# On the Design of High Accuracy Rail Digital Maps based on Sensor Fusion

Sara Baldoni, *Roma Tre University and RadioLabs*
Federica Battisti, *University of Padova*
Michele Brizzi, *Roma Tre University and RadioLabs*
Giusy Emmanuele, *RFI*
Alessandro Neri, *Roma Tre University and RadioLabs*
Luca Pallotta, *Roma Tre University*
Agostino Ruggeri, *RadioLabs*
Alessia Vennarini, *RadioLabs*

## BIOGRAPHY

**Sara Baldoni** is a Ph.D. candidate in Applied Electronics at Roma Tre University. Her main research interests are in the area of Communication Security and Navigation and Localization Systems.

**Federica Battisti** is currently tenure-track Assistant Professor in the Department of Information Engineering at University of Padova. Her main research interests are in the field of signal and image processing.

**Michele Brizzi** is a Ph.D. candidate in Applied Electronics at Roma Tre University. His main research interests are in the field of image processing and imaging sensors for visual navigation.

**Giusy Emmanuele** MSc. in Automation Engineering and Control of Complex Systems. She is a RAILGAP project coordinator. She was involved in ERSAT Program, made of several satellite project ERSAT GGC, DB4Rail, SBS phase 2 and Gate4Rail.

**Alessandro Neri** is full professor in Telecommunications at Roma Tre University. Since 2009, he is the President of RadioLabs, a not-for-profit research center based on the partnership between universities and industries. His research activity has mainly been focused on information theory, signal and image processing, location and navigation technologies.

**Luca Pallotta** is a non-tenured Assistant Professor at Department of Engineering of Roma Tre University. His research interests lie in the field of radar target detection, automatic target recognition, polarimetric SAR image classification, and statistical signal processing with emphasis on radar/SAR signal processing.

**Agostino Ruggeri** is a senior network engineer at RadioLabs, Italy. He has experience in applied research programs and in the field of GNSS and TLC applications for railway safety-related systems. He participated to several international projects funded by ESA and the EC concerning the employment of Next Generation Communication Systems in ERTMS.

**Alessia Vennarini** is a Satellite Navigation Systems Engineer at Radiolabs. Her fields of research include the characterization and analysis of GNSS signals and software development for the use of GNSS data in tracking applications. She was involved in several European projects related to satellite navigation applied in railway transport systems.

## ABSTRACT

Recently, multi-sensor localization strategies are gaining attention in the railway scenario. In fact, the current trend is to reduce or remove the physical equipment deployed along the track for positioning purposes and to exploit on-board sensors to realize the same functionalities. Although GNSS is one of the major resources to perform this task, its performances dramatically decrease in presence of sources of local hazards like multipath, shadowing and blockage. For this reason, multi-sensor positioning methods are under study. Among them, those based on the detection of landmarks constituted by georeferenced trackside infrastructure elements like rail signs, and the estimation of the relative position of the train with respect to them are rather promising. Thus, in this paper we focus on the construction of the section of a Rail Digital Map related to these infrastructure elements on the basis of the fusion of the outputs of a stereo video camera and a LIDAR. In particular, the algorithms for object detection, single epoch landmark position estimation and landmark tracking are discussed. Results of the performance assessment based on Monte Carlo simulations are also reported.

## I. INTRODUCTION

Railways can play a significant role to achieve a smarter and more sustainable way of mobility. However, as the traffic grows, safe and accurate localization becomes crucial for railway traffic management, for example, for increasing the capacity by reducing the headway between trains, or for discriminating the track on which the train is operating, while guaranteeing safety. The current train localization systems are based on the use of odometers and balises. Balises are physical devices deployed along the track which allow the determination of the train's position when it passes over them. However, in order to satisfy the safety integrity requirements, a huge number of balises should be installed thus leading to increasing costs [1]. For this reason, the Global Navigation Satellite System (GNSS) technology has been classified as one of the game-changers for the evolution of the European Rail Traffic Management System (ERTMS)/European Train Control System (ETCS) and Command and Control System (CCS). The introduction of a GNSS-based functionally equivalent to the physical balises, i.e. the Virtual Balises, in fact, should allow a significant cost reduction. Moreover, the satellite-related assets should operate seamlessly with the current signalling standards thus ensuring end-to-end compatibility. However, in order to satisfy the strict accuracy and integrity navigation requirements, local hazards have to be mitigated. In fact, the railway environment presents its own particular challenges when it comes to satellite navigation, like urban canyons, tunnels or forested areas. To account for this issue, the restricted motion of the train can be exploited. More in detail, the train location can be constrained to the track path then increasing Position, Velocity and Time (PVT) accuracy and integrity. Nevertheless, combination of smoothed code pseudoranges with (differential) carrier phase and/or fusion with Inertial Measurement Unit (IMU) outputs are ineffective against the impairments originated by multipath low frequency components. Therefore, recently the integration of GNSS and IMU with on-board visual sensors has been addressed. For instance, the ESA NAVISP 2 project "VOLIERA - Video Odometry with LIDAR and EGNSS for ERTMS applications" [2] is currently investigating a multi-sensor positioning system for trains. More specifically, images, depth maps, and pointclouds are used to complement an IMU/GNSS localization module. Among the functionalities provided by the VOLIERA framework, the train absolute position is supplied. To achieve this goal, conspicuous points such as railway infrastructure elements are employed. More specifically, the train relative position with respect to the identified landmarks is computed, and the absolute location is obtained by exploiting the a-priori knowledge about landmark coordinates. However, to allow the deployment of positioning procedures based on the track constraint and on the rail landmarks, a database containing a digital map of the railway is essential. The availability in real-time of a reliable, authoritative map, in fact, would improve the reliability and integrity of the navigation system.

