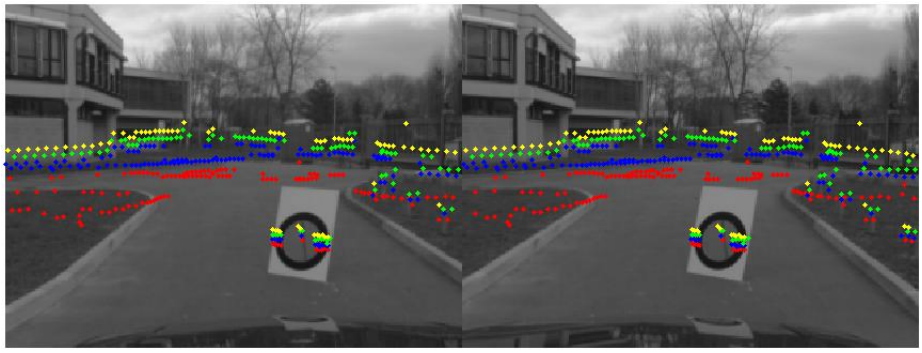


Fig. 12 ML Lidar data re-projection on stereo images

5.1 Identifying the Zone of Interest

Urban environments constitute a very complex 3D scene because of the presence of a large number of static and mobile objects. Different approaches may be used to reduce the complexity of the scene (i.e. the number of tracked objects) using, for instance, the temporal track persistence (i.e. forgetting factor), the dynamic of the tracks and the uncertainty of the track localization. For this study, we reduce the 3D observation space by detecting a zone of interest (ZOI) based on prior knowledge of the scene. This function was proposed and implemented in real time by Fayad et al. in [11]. The method is based principally on lidar scan histogram maxima detection.

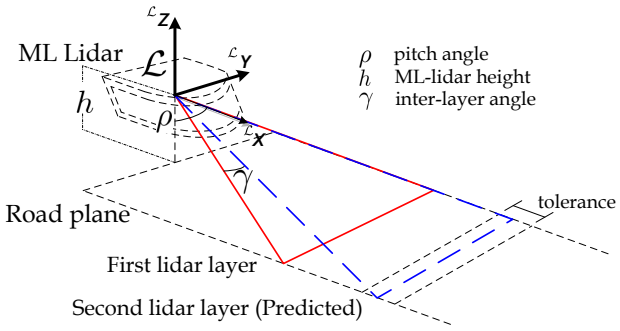


Fig. 13 Road lidar data filtering

A first lidar data filtering step is performed with the aim not only of improving the ZOI detection, but also of significantly attenuating scene clustering issues. This filtering involves detecting the 3D lidar data corresponding to the road surface, using the characteristic pattern observed when the two lower layers intersect the road plane at different angles, as illustrated in Fig. 13. The second lidar layer can be predicted from the first lidar layer and the geometrical constraints. The measurements belonging to the road plane are detected and excluded from further processing when the Euclidean error between the predicted and the measured layer falls below a predefined threshold. The pitch angle of the ML-lidar, denoted ρ, is estimated by a temporal filtering and is updated using the detected road impacts. The parameters h (ML-lidar height) and γ (inter-layer angle) are assumed to be known.

The zone of interest is characterized by two local limits in the x-axis direction of the lidar frame. As illustrated in Fig. 14, a 4-layer data scan is projected onto the ${}^{\mathcal{L}}xy$ plane (see upper subplot) and provides a convenient histogram ${}^{\mathcal{L}}y$ axis (see lower subplot). Objects such as security barriers, walls and parked vehicles limit the zone of interest effectively. The detected limits are then filtered using a fixed-gain Luenberger observer in order to reduce oscillations produced by significant changes in pitch. Intersections and roundabouts may lead to the loss of histogram peaks. In such cases no further updates of the ZOI are performed, meaning that it is the last ZOI estimation which is retained.

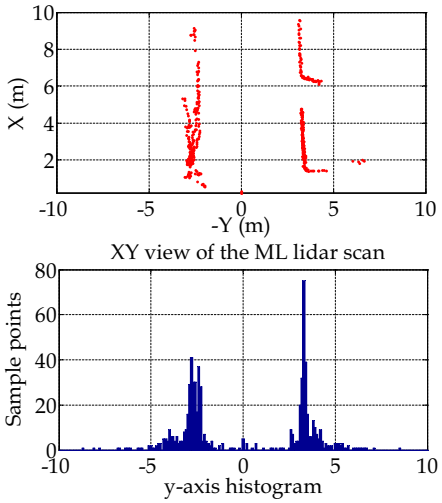


Fig. 14 Identifying the zone of interest using a y-axis histogram

5.2 Object Detection

Based on each ML-lidar scan, an object detection function delivers a set of neighboring objects obtained by a 3D Euclidean inter-distance clustering. Predefined geometric features [22] would be an alternative, but these require prior knowledge of the object. Detected objects are characterized by their planar location (i.e. ${}^{\mathcal{L}}Z = 0$) in the lidar frame $\mathcal{L}$, their dimension (i.e. bounding circle) and a detection confidence indicator [12], estimated using the following criteria:

- The ability of the ML sensor to detect vertical objects with respect to its range position.
- The beam divergence which worsens the measurement precision, particularly in situations of a non-perpendicular incidence angle.
- The theoretical maximum number of laser impacts (per layer) lying on a detected object. This factor can be computed as a function of the object dimension, the detection range and the laser scanner resolution.

Knowing the 3D localization of the vehicle, the detected objects can be referenced with respect to the world frame, $\mathcal{W}$. For instance, let ${}^{\mathcal{L}}\widetilde{\mathbf{y}}(t) = [x \;\; y \;\; 0]^T$ be the coordinates of a detected object at time t. Its corresponding localization in $\mathcal{W}$ can be computed as follows:

$$ {}^{\mathcal{S}(t)}\underline{\widetilde{\mathbf{y}}}(t) = \left({}^{\mathcal{S}}\mathbf{q}_{\mathcal{L}} \star^{\mathcal{L}} \underline{\widetilde{\mathbf{y}}}(t) \star^{\mathcal{S}} \bar{\mathbf{q}}_{\mathcal{L}}\right) +^{\mathcal{S}} \underline{\mathbf{t}}_{\mathcal{L}} \tag{10} $$

$$ {}^{\mathcal{W}}\underline{\widetilde{\mathbf{y}}}(t) = \left({}^{\mathcal{W}}\mathbf{q}(t) \star^{\mathcal{S}(t)} \underline{\widetilde{\mathbf{y}}}(t) \star^{\mathcal{W}} \bar{\mathbf{q}}(t)\right) +^{\mathcal{W}} \underline{\mathbf{p}}(t) \tag{11} $$

where ${}^{\mathcal{S}(t)}\widetilde{\mathbf{y}}(t)$ and ${}^{\mathcal{W}}\widetilde{\mathbf{y}}(t)$ are the corresponding coordinates of the detected object in the SVS and the world frame respectively.

Only the detected objects lying in the ZOI are localized w.r.t. the world frame using Eq. 10 and 11. They are then tracked independently using Kalman filters [13]. This tracking step improves the robustness of the perception scheme.

5.3 Track Prediction

The track state is described by ${}^{\mathcal{W}}\mathbf{y}(t|t)$ consisting of the ${}^{\mathcal{W}}XZ$ plane coordinates $(x(t),\, z(t))$ in meters and the planar velocity $(v_x(t),\, v_z(t))$ in m/s as follows:

$$ {}^{\mathcal{W}}\mathbf{y}(t|t) = \left[\, x(t) \;\; z(t) \;\; v_x(t) \;\; v_z(t) \,\right]^T \tag{12} $$

The object size is considered as an attribute of the track but is not included in the state. Assuming that the motion of the objects is linear and uniform, the prediction of the track state is given by the following evolution equation:

$$ {}^{\mathcal{W}}\mathbf{y}(t|t-1) = \mathbf{B}(t) \cdot {}^{\mathcal{W}}\mathbf{y}(t-1|t-1) \tag{13} $$

$$ \text{with} \quad \mathbf{B}(t) = \begin{bmatrix} I_{2\times 2} & \Delta t \cdot I_{2\times 2} \\ \mathbf{0}_{2\times 2} & I_{2\times 2} \end{bmatrix} $$

where ${}^{\mathcal{W}}\mathbf{y}(t|t-1)$ is the predicted state of the tracked object, $\mathbf{B}(t)$ is the state transition matrix and Δt is the sampling time period (which is not constant).

5.4 Track-Object Association and Updating

At each ML-lidar sampling, the tracks are updated. This involves matching the new detected objects and tracks using a mono-hypothesis assumption. The implemented association test relies on a nearest neighbor criterion [1] (i.e. $min(\eta)$) using the following normalized metric:

$$ \eta^2 = \mu(t)^T \left(\mathbf{M}(t|t-1) + \mathbf{N}\right)^{-1} \mu(t) + \ln\left(\det\left(\mathbf{M}(t|t-1) + \mathbf{N}\right)\right) \tag{14} $$

$$ \text{with} \quad \mu(t) = \mathbf{C} \cdot {}^{\mathcal{W}}\mathbf{y}(t|t-1) - {}^{\mathcal{W}}\widetilde{\mathbf{y}}(t)_{(x,z)} $$

$$ \text{and} \quad \mathbf{C} = \left[\, I_{2\times 2} \; \mathbf{0}_{2\times 2} \,\right] $$

where ${}^{\mathcal{W}}\widetilde{\mathbf{y}}(t)_{(x,z)}$ represents, with a slight abuse of the notation, the XZ coordinates of the detected object in the $\mathcal{W}$ frame, $\mathbf{C}$ is the observation matrix and $\mathbf{M}(t|t-1)$ the covariance matrix of the predicted state, ${}^{\mathcal{W}}\mathbf{y}(t|t-1)$. The first term of Eq. 14, corresponds to the classical Mahalanobis metric, and the second corresponds to a weighting factor, $\ln\left(\det\left(\mathbf{M}(t|t-1) + \mathbf{N}\right)\right)$, computed from the track imprecision. The uncertainties in the lidar localization of the objects and the object motion model are taken into account through the covariance of the measurement noise, $\mathbf{N}$ and the covariance of the state transition model, $\mathbf{O}$.

In order to cope with temporal object occlusions, the unmatched tracks are retained for a fixed time duration (for example, 2 seconds). However, setting a long prediction time may result in retaining track artifacts. The unmatched objects in the zone of interest generate new tracks until the algorithm reaches a previously defined maximum number of tracked objects. Here, the fixed number of tracks is set sufficiently high to track all the detected objects in the ZOI.

Finally, the states of the tracks and their corresponding covariances are improved by combining the information provided by the associated lidar objects positions and the predicted tracks [1]. For this we use the Kalman filter update equations.

It is worth recalling that the object tracking stage increases the robustness of the system by allowing the occlusion of objects, since tracks contain information that has been confirmed several times by the same source (the ML lidar here).

5.5 Experimental Results

In the experimental study of the multi-modal object localization and tracking system, the 3D ego-localization, the identification of the zone of interest and the detection of objects functions were processed in real time and their outputs were logged. The object tracking function was implemented in Matlab. The reported results were obtained in offline conditions, taking as the input the logged data.

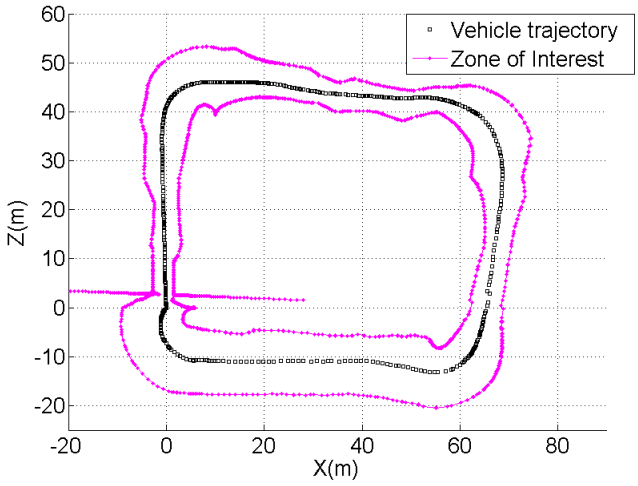


Fig. 15 Reconstruction of the Zone of Interest ($^{\mathcal{W}}XZ$ plane view)

Fig. 15 illustrates the XZ view of the reconstructed zone of interest in the fixed-reference frame. The reconstructed ZOI was obtained by exchanging geometrical data between the lidar and vision-based functions through the calibration information. It is important to remark that at the beginning of the test sequence (i.e. initial position (0,0) on XZ view), the vehicle remains stationary, which shows how the boundaries of the zone of interest converge. These results constitute a very interesting feature which can be linked to GIS (Geographic Information System) for map-matching applications.