More specifically, the map should include both the track geometry and the elements of the rail infrastructure (e.g., panels, signals, signal gantries). Currently, building a database with the required quality and performance attributes is quite expensive, since conventional mapping techniques are laborious, and establishing an exhaustive map of all tracks in a particular railway network will require a huge effort of expert labour. In contrast, it is possible to aggregate navigation data collected through the repeated use of the track network by properly equipped trains. These aggregated data, correctly analysed and manipulated, will yield the mapping information as a side benefit from regular operations. More in general, the need for digital maps is part of the digitalization trend of the railways that began in the 90's. Additionally, in the field of signalling, mainly pushed by the ERTMS signalling system, the necessity for creating a "digital twin" platform has increased. It will be of paramount importance in test procedures and it will facilitate the detection and decision making process, on ERTMS/ETCS on-board and track-side subsystems, in order to ensure interoperability of the railway systems.

As a consequence, the availability of a high accuracy rail digital map represents one of the key points for supporting the process of European GNSS (EGNSS) uptake into rail signalling and control systems. To fill this gap, we present a technique based on the fusion of a stereo video camera and a Light Detection And Ranging (LIDAR) sensor for the construction of the section of a rail digital map related to the rail infrastructure elements. This approach could be beneficial for the development of projects such as RAILGAP (RAILway Ground truth and digital mAP) [3]. One of the main objectives of RAILGAP, in fact, is the definition of innovative and advanced methodologies and related tools for designing accurate and reliable digital maps of the railway environment relying on the processing of multi-sensor data. To this aim, RAILGAP foresees the exploitation of sensors such as IMUs, LIDARs and stereo cameras, all fused with Dual-Frequency, Multi-Constellation GNSS to improve the map accuracy in challenging environments (e.g., urban areas, tree canopies, etc.), thus extending the coverage of GNSS on rails.

In this work, along with a synthetic description of the railway digital map design, we focus on the computation of the railway infrastructure elements' position for georeferencing purposes. More in detail, we describe the procedures for computing the relative position of a landmark with respect to a train starting from the data acquired from the on-board visual sensors like a stereo camera and a LIDAR. This position can then be combined with the absolute position of the train obtained by post processing GNSS and IMU recorded data in order to georeference the landmark.

The remainder of the paper is organized as follows. In Section II the related works are briefly reviewed and in Section III the proposed method is described. More specifically, both the digital map design and the landmark position computation are detailed. Then, in Section IV the experimental results are provided and in Section V the conclusions are drawn.

## II. RELATED WORKS

In the literature, several studies have focused on the exploitation of multi-sensor positioning systems based on map matching. For instance, in [1], a train positioning system based on Inertial Navigation System (INS), odometer, GNSS and a digital map is proposed. More in details, once a coarse train position is available, the two nearest points in the track map are identified and the line passing through them is computed. The refined train position is then obtained as the perpendicular projection of the coarse position with respect to that line. Moreover, several researchers focused on landmark-based positioning systems. To do so, two types of approaches have been exploited. The former employs active beacons, whereas the second uses visual sensors in order to detect passive landmarks. In addition, two classes of passive landmarks exist: artificial and natural. The former consist in on-purpose built landmarks, whereas the second category includes elements that, despite being artificial objects, are naturally present in the localization environment (e.g., traffic signs, traffic lights, buildings). While the detection and identification of artificial landmarks is usually easier since their features are designed on purpose, the use of natural landmarks avoids the deployment of additional infrastructure.

Some examples of the use of active beacons are provided in [4] and [5]. More in details, in [4] WiFi-based landmarks are employed for indoor positioning. First of all, a landmark extraction is needed to build the landmark database, then, thanks to the obtained database, the user's position is computed based on a matching approach. On the other hand, in [5], an indoor localization procedure based on Bluetooth beacons is presented. As for the vision-based approaches, in [6] a landmark-based positioning system which exploits a single camera for warehouse positioning applications is proposed. To perform this task, artificial landmarks have been placed in known locations. Moreover, in [7], the authors propose a landmark-based positioning framework which exploits on-board LIDAR for the road environment. Concerning the railway scenario, in [8] the use of a set of Radio-Frequency IDentification (RFID) tags deployed along the railway has been investigated to realize a projection-based map matching positioning procedure. First of all, the RFID tags have to be deployed and their position has to be recorded. Then, when the reader installed on-board is close enough, the tag transmits a signal containing its identity. Given a train position estimate, the two closest map points are extracted and the initial position is projected along the track to obtain a fine estimate.

In order to allow for landmark-based positioning, a digital map containing all the georeferencing information is needed. Some examples are provided in [9] and [10]. More in details, Tschopp *et al.* in [9] propose a map building method for the railway environment based on the use of a Dynamic Vision Sensor (DVS) and on the exploitation of the Hough Transform to detect landmarks. The obtained results, although promising, show high false negative and false positive rates. In addition, in [10], a continuous map learning is proposed by building short term maps which are then integrated in a long-term map in the road scenario. Differently from the mentioned approaches, here we propose a railway mapping method based on the use of a stereo camera and a LIDAR which employs distance and angular measurements to compute the landmark coordinates.

## III. PROPOSED METHOD

### 1. Digital Map definition

A Rail Digital Map describes the railway network infrastructure. As such, it could be considered as an extension of the track topology and geometry 2D format by introducing additional information concerning infrastructure elements. The definition of the content of a Rail Digital Map constitutes the subject of the Rail Topo Model Expert Modelling Group (RTMEMG), a continuous working group of UIC, the International Union of Railways. In particular the standardization activities have been focused on the definition of the Topological data model (RailTopoModel) and on the Data exchange format (railML). The RailTopoModel is described in the IRS 30100 standard, [11], while detailed documentation on the railML markup language is available at [12] .