Fig. 16 illustrates a zoomed area of the dynamic map. In this figure we focus on a tracked vehicle. The size of the track is represented by its bounding circle in red and its center as a red triangle. The detected track size changes as the surface is impacted by the ML lidar. The corresponding image track projections (3D red boxes) and their speed vector (green line) are also illustrated in the upper part of the figure. Observing the image projection of the track speed vector, one can see that the multi-modal system performs quite well.

Fig. 17 shows another section of the dynamic map. No ground truth for the track localization was available during the experiment. However, the reconstructed trajectory corresponds to the pedestrian's observed trajectory in the snapshot sequence.

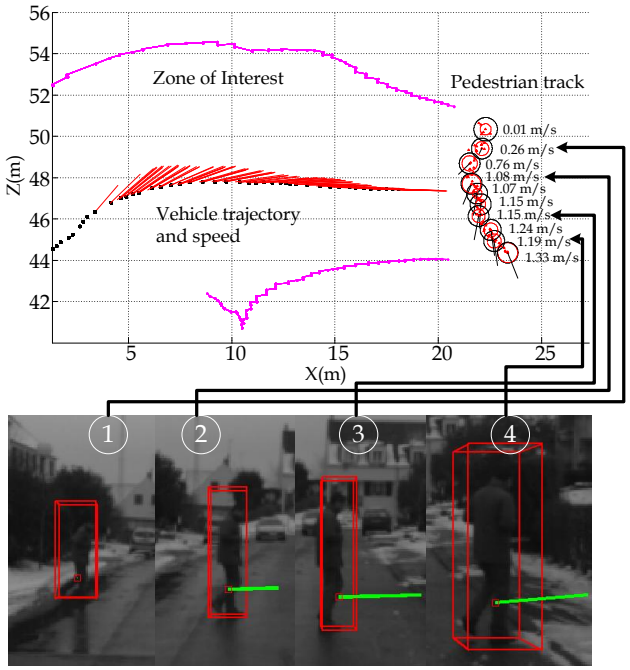


Fig. 17 Trajectory of a pedestrian

6 Visual Track Confirmation

The multi-modal system presented thus far provides a precise localization, with respect to a fixed-reference frame, of the vehicle and the surrounding objects that represent potential obstacles. However, the object detection relies on a single source: the ML lidar. This section presents visual track confirmation, which represents a way of increasing the integrity of the information provided by the system. Integrity is a key attribute of a perception system when dealing with ADAS [19]. For the system proposed in this paper, integrity means that the intended perception functions provide trustworthy information without false alarms or ambiguity.

Visual track confirmation is performed using the following strategy. First, each lidar-tracked object is transformed into the ego frame $\mathcal{E}(t)$. Its corresponding bounding cylinder (lidar bounding circle with an arbitrary height) is reprojected onto the stereo images. In each image, the track projection provides a Region Of Interest (ROI). Secondly, the pixels composing the ROI are reconstructed by stereopsis in the 3D space in order to provide a 3D point cloud. Thirdly, this set of 3D points is segmented into 2 clusters : the object and

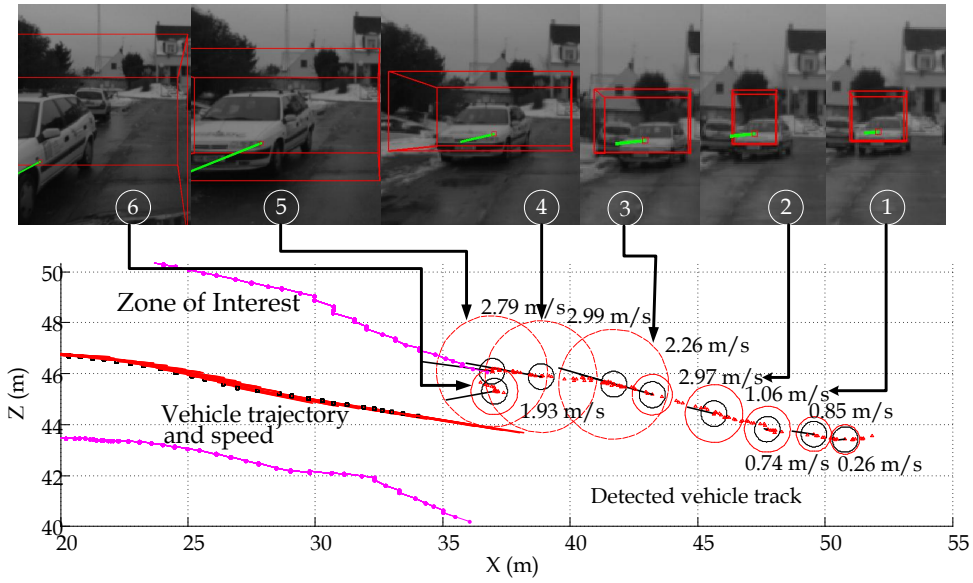


Fig. 16 Trajectory of a tracked vehicle

the background. Finally, the track is confirmed if one of the 3D point cluster centroids is validated using a Mahalanobis distance given a confidence level.

6.1 Region Of Interest in the images

The projection of an observed track onto the image plane can be done once the track position is referenced w.r.t. the camera frame. But, the 3D track fix must be firstly reconstructed, since the state vector provides its planar coordinates (i.e. ${}^{\mathcal{W}}XZ$ coordinates) only. To this end, the last object altitude (i.e. ${}^{\mathcal{W}}Y$ coordinate) associated to the track is used. In the following, the reconstructed track fix is denoted ${}^{\mathcal{W}}\mathbf{y}^{+}(t|t-1)$. The track is then localized in the ego frame, $\mathcal{E}(t)$ as follows:

$$ {}^{\mathcal{S}(t)}\underline{\mathbf{y}}^{+}(t|t-1) = {}^{\mathcal{W}}\bar{\mathbf{q}}(t) \star \left({}^{\mathcal{W}}\underline{\mathbf{y}}^{+}(t|t-1) - {}^{\mathcal{W}}\underline{\mathbf{p}}(t) \right) \star {}^{\mathcal{W}}\mathbf{q}(t) \quad (15) $$

$$ {}^{\mathcal{E}(t)}\underline{\mathbf{y}}^{+}(t|t-1) = {}^{\mathcal{E}}\mathbf{q}_{\mathcal{S}} \star {}^{\mathcal{S}(t)}\underline{\mathbf{y}}^{+}(t|t-1) \star {}^{\mathcal{E}}\bar{\mathbf{q}}_{\mathcal{S}} \quad (16) $$

where ${}^{\mathcal{S}(t)}\underline{\mathbf{y}}^{+}(t|t-1)$ is the reconstructed track fix in the $\mathcal{S}(t)$ frame and ${}^{\mathcal{E}(t)}\mathbf{y}^{+}(t|t-1)$ is the resulting position of the track in $\mathcal{E}(t)$.

The ROI is characterized by re-projecting the bounding box vertices of the track on the images. These vertices are estimated from the track size (the track height is known *a priori*) and its 3D centroid position (see Fig. 19).

The track position is projected onto the image plane by the following equation:

$$ \begin{bmatrix} u \\ \mathrm{v} \\ 1 \end{bmatrix} \sim \mathbf{K} \cdot \left(\left({}^{\mathcal{S}}\mathbf{q}_{\mathcal{E}} \star {}^{\mathcal{E}(t)}\underline{\mathbf{y}}^{+}(t|t-1) \star {}^{\mathcal{S}}\bar{\mathbf{q}}_{\mathcal{E}} \right) + {}^{\mathcal{G}}\underline{\mathbf{t}}_{\mathcal{S}} \right) \quad (17) $$

where $[u\ v\ 1]^T$ are the image coordinates and $\mathbf{K}$ is the intrinsic camera matrix. ${}^{\mathcal{S}}\mathbf{q}_{\mathcal{E}}$ corresponds to the conjugate quaternion of ${}^{\mathcal{E}}\mathbf{q}_{\mathcal{S}}$ and ${}^{\mathcal{G}}\mathbf{t}_{\mathcal{S}} = -{}^{\mathcal{S}}\mathbf{t}_{\mathcal{G}}$. The operator $\sim$ means up to a scale factor.

6.2 3D dense reconstruction of the ROI

Every pair of images contains 3D dense information of the scene since the pixel images correspondence and the camera parameters are known. This information can be represented by a disparity map which in this study is assumed to be referenced w.r.t. the left camera of the SVS.

The 3D dense reconstruction of the ROI consists in overlapping the ROI and the disparity map as illustrated in Fig. 18. The set of corresponding disparity values is then extracted. Finally, the 3D coordinates, $[x\ y\ z]^T$, of each pixel are estimated by performing a classical triangulation process [16].

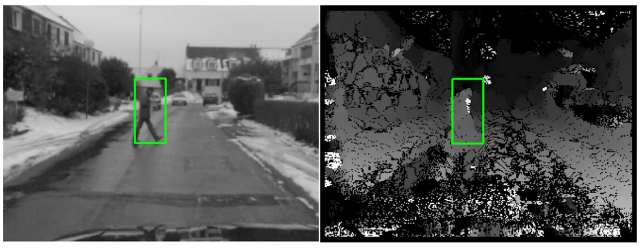


Fig. 18 3D dense reconstruction of the ROI

6.3 Track confirmation

Conceptually, the existence of a tracked object is confirmed if its position matches the visual 3D points. One

possible strategy would be to test all the points in order to determine the percentage that match with the track. However, this approach may be time consuming and is not compatible with an embedded system. An alternative is to cluster the visual points to apply a feature-like test. Because of the imprecision of the lidar object model, the ROI usually contains the imaged object and some scene background, as illustrated in Fig. 18. For this reason, clustering the 3D points can be seen as an ideal solution for simplifying and speeding up the process.

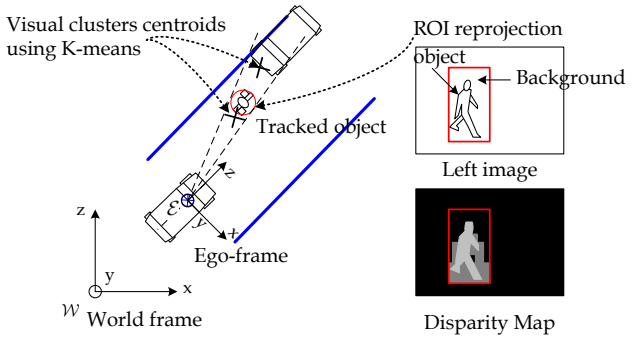


Fig. 19 Visual 3D Track Confirmation

Assuming that the objects and the scene background in the ROI are distinguishable in the 3D space (see the example shown in Fig. 19), the reconstructed 3D points can be segmented into two clusters: track and background. This assumption is usually justifiable, because the tested objects are located on the ZOI which corresponds to the navigable space, as shown in Fig. 20. For this kind of clustering, a K-means method [20] is particularly suitable, since the number of classes is known. The clustering is based on the Euclidean distance between 2 clusters characterized by theirs centroids, ${}^{\mathcal{E}(t)}\mathbf{c}_j$ for all $j = \{1, 2\}$, and their associated points. Two particular cases can occur, but they have no impact. If the ROI contains only the object (i.e. no background is included), the two clusters obtained will remain close to the real track. If no real object is included in the ROI (i.e. a lidar false detection), the clusters will form part of the background.