Traditionally, given a Rail Network, the track description of a line is defined by the track layout scheme, while the attributes of the trackside elements are contained in tables and characterized by unique identifiers and their location on the track. Each track is partitioned into track segments that, by definition, are sections of track without switches, where the train movement is unambiguous. In RailTopoModel, the topology of a Rail Network is described in term of a topological node edge model. As such, no information about distances and locations is available at this level. Connection of the topology objects to the physical word is carried out by the geometry. For instance, the geometry of a track segment consists of an ordered list of points with increasing mileage laying on the track centerline. For each point of the list, the Digital Map includes its position with respect to the linear reference system, defined as the distance (mileage) from the start of the element measured along the track, as well as its geographical coordinates. Geographical coordinates can be either Cartesian, or spherical (i.e., latitude, longitude, altitude). In general, the geographic reference system and the datum need to be unique within a given Digital Map. In addition to the list of points, a Digital Map should support the description of the track geometry bed in terms of horizontal and vertical geometry. To this aim, the horizontal geometry can be approximated by a sequence of segments of curves. Usually, three types of curves are employed: straight line, arc (with constant radius that is neither infinite nor zero), and transition curve. Transition curves may be used for connecting straight lines and/or circular arcs. Transition curves supported by the railML open-source markup language [12] include: sinusoid, doucine, curveWiener, curveBloss, cubicParabola, cosinusoid and clothoide. Thus, a compact description of the horizontal geometry can be obtained by resorting to the parameters defining the analytic expression of each curve segment. Vertical geometry is described in terms of changes on the slope of a track.

In addition to the network topology and to the track geometry, a Rail Digital Map may include information about a variety of railway relevant assets, addressed in the following as railway infrastructure elements, that can be found on, under, over or next to the railway track, like balises, platform edges, rail signs, and traffic lights. Attributes of these elements, including their geometry, can be stored in tables characterized by unique element identifiers. The geometry of infrastructure elements whose location can be represented as a single point of a map, (i.e., boards, poles, etc.) will include the linear and the geographic coordinates of their representative point. In this case, the coordinates with respect to the linear reference system will include, in addition to the mileage, the across track distance from the track centerline and the height with respect to the rail plane. On the contrary, infrastructure elements for which the physical size or the size of the region of interest is not negligible cannot be represented as single points on the map. For those applications for which a simplified description of their geometry can be considered sufficient we can store in the Digital Map the coordinates of a bounding box representing the volume occupied by the infrastructure element. Depending on the considered element, the bounding box may degenerate into a rectangle or a segment.

Considering that the best practices for surveys aimed at building the network topology and track geometry have already been established since a long time, in this paper , we focus our contribution on the theoretical background and on the algorithms used to build the section of the Digital Map related to the railway infrastructure elements. More specifically, the methodology behind the design and generation of this component of the Digital Map is based on:

- The post-processing of imaging and ranging sensors' records (stereo cameras and LIDAR) aimed at detecting the railway infrastructure elements and estimating their relative position with respect to the train.
- The post-processing of the GNSS records aimed at detecting the presence of local phenomena (multipath, electromagnetic interferences, etc.) that may impair the estimation of the location of the train, and estimating their temporal occurrence during a run.
- The combination of the train location derived from GNSS data with the object's relative position and the absolute (geographic) object's location based on previous runs stored in the Digital Map in order to set/update the absolute position of each infrastructure element. The process of building the Digital Map is incremental and the accuracy of the location of each Network Element improves by increasing the number of runs.

## 2. Landmark position estimation

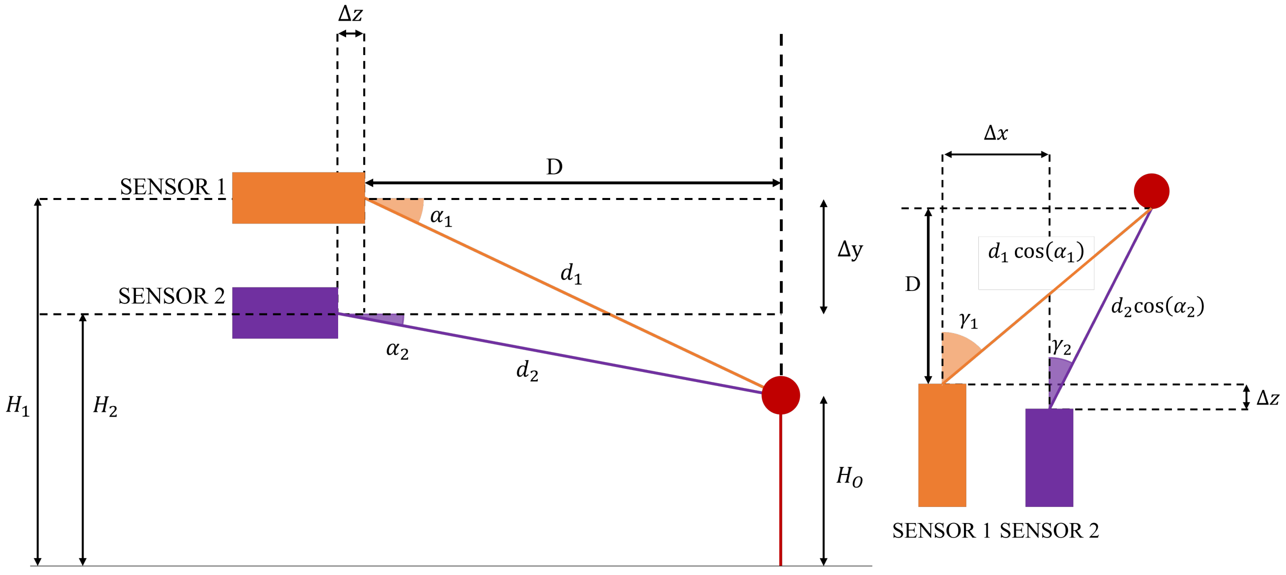
In this work, we deal with natural landmarks, so that no additional infrastructure has to be deployed along the track, and with computing their position by exploiting on board visual sensors. As described in [7], landmarks have to show at least the following properties: they have to be common objects, they have to be time-invariant both in terms of location and appearance, their observability has to be viewpoint invariant (i.e., their centroid computation has to be always possible), and their frequency of occurrence has to be high. To satisfy these requirements, in [7], the pole-like objects (i.e., traffic lights, street signs, streetlamps) are selected as primary landmarks. The same assumption holds for the analysis reported in the following. Moreover, in order to estimate the landmark position two substasks have to be performed: landmark detection and landmark coordinate extraction.