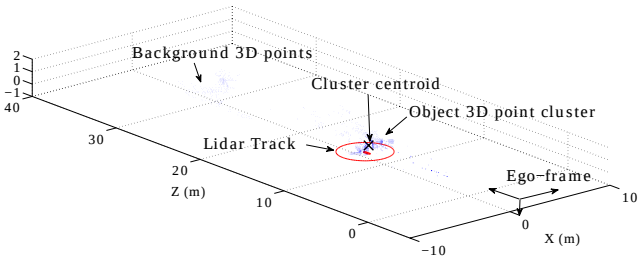


Fig. 20 3D visualization of a confirmed track using 3D point clusters

Once the clusters have been determined, the uncertainty associated with each cluster needs to be estimated. As the centroid of every cluster is localized from a set of reconstructed points, its uncertainty is strongly correlated to the triangulation process quality. As shown in [2], the confidence in the range of each centroid, ${}^{\mathcal{E}(t)}\mathbf{c}_j$ can be estimated, up to a tolerance factor α, by the following:

$$\tau_j = \begin{cases} 1 - \left(1 - \left(\alpha \cdot \frac{b \cdot f}{{}^{\mathcal{E}(t)}\mathbf{c}_{(z),j}}\right)^2\right) & , \alpha < \frac{{}^{\mathcal{E}(t)}\mathbf{c}_{(z),j}}{b \cdot f} \\ 1 & , \alpha \geq \frac{{}^{\mathcal{E}(t)}\mathbf{c}_{(z),j}}{b \cdot f} \end{cases} \tag{18}$$

where ${}^{\mathcal{E}(t)}\mathbf{c}_{(z),j}$ refers to the z metric coordinate of the centroid ${}^{\mathcal{E}(t)}\mathbf{c}_j$ and τ_j is a score which expresses the confidence in the cluster centroid fix in depth w.r.t the SVS. This score is directly derived from the first-order approximation of the joint probability distribution of the centroid image projection on the left and right image planes. The error tolerance factor, α, is set in meters, taking into account the image resolution, the focal distance, the baseline of the SVS and the accuracy of the SVS calibration.

In order to perform a normalized test between the clusters and the track to be verified, the covariance of the 2D position of the cluster centroid $\mathbf{U}$ can be estimated using this confidence score τ_j:

$$\mathbf{U} = \begin{bmatrix} \frac{1}{k_1 \cdot \tau_j} & 0 \\ 0 & \frac{1}{k_2 \cdot \tau_j} \end{bmatrix} \tag{19}$$

where the weighting parameters k_1 and k_2 can be chosen on the basis that the reconstruction error regarding depth has more impact in the longitudinal direction (${}^{\mathcal{E}(t)}Z$ axis) than transversely (i.e. $k_2 > k_1$: typically one can choose $k_2 = 2k_1$).

Each cluster is tested w.r.t the ML-lidar tracked object using a Mahalanobis distance, ξ, with respect to the track position:

$$\xi_j^2 = \kappa \cdot (\mathbf{U} + \mathbf{M}(t|t-1))^{-1} \cdot \kappa^T \tag{20}$$

$$\text{with} \quad \kappa = {}^{\mathcal{E}(t)}\mathbf{c}_{(x,z),j} - {}^{\mathcal{E}(t)}\mathbf{y}^+(t|t-1)_{(x,z)}$$

where ${}^{\mathcal{E}(t)}\mathbf{c}_{(x,z),j}$ and ${}^{\mathcal{E}(t)}\mathbf{y}(t|t-1)_{(x,z)}$ are respectively the ${}^{\mathcal{E}(t)}XZ$ coordinates of the centroid cluster to be tested and the tracked object. The matrix $\mathbf{M}(t|t-1)$ is the covariance of the tracked object location. If $min(\xi)$ is below a certain threshold (typically corresponding to 3 standard deviations), this means that one visual cluster matches with the track in position. A match confirms the existence of the tracked object. Integrity is automatically satisfied, since two independent sources have been used.

6.4 Sampling synchronization issues

The visual track confirmation scheme relies on three main functions interacting together: the disparity map computation, the 3D-ego localization and the object detection. These functions are asynchronous and run in different threads at different frequencies (26, 16 and 15 Hz respectively) which are quite constant as illustrated in Fig. 21.

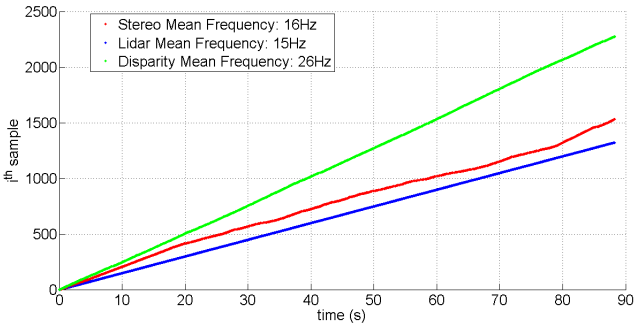


Fig. 21 Sampling function frequencies

In order to address these synchronization issues, predictive filters are used based on accurate stamped data. The possible processing tasks over time are:

- A new 3D-ego localization is available: the predictive filter of the ego-motion is time-updated and its state is corrected.
- New objects are detected by the ML-lidar: the last known vehicle localization is predicted up to this time. Then, these objects are localized in the world frame and the tracked objects are updated.
- A new disparity map is available: the vehicle localization and the tracked objects are extrapolated at this time. Predicted objects are localized in the ego-frame using the predicted vehicle pose. Candidate tracks are proposed, to be confirmed using the disparity map.

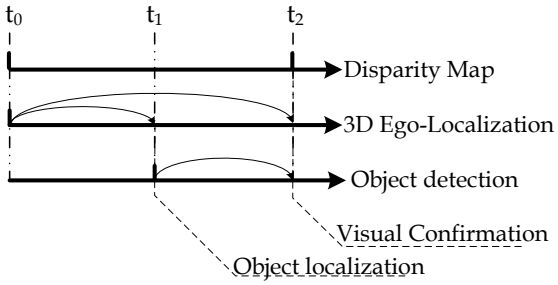


Fig. 22 Example of possible data arrival

Fig. 22 shows an example of possible measurements arrivals. At t_0, the disparity map and the localization information are available but there is no object. Thus, only the vehicle pose is updated. At t_1, a set of objects have been detected. They are localized using the predicted vehicle pose. At t_2, tracks can be confirmed using the predictions of the objects and the vehicle pose.

We remarked during our experiments that this mechanism dramatically increases the performance of the system, particularly when the scene includes a significant dynamic content.

6.5 Experimental results

The data output log of the disparity map, 3D-ego localization and object detection functions became the input into the object tracking and track confirmation functions that were post-processed using Matlab.

Use Cases and Evaluation Methodology In order for the performance of the system to be evaluated, we report experimental results for the visual confirmation function using five sequences. These use cases are relevant to common scenarios in urban environments. Fig. 23 gives a graphical description of the evaluated situations involving three kinds of mobile objects: pedestrians, wheelchair pedestrians and cars.