### *a). Landmark detection*

During the last years, several approaches have been proposed for object detection. In addition, due to the increasing interest in autonomous driving, many works deal with the detection of road infrastructure elements as well as road users and pedestrians. In this work, focusing on the railway scenario, we consider as suitable landmarks for position estimation the signalling equipment available along the track (e.g., traffic lights or traffic signs). To perform this task, deep learning methods, such as Convolutional Neural Networks (CNNs), have proven to be effective [13]. Some examples are the Fast R-CNN [14] and the Faster R-CNN [15]. The former extracts a feature map exploiting a Region of Interest (ROI) pooling layer to obtain the feature vectors from the proposed regions, and fully connected layers to produce softmax probabilities for each class and refined bounding box positions. The latter replaces the selective search algorithm, which is commonly employed for generating the region proposals, with a region proposal CNN. Typical object detection techniques provide as output a bounding box containing the detected objects. In order to estimate the landmark location in the image, however, a more accurate information is needed. To this aim, we decided to employ the Mask R-CNN [16] which is an extension of the Faster R-CNN architecture. More in details, a semantic segmentation branch is included for estimating the object masks inside the bounding boxes. The Mask R-CNN provides as output both the masks and the bounding boxes.

Since images usually enable an easier scene segmentation and understanding [17], landmark detection will be performed employing only the images. Then, the 2D bounding boxes provided by the detection algorithm are mapped to the corresponding 3D bounding boxes exploiting the depth information provided by the stereo camera. To do so, a registration step is required in order to represent the output provided by the different sensors in a single reference system. To this aim, first of all both sensors have to be calibrated. Park *et al.* proposed to use polygonal planar boards as targets that can be detected by both modalities to generate 3D-2D correspondences and obtain a more accurate calibration [18]. However, spatial targets make this method laborious for on-site calibration. As an alternative, Ishikawa *et al.* [19] devised an online calibration method without spatial targets which uses the odometry estimation of the sensors with respect to the environment to iteratively calibrate them. In order to obtain a registration method independent both from planar boards and odometry, we decided to employ a registration procedure similar to the one proposed in [20]. More specifically, De Silva *et al.* propose to align the outputs of a LIDAR and a monocular camera by performing a geometric transformation and a resolution matching procedure. In their work, the resolution matching was needed for obtaining a distance value for all the image pixels.

In our application, thanks to the depth map provided by the stereo camera, a distance value is already available for each image pixel. This allows to reduce the registration complexity. The goal of the geometric alignment is to find the transformation between the stereo camera and the pointcloud source reference systems. In the following, we indicate the main source as sensor 1, and the sensor which needs to be aligned as sensor 2. Let us consider an object $O$ of height $H_O$ at a depth $D$ from sensor's 1 origin. Let us call $\Delta x$, $\Delta z$ and $\Delta y$ the lateral, frontal and vertical displacements of the centers of sensor 1 and sensor 2, respectively. In addition, let $H_1$ and $H_2$ be the heights of sensor 1 and sensor 2, respectively, so that $\Delta y = H_1 - H_2$. Moreover, let $d_1$ be the distance measure of sensor 1, and $\alpha_1$ and $\gamma_1$ the latitude and longitude of object $O$ as measured by sensor 1. Similarly, let $d_2$, $\alpha_2$ and $\gamma_2$ be the same quantities with respect to sensor 2, and let us assume that the main axes of the two sensors are aligned with each other. The sensor geometry is shown in Figure 1, where a large displacement between the sensors has been employed for sake of clarity.



**(a)** Lateral view

**(b)** Top view.

**Figure 1:** Sensor geometry.

The distance $D$ can be expressed as follows

$$D = d_1 \cos\alpha_1 \cos\gamma_1 = d_2 \cos\alpha_2 \cos\gamma_2 + \Delta z. \tag{1}$$

Let us note that if the origin for computing $\Delta x$, $\Delta z$ and $\Delta y$ is set in sensor's 1 origin, in the case shown in Figure 1(b), $\Delta z$ is negative. Moreover, the object's height can be written as

$$H_O = H_1 + d_1 \sin\alpha_1 = H_2 + d_2 \sin\alpha_2, \tag{2}$$

where positive angles are measured in the upward direction. From the above equations we have

$$\tan\alpha_1 = \frac{((H_1 - H_2) + d_2 \sin\alpha_2)\cos\gamma_1}{d_2 \cos\alpha_2 \cos\gamma_2 + \Delta_z} = \frac{(\Delta y + d_2 \sin\alpha_2)\cos\gamma_1}{d_2 \cos\alpha_2 \cos\gamma_2 + \Delta_z}. \tag{3}$$

In addition, due to the horizontal displacement the following relation holds:

$$d_1 \cos\alpha_1 \sin\gamma_1 = d_2 \cos\alpha_2 \sin\gamma_2 + \Delta x. \tag{4}$$

As a consequence, we can write:

$$\tan\gamma_1 = \frac{d_2 \cos\alpha_2 \sin\gamma_2 + \Delta x}{d_2 \cos\alpha_2 \cos\gamma_2 + \Delta z}. \tag{5}$$

The equations of the latitude and longitude with respect sensor 1 allow to align sensor 2 reference system to the first one.

Once the pointcloud is represented in the stereo camera reference system, the 3D bounding box can be obtained by selecting the horizontal and vertical extensions of the 2D bounding box and defining a depth interval centered at the depth value of the object provided by the depth map.

*b). Single-epoch position estimation*

The landmark coordinates have to be provided with respect to an application dependent Conventional Terrestrial Reference System (CTRS), like the Earth-Centered, Earth-Fixed (ECEF). However, the position information provided by the IMU/GNSS module will be referred to the Antenna Reference Point (ARP), while measures extracted from the visual sensors are referred to the sensor's reference systems. As a consequence, a set of three transformations is needed in order to obtain the landmark positions:

1. a transformation from the sensor's reference system ($S$) to the train body reference system ($B$), associated to a rotation matrix $\mathbf{R}_S^B$ and a translation vector $\mathbf{t}_S^B$;
2. a transformation from the train body ($B$) system to the navigation reference system ($N$), associated to a rotation matrix $\mathbf{R}_B^N$ and a translation vector $\mathbf{t}_B^N$;
3. a transformation from the navigation reference system to the CTRS, associated to a rotation matrix $\mathbf{R}_N^C$ and a translation vector $\mathbf{t}_N^C$.

Given the landmark position in the sensor's reference frame, the coordinates in the body reference system will be:

$$\mathbf{P}_B = \mathbf{R}_S^B \mathbf{P}_S + \mathbf{t}_S^B, \tag{6}$$

where $\mathbf{P}_S$ is the coordinate vector in the sensor's reference system and $\mathbf{P}_B$ is the corresponding vector in the body reference system. Then, the coordinates in the navigation reference system can be obtained as

$$\mathbf{P}_N = \mathbf{R}_B^N \mathbf{P}_B + \mathbf{t}_B^N = \mathbf{R}_B^N \mathbf{R}_S^B \mathbf{P}_S + \mathbf{R}_B^N \mathbf{t}_S^B + \mathbf{t}_B^N. \tag{7}$$

Consequently, the position with respect to the CTRS is:

$$\mathbf{P}_C = \mathbf{R}_N^C \mathbf{P}_N + \mathbf{t}_N^C = \mathbf{R}_N^C \mathbf{R}_B^N \mathbf{R}_S^B \mathbf{P}_S + \mathbf{R}_N^C \mathbf{R}_B^N \mathbf{t}_S^B + \mathbf{R}_N^C \mathbf{t}_B^N + \mathbf{t}_N^C. \tag{8}$$

In order to describe the landmark position estimation in the sensor's frame, let us consider a reference system aligned to the one of the on-board sensor. When more than one sensor is present, one of them can be selected independently. More in details, the reference system $\{\mathbf{P}, \mathbf{e}_X, \mathbf{e}_Z, \mathbf{e}_Y\}$, can be defined as follows

- $\mathbf{e}_X$ is the unit vector whose direction is orthogonal with respect to the direction of motion;
- $\mathbf{e}_Z$ is the unit vector whose direction coincides with the direction of motion;
- $\mathbf{e}_Y$ is the unit vector pointing up given by the cross product of $\mathbf{e}_X$ and $\mathbf{e}_Z$;

and $\mathbf{P}$ coincides with the sensor position. In the following, without loss of generality, we assume that the selected reference system is the one of the stereo camera sensor.

In order to compute the landmark location, two types of measurements are needed: angle and distance estimation. Concerning the angle estimation, the image provided by one of the camera sensors of the stereo camera can be employed. Given the Field of View (FOV) of the sensor, it is possible to associate to each pixel the corresponding angular extension. Therefore, by computing the number of pixels between the optical axis of the camera and the Line of Sight (LoS) between camera and landmark, the corresponding angular measurement is obtained. To estimate the LoS, we defined it as the segment whose end-points are the camera's optical center and the landmark centroid. In addition, we estimated this last as the centroid of the portion corresponding to the landmark in the masks provided by the landmark detection algorithm.

As for the distance measurement, both the pointcloud provided by the LIDAR and the depth map provided by the stereo camera can be employed. More in details, the distance value for an object can be obtained as the mean value of the pixels associated to the object in the depth map. As for the pointlcoud, the distance value can be computed as the mean distance of the points falling inside the 3D bounding box. In addition, to provide a robust estimation of the distance value, the two quantities can be averaged by employing a weighting procedure based on the sensor confidence. The two types of sensors, in fact, have different ideal operating conditions. Therefore, depending on the specific scenario, the depth values provided by the stereo camera and the LIDAR can be weighted differently.

Given the angle and the distance, the landmark's position can be computed as:

$$\begin{cases} x_L = d_1 \cos\alpha_1 \sin\gamma_1 \\ z_L = d_1 \cos\alpha_1 \cos\gamma_1 \\ y_L = d_1 \sin\alpha_1 \end{cases} . \tag{9}$$

The graphical representation of the landmark coordinates computation is provided in Figure 2.

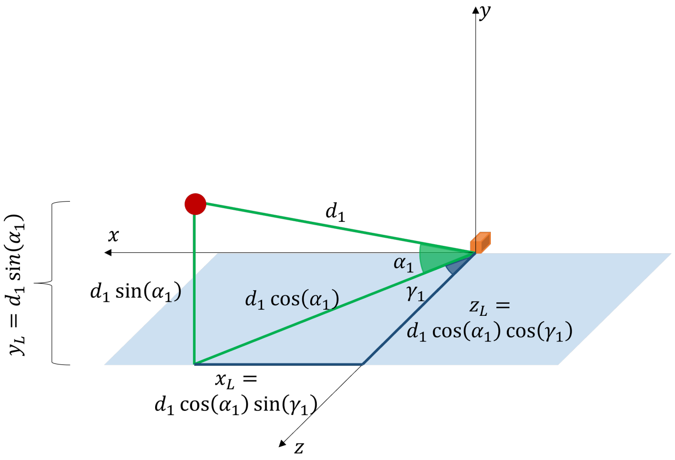

y
x
z
$d_1$
$\alpha_1$
$\gamma_1$
$y_L = d_1\, sin(\alpha_1)$
$d_1\, \sin(\alpha_1)$
$d_1\, \cos(\alpha_1)$
$z_L =$
$d_1 \cos(\alpha_1) \cos(\gamma_1)$
$x_L =$
$d_1 \cos(\alpha_1) \sin(\gamma_1)$

**Figure 2:** Landmark position computation.

*c). Landmark tracking*

In order to properly handle occlusions and false positives, the positions of the landmarks' projection on the image planes are tracked over time using a multiple-object tracking algorithm, which creates, prunes, and updates the tracks. Therefore, the output of the landmark tracking block is a set of temporally-consistent tracks, which are updated using the detections from the landmark detection.