In the reported experiments, the ML-lidar did not give rise to any misdetections. The evaluation methodology aims at quantifying the percentage of time during which the object tracking function becomes unavailable because of visual non-confirmations.

The ground truth was referenced manually in the left image plane of the SVS: the center point coordinates of the observed objects of interest were selected, frame by frame. All the objects considered in the ground truth were localized in a common perception region for the SVS and the ML-lidar. The confirmation track rate was obtained by counting the number of times where the bounding box of the confirmed track contains the ground truth.

Real Data Results The results obtained using the ground truth are reported in Table 1. A total of 650 frames in 5 different situations showed that at least 81% of the time, the detected objects of interest were confirmed by the two modalities. Although one may conclude that the visual track confirmation may sometimes reduce the number of true positives, it should nevertheless be remarked that the confirmed tracks enhance the integrity of the perception process.
The false alarm rejection rate was also evaluated by providing manually a phantom track located four meters in

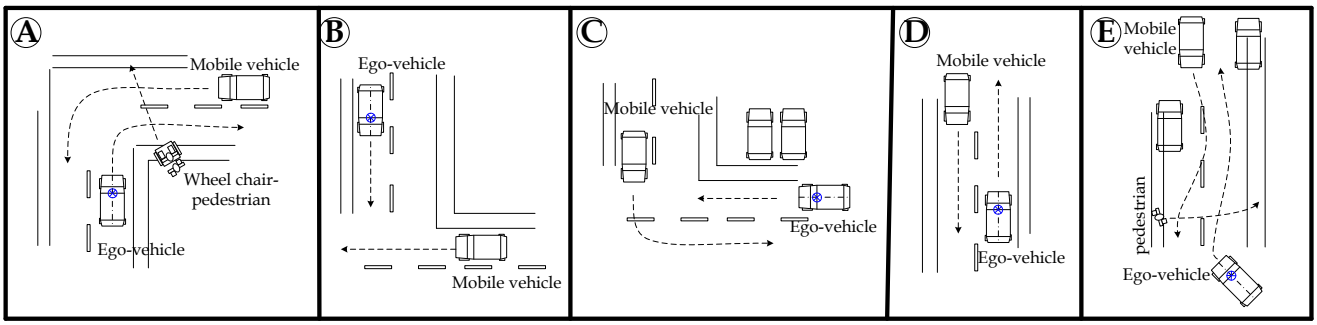


Fig. 23 Use cases considered in the evaluation test

front of the vehicle at every sample time (which means that the lidar is providing a 100% FA rate). Using a sequence of 2856 frames, with no objects on the road, visual confirmation only dealt with spurious tracks. Five false alarm conditions have been tested by changing the phantom track size from 80cm to 2m (see Table 2. As the proposed method can eliminate up to 94% of FA as a result.

Table 1 Rate of detected objects confirmed by vision

Video Sequence	A	B	C	D	E
Duration (s)	5	4	5	6	10
Number of Analyzed Frames	110	90	90	125	235
Number of scans	78	51	62	94	137
Positioning updates	72	37	65	121	160
Visual Confirmation Rate (%)	100	100	81.8	98.5	83.5

In use cases C and E, it was observed that large changes in vehicle pitch angle can influence the precision of object tracking, since object motion is considered to be planar and the vehicle pitch angle is unknown.

Table 2 False Alarm Rejection for 2856 frames

Track size (m)	0.8	1	1.2	1.5	2
Erroneous Confirmations	170	207	208	219	184
FA Rejection Rate (%)	94.04	92.75	92.72	92.33	93.55

In Fig. 24, the left-hand side illustrates the world map where the ego-vehicle and the detected objects are localized and tracked. The right-hand side of the figure shows the reconstructed points of the ROI image. Looking at the ego-map, it will be remarked that one of the centroids of the clustered point cloud has been matched with the track. This match confirms the existence of the detected object, as illustrated in the upper image in Fig. 25.

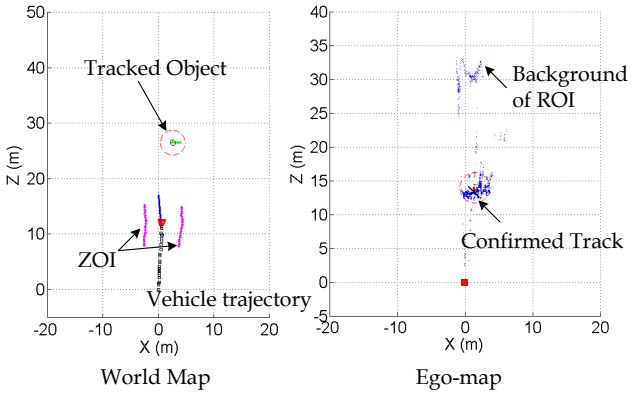


Fig. 24 Example of a confirmed tracked object using the SVS

Fig. 25 presents some examples of confirmed objects. Their bounding boxes (in red) and their speed vector projections (in green) show a good localization, even for fast-moving objects. These results validate the synchronization strategy.

7 Conclusion and Future Work

An asynchronous embedded multi-modal system for object localization and tracking has been proposed and experimentally validated. The approach presented here provides a 3D *dynamic* map of a vehicle's surroundings. The method merges sensor data to achieve a robust, accurate 3D ego-localization. This function is combined with a lidar-based object tracking focused on a zone of interest providing the trajectories and speeds of objects as they move in the space. A visual confirmation function integrated into the tracking system checks the integrity of the provided information. To this end, ROI are processed in a dense way. 3D points are reconstructed and compared to the lidar-tracked object. This scheme makes full use of the broad functional spectrum of stereo vision systems. Synchronization issues are taken into account to ensure the temporal consistency of the system.