The algorithm maintains two sets of tracks: the set of tentative tracks $\mathcal{T}$, where newly created tracks characterized by higher uncertainty are inserted, and the set of confirmed tracks $\mathcal{C}$, which contains the tracks that have been assigned a sufficient number of detections for them to be considered more probable. Each track is described by a state $\boldsymbol{\upsilon}$, a covariance matrix $\boldsymbol{\Upsilon}$, and by the subset of detections which have been assigned to it. For each track, an incremental counter $\eta$, representing the *lifetime* (i.e., the number of frames elapsed since the creation of the track) and a visibility score $\zeta$, which is given by the fraction of frames containing a measurement over the total $\eta$, are also kept. For each track, a Kalman filter makes the prediction from the previous step to the current step and uses the detections to correct that prediction.

New tracks are initialized with state $\boldsymbol{\upsilon} = [u, v, w, h, vu, vv, vw, vh]^T$. The first part of the state, namely $u$, $v$, $w$, and $h$, denotes the information on the bounding box associated to the detection that led to the creation of the new track, while the second part, namely $vu$, $vv$, $vw$, and $vh$, represents the velocity components. For the former, which can be measured directly, the covariance matrix $\boldsymbol{\Upsilon}$ is initialized as

$$\boldsymbol{\Upsilon} = \text{diag}\left(\sigma_u^2, \sigma_v^2, \sigma_w^2, \sigma_h^2\right), \tag{10}$$

while for the rest a large initial covariance is assumed.

Kalman filter state estimates and covariances predicted by the $k^{th}$ tracker from the previous time step $t-1$ to the current time step $t$, denoted by $\hat{\boldsymbol{\upsilon}}_k^-$ and $\hat{\boldsymbol{\Upsilon}}_k^-$ respectively, are computed as follows:

$$\hat{\boldsymbol{\upsilon}}_k^- = \boldsymbol{\Phi}\hat{\boldsymbol{\upsilon}}_k^+, \tag{11}$$

$$\hat{\boldsymbol{\Upsilon}}_k^- = \boldsymbol{\Phi}\hat{\boldsymbol{\Upsilon}}_k^+\boldsymbol{\Phi}^T + \mathbf{Q}, \tag{12}$$

where

$$\boldsymbol{\Phi} = \begin{bmatrix} \mathbf{I}_4 & \delta t\mathbf{I}_4 \\ \mathbf{0}_4 & \mathbf{I}_4 \end{bmatrix} \tag{13}$$

is the transition matrix for a constant velocity model, and $\mathbf{Q}$ is the system noise covariance matrix. For sake of compactness, the Kalman filter matrices are expressed in terms of submatrices corresponding to the vector components of the state (i.e., position and velocity). $\mathbf{I}_n$ denotes the identity matrix of size $n$, while $\mathbf{0}_n$ denotes the square null matrix of size $n$.

From the output of the landmark detection network, the set of bounding boxes $\{\mathbf{d}\}_i$ (i.e., the position of the upper left corner, the width, and the height of each box) are extracted. The corrected states $\hat{\boldsymbol{\upsilon}}_k^+$ and covariance matrices $\hat{\boldsymbol{\Upsilon}}_k^+$ at time step $t$ updated using the assigned measurements are given by

$$\hat{\boldsymbol{\upsilon}}_k^+ = \hat{\boldsymbol{\upsilon}}_k^- + \mathbf{K}_k(\mathbf{d} - \mathbf{H}\hat{\boldsymbol{\upsilon}}_k^-), \tag{14}$$

$$\hat{\boldsymbol{\Upsilon}}_k^+ = (\mathbf{I}_8 - \mathbf{K}_k\mathbf{H})\hat{\boldsymbol{\Upsilon}}_k^- \tag{15}$$

where

$$\mathbf{H} = \left[\mathbf{I}_4 \quad \mathbf{0}_4\right], \tag{16}$$

$$\mathbf{K}_k = \hat{\boldsymbol{\Upsilon}}_k^-\mathbf{H}^T(\mathbf{H}\hat{\boldsymbol{\Upsilon}}_k^-\mathbf{H}^T + \mathbf{R})^{-1}, \tag{17}$$

and $\mathbf{R}$ is the covariance matrix of the measurement noise.

The algorithm that oversees the tracks performs the following steps:

- the state and covariance of all the tracks are updated using (11), (12);
- detections are assigned to existing tracks using Munkres's variant of the Hungarian algorithm, minimizing the normalized distances between the $i^{th}$ detection $\mathbf{d}_i$ and the predicted measurement $\hat{\mathbf{d}}_k = \mathbf{H}\hat{\boldsymbol{\upsilon}}_k^-$ computed by the $k^{th}$ tracking filter, weighted by the residual covariance $\mathbf{S}_k$ and incremented by a penalty term $C$ for assignments with different classes:

  $$\chi_{i,k} = (\mathbf{d}_i - \hat{\mathbf{d}}_k)^T\mathbf{S}_k^{-1}(\mathbf{d}_i - \hat{\mathbf{d}}_k) + \log(|\mathbf{S}_k|) + (1 - \delta_{c_i,c_j})C, \tag{18}$$

  where

  $$\mathbf{S}_k = \mathbf{H}\hat{\boldsymbol{\Upsilon}}_k^-\mathbf{H}^T + \mathbf{R}, \tag{19}$$

  and

  $$\delta_{c_i,c_j} = \begin{cases} 1 & \text{if } c_i = c_j \\ 0 & \text{if } c_i \neq c_j \end{cases}. \tag{20}$$

  By setting a threshold over $\chi_{i,k}$, a validation gate around the track is formed, and only the detections falling within the gate are considered as candidates for the assignment;
- new tentative track are created from unassigned detections;
- tracks in $\mathcal{T}$ are confirmed if at least $t_{conf}$ detections are assigned to them, and are moved from to $\mathcal{C}$;
- tracks whose visibility falls below a threshold $t_{del}$ are deleted.