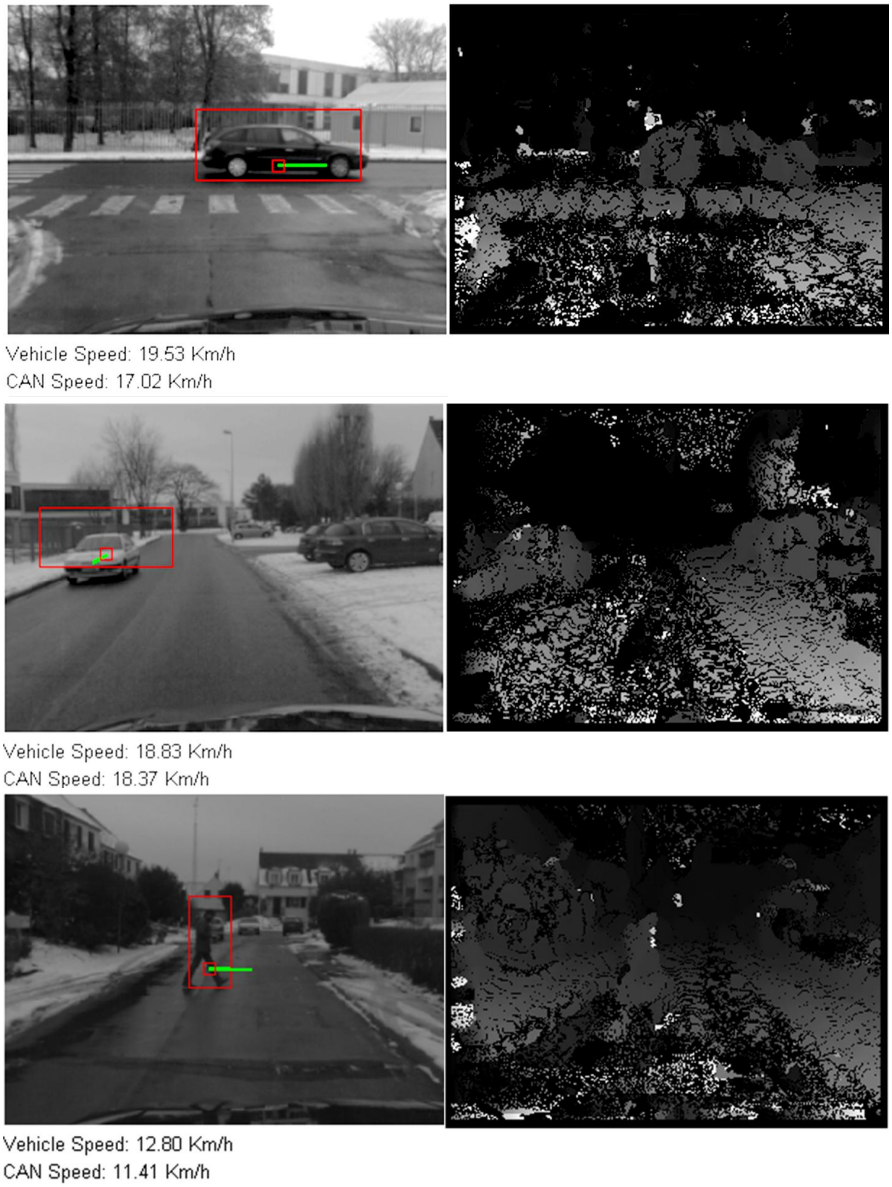


Fig. 25 Confirmed objects. The top figure illustrates the visual confirmation of a highly dynamic object (Scenario B in Fig. 23). The middle figure shows a vehicle having just turned. Here the ROI is delocalized because of pitch changes, given that the ego-vehicle is accelerating (Senario C in Fig. 23). The bottom figure shows the confirmed pedestrian when crossing the road in Scenario E (Fig. 23)

The results obtained show the effectiveness and the integrity of the proposed strategy. The visual confirmation function was tested in five different scenarios, demonstrating a good confirmation rate. Although visual confirmation inevitably reduces the availability of the detection function, the obtained rate would appear to be compatible with the development of ADAS functions (e.g. collision detection and avoidance). The inclusion of a visual object recognition function for selecting the most suitable object motion model might improve the tracking process. This is one perspective of our research.

Acronyms

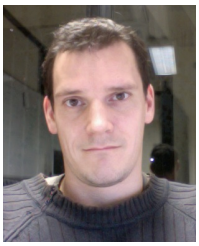
ADAS	Advanced Driver Assistance System
GPS	Global Positioning System
SVS	Stereo Vision System
FA	False alarms
FOV	Field-Of-View
WSS	Wheel Speed Sensors
ESP	Electronic Stability Programme
ML	Multi-Layer
CAN	Controller Area Network
ROI	Region Of Interest
ZOI	Zone Of Interest

Acknowledgements The authors wish to thank Fadi Fayad for the real time implementation of the ML lidar detection functions and Gerald Dherbomez and Thierry Monglon for their experimental support.