## IV. EXPERIMENTAL RESULTS

In order to evaluate the performances of the landmark positioning procedure, we considered the experimental scenario depicted in Figure 3. More specifically, the train is assumed to move on a straight line along the $z$ axis, going from 0 to 20 m. In addition, three landmarks have been considered: two lateral landmarks and a frontal landmark situated above the railway. As for the vertical coordinates, in compliance with the user requirements specified in the VOLIERA project concerning on-board sensor setup and measurement procedures for railway objects detection, we assumed a sensor installation height of 3.7 m, and we considered heights equal to 3.5 m and 5 m for the lateral and frontal landmarks, respectively.

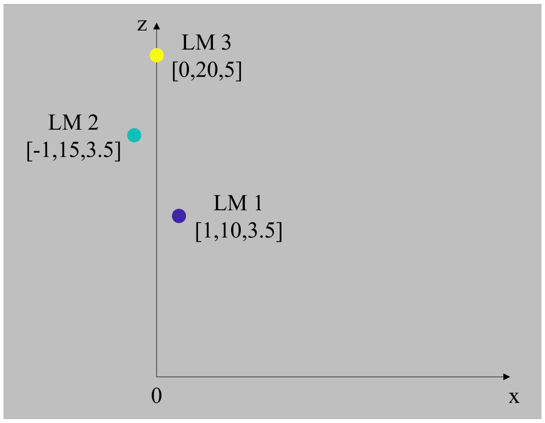


**Figure 3:** Landmark configuration for the performance assessment.

**Table 1:** Nerian calculator parameters.

| | |
|---|---|
| Sensor size | 2/3” |
| Sensor resolution | $2432 \times 2048$ |
| Focal length | 8 mm |
| Baseline | 25 cm |

For the performance assessment we employed a Monte Carlo simulation. We modelled the measurement errors by considering state of the art sensors that can be employed for building the Digital Map and the results obtained in our previous work on angular-based positioning [21] for the angular measurement. More in details, for the latter, we considered a standard deviation equal to $0.19°$. As for the distance measurement, we exploited the publicly available information concerning the Livox Horizon [22] and the Neuvition Titan M1-Pro [23] LIDARs. For both sensors, a 2 cm standard deviation is claimed. In addition, a minimum depth value of 1 m is reported for the Neuvition Titan M1-Pro, while this information is not available for the Livox Horizon.

**Table 2:** Nerian calculator outputs.

| | |
|---|---|
| Horizontal FOV | $55.5°$ |
| Vertical FOV | $47.8°$ |
| Minimum depth | 1.13 m |

Concerning the stereo camera, we considered the Nerian Scarlet 3D depth camera [24]. More in details, we employed the camera parameters shown in Table 1 in the online performance calculator provided by Nerian at [25]. The results provided by the calculator are shown in Table 2 and the obtained distance error is provided in Figure 4.

Considering the obtained horizontal FOV, and the lateral distance between the track and the landmarks, a minimum distance of 1.9 m is required to ensure landmark visibility. This value is compatible with the minimum depth which can be measured both by the stereo camera and by the LIDAR so that a minimum distance equal to 2 m has been considered in the simulations.

Furthermore, for the distance measurement we employed a weighted average of the values provided by the LIDAR and the stereo camera. To do so, we defined the weights according to the standard deviations as follows:

$$\begin{cases} w_L = \frac{\sigma_S}{\sigma_L + \sigma_S} \\ w_S = \frac{\sigma_L}{\sigma_L + \sigma_S}, \end{cases} \tag{21}$$

where $\sigma_L$ is the LIDAR standard deviation which is assumed to be constant and equal to 2 cm according to the information reported in [22, 23], and $\sigma_S$ is the stereo standard deviation which has been approximated for every train position according to Figure 4.

The obtained performances are shown in Figure 5, where the results for $z$ values between 0 and the selected minimum distance from the landmark are provided. The Monte Carlo simulation has been performed considering 10 000 runs and a candidate train position every 0.2 m between 0 and 20 m along $z$. The results shown in Figure 5 represent the overall error standard deviation

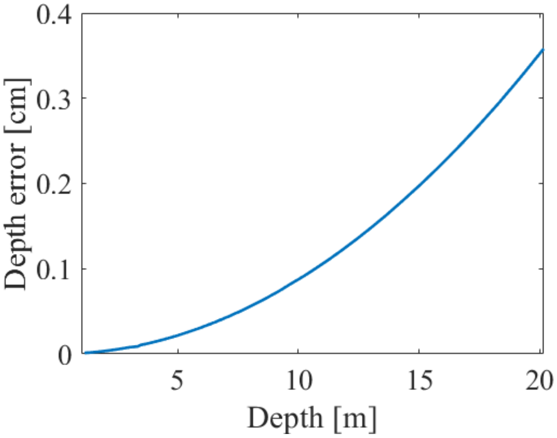


**Figure 4:** Landmark configuration for the performance assessment.

obtained as $\sqrt{\sigma_x^2 + \sigma_z^2 + \sigma_y^2}$. As clearly shown from the figure, the error decreases when the train gets closer to the landmark, reaching a minimum value of 1.27 cm. Additionally, the maximum obtained error equals 10.15 cm, thus demonstrating the suitability of the proposed framework to build the Digital Map.

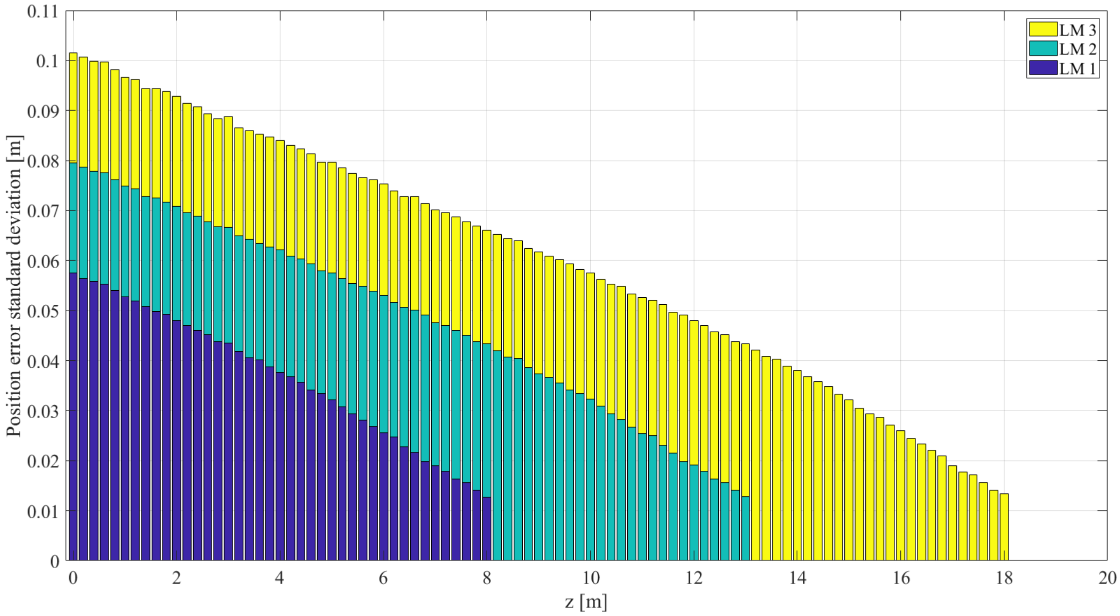


**Figure 5:** Landmark positioning performances.

## V. CONCLUSIONS

In this work, we proposed a methodological approach for the design of a high accuracy digital map of the railway environment, to be used for positioning purposes. The elements constituting the topology of the digital map have been defined, and a sensor fusion technique for determining the position of railway infrastructure elements has been developed. The experimental results prove the applicability of the proposed framework. More in details, a Monte Carlo simulation has been performed considering three landmark positions, and an error standard deviation between 1.27 and 10.15 cm has been obtained.

## REFERENCES

[1] W. Jiang, S. Chen, B. Cai, J. Wang, W. ShangGuan, and C. Rizos, “A Multi-Sensor Positioning Method-Based Train Localization System for Low Density Line,” *IEEE Transactions on Vehicular Technology*, vol. 67, no. 11, pp. 10 425–10 437, 2018.

[2] “VOLIERA Project,” [Available at: https://voliera.eu/ - accessed 03-November-2021].

[3] “RAILGAP Project,” [Available at: https://railgap.eu/ - accessed 03-November-2021].

[4] M. Youssef, P. Robertson, H. Abdelnasir, M. Puyol, E. Le Grand, and L. Bruno, “Lighthouse: Enabling Landmark-Based Accurate and Robust Next Generation Indoor LBSs on a Worldwide Scale,” in *2019 20th IEEE International Conference on Mobile Data Management (MDM)*, 2019, pp. 8–17.

[5] B. Jang, H. Kim, and J. W. Kim, “IPSCL: An Accurate Indoor Positioning Algorithm Using Sensors and Crowdsourced Landmarks,” *Sensors*, vol. 19, no. 13, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/13/2891

[6] Y. Y. Yap and B. E. Khoo, “Landmark-Based Automated Guided Vehicle Localization Algorithm for Warehouse Application,” in *Proceedings of the 2019 2nd International Conference on Electronics and Electrical Engineering Technology*, ser. EEET 2019. New York, NY, USA: Association for Computing Machinery, 2019, p. 47–54. [Online]. Available: https://doi.org/10.1145/3362752.3365285

[7] J. Gim, C. Ahn, and H. Peng, “Landmark Attribute Analysis for a High-Precision Landmark-Based Local Positioning System,” *IEEE Access*, vol. 9, pp. 18 061–18 071, 2021.

[8] K. Kim, S. Seol, and S.-H. Kong, “High-speed Train Navigation System based on Multi-sensor Data Fusion and Map Matching Algorithm,” *International Journal of Control, Automation and Systems*, vol. 13, 06 2015.

[9] F. Tschopp, C. von Einem, A. Cramariuc, D. Hug, A. W. Palmer, R. Siegwart, M. Chli, and J. Nieto, “Hough$^2$Map – Iterative Event-Based Hough Transform for High-Speed Railway Mapping,” *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2745–2752, 2021.

[10] M. Stübler, S. Reuter, and K. Dietmayer, “A Continuously Learning Feature-based Map using a Bernoulli Filtering Approach,” in *2017 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*, 2017, pp. 1–6.

[11] “RAILTOPOMODEL,” [Available at: https://www.railtopomodel.org/en/ - accessed 03-November-2021].

[12] “RAILML,” [Available at: www.railml.org - accessed 03-November-2021].

[13] M. Weber, P. Wolf, and J. M. Zöllner, “DeepTLR: A single deep convolutional network for detection and classification of traffic lights,” in *2016 IEEE Intelligent Vehicles Symposium (IV)*, 2016, pp. 342–348.

[14] R. Girshick, “Fast R-CNN,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.

[15] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[17] L. Heng, B. Choi, Z. Cui