References

1. Samuel S. Blackman and Robert Popoli. *Design and Analysis of Modern Tracking Systems*. Artech House, Incorporated, 1999.
2. Steven D. Blostein and Thomas S. Huang. Error analysis in stereo determination of 3d point positions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:752–765, 1987.
3. Jean-Yves Bouguet. Pyramidal implementation of the lucas kanade feature tracker description of the algorithm. Technical report, Intel Corporation Microprocessor Research Labs, 2002.
4. A. Broggi, P. Cerri, S. Ghidoni, P. Grisleri, and H.G. Jung. A new approach to urban pedestrian detection for automatic braking. *Journal of Intelligent Vehicles Systems*, 10(4):594–605, 2009.
5. Cindy Cappelle, Maan E. El Najjar, Denis Pomorski, and Francois Charpillet. Multi-sensors data fusion using dynamic bayesian network for robotised vehicle geolocalisation. *IEEE Fusion*, 2008.
6. Andrew Comport, Ezio Malis, and Patrick Rives. Accurate quadrifocal tracking for robust 3d visual odometry. *IEEE International Conference on Robotics and Automation*, pages 40–45, April 2007.
7. Gregory Dudek and Michael Jenkin. *Springer Handbook of Robotics*, chapter Inertial Sensors, GPS, and Odometry, pages 477–490. Springer Berlin Heidelberg, 2008.
8. Romain Dupont, Renaud Keriven, and Philippe Fuchs. An improved calibration technique for coupled single-row telemeter and ccd camera. *The International Conference on 3-D Digital Imaging and Modeling*, 2005.
9. Hugh Durrant-Whyte and Tim Bailey. Simultaneous localisation and mapping (slam). *IEEE Robotics & Automation Magazine*, 13:99–110/108 – 117, 2006.
10. Hugh Durrant-Whyte and Thomas C. Henderson. *Springer Handbook of Robotics*, chapter Multisensor Data Fusion, pages 585–610. Springer Berlin Heidelberg, 2008.
11. Fadi Fayad and Veronique Cherfaoui. Tracking objects using a laser scanner in driving situation based on modeling target shape. *IEEE Intelligent Vehicles Symposium*, 1:44–49, 2007.
12. Fadi Fayad and Veronique Cherfaoui. Object-level fusion and confidence management in a multi-sensor pedestrian tracking system. *IEEE Int. Conf. on Multisensor Fusion and Integration for Intelligent Vehicles*, 1:58–63, 2008.
13. Mohinder S. Grewal and Angus P. Andrews. *Kalman Filtering: Theory and Practice Using Matlab*. Wiley-Interscience Publication, 2001.
14. Andrew J. Hanson. *Visualizing Quaternions*. Morgan Kaufmann, 2006.
15. Chris Harris and Mike Stephens. A combined corner and edge detector. *Proceedings fo The Fourth Alvey Vision Conference*, 1:147–151, 1988.
16. R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision. Second Edition*. Cambridge, 2003.
17. Raphael Labayrade, Cyril Royere, Dominique Gruyer, and Didier Aubert. Cooperative fusion for multi-obstacles detection with use of stereovision and laser scanner. *Autonomous Robots*, 19:117–140, 2005.
18. Bastian Leibe, Nico Cronelis, Kurt Cornelis, and Luc Van Gool. Dynamic 3d scene analysis from a moving vehicle. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR*, 1, 2007.
19. Martin E. Liggins, David L. Hall, and James Llinas. *Handbook of Multi-Sensor Data Fusion*. CRC Press, 2008.
20. J. MacQueen. Some methods for classification and analisys multivariate observations. *Berkeley Symposium on Mathematical Statistics and Probability*, 1:281–297, 1967.
21. T. Miyasaka, Y. Ohama, and Y. Ninomiya. Ego-motion estimation and moving object tracking using multi-layer lidar. *IEEE Intelligent Vehicles Symposium*, 1:151–156, 2009.
22. Fawzi Nashashibi, Ayoub Khammari, and Claude Laurgeau. Vehicle recognition and tracking using a generic multisensor and multialgorithm fusion approach. *International Journal of Vehicle Autonomous Systems*, 6:134–154, 2008.
23. Anna Petrovskaya and Sebastian Thrun. Model based vehicle tracking in urban environments. *IEEE International Conference on Robotics and Automation, Workshop on Safe Navigation*, 1:1–8, 2009.
24. Sergio A. Rodriguez, Vincent Fremont, and Philippe Bonnifait. Influence of intrinsic parameters over extrinsic calibration between a multi-layer lidar and a camera. In *IEEE IROS 2nd Workshop on Planning, Perception and Navigation for Intelligent Vehicles*, volume 1, pages 34–39, 2008.
25. Sergio A. Rodriguez, Vincent Fremont, and Philippe Bonnifait. An experiment of a 3d real-time robust visual odometry for intelligent vehicles. In *IEEE International Conference on Intelligent Transportation Systems*, volume 1, pages 226 – 231, Saint Louis, USA, 2009.
26. SAFESPOT. Cooperative vehicles and road infraestructure for road safety. http://www.safespot-eu.org/.
27. D. Scaramuzza, A. Harati, and R. Siegwart. Extrinsic self calibration of a camera and a 3d laser range finder from natural scenes. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1:4164–4169, 2007.
28. Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47:7–42, 2001.
29. Yaakov Bar Shalom and William Dale Blair. *Multitarget/Multisensor Tracking: Applications and Advances*. Artech House Publishers, 2000.
30. A. Shashua and M. Werman. On the trilinear tensor of three perspective views and its underlying geometry. In *In Proc. of the International Conference on Computer Vision (ICCV)*, June 1995.
31. Charles V. Stewart. Robust parameter estimation in computer vision. *Society for Industrial and Applied Mathematics*, 41(3):513–537, 1999.
32. Chieh-Chih Wang, Charles Thorpe, Martial Herbert, Sebastian Thrun, and Hugh Durrant-Whyte. Simultaneous localization, mapping and moving object tracking. *International Journal of Robotics Research*, 26:889–916, 2007.
33. Qilong Zhang and Robert Pless. Extrinsic calibration of a camera and laser range finder (improves camera calibration). *IEEE/RSJ. Intelligent Robots and Systems, 2004.*, 3:2301–2306, 2004.

Sergio A. Rodríguez F. was born in Bucaramanga, Colombia, in 1982. He received his Engineer's degree in Mechatronics *cum laude* from the Universidad Autonoma de Bucaramanga, Colombia, in 2005 and his M.S. degree in Control of Systems from the University of Technology of Compiegne, France, in 2007. He is currently a Ph.D. student at the HEUDIASYC CNRS Laboratory, France.



Vincent Fremont received his M.S. degree in Automatic Control and Computer Science from the Ecole Centrale de Nantes, France, in 2000 and his Ph.D. in Automatic Control and Computer Science from the Ecole Centrale de Nantes, France, in 2003. He is an associate professor in the Department of Computer Engineering at the UTC. His research interests in the HEUDIASYC CNRS Laboratory are computer vision, camera-based calibration, 3D reconstruction and motion analysis for dynamic scenes.

Philippe Bonnifait graduated from the Ecole Superieure d'Electronique de l'Ouest, France, in 1992 and gained his Ph.D. in Automatic Control and Computer Science from the Ecole Centrale de Nantes, France, in 1997. In December 2005 he obtained the Habilitation à Diriger des Recherches from the University of Technology of Compiègne. He joined the HEUDIASYC CNRS Laboratory, France, in September 1998. Since Sept. 2007 he has been professor and head of a research group in Robotics, Automation and Embedded Systems. His current research interests are in Intelligent Vehicles and Advanced Driving Assistance Systems, with particular emphasis on dynamic ego-localisation based on multisensor-fusion and tracking.

Véronique Cherfaoui received her M.S. degree in computer science from Lille University, France, in 1988 and her Ph.D. degree in control of systems from the University of Technology of Compiegne, France in 1992. She defended an "*Habilitation à Diriger des Recherches*" in 2009. She is now an associate professor in the Computer Engineering Department at the University of Technology of Compiegne. Her research interests in the HEUDIASYC CNRS Laboratory are data fusion algorithms in distributed architecture, data association and real-time perception systems for intelligent vehicles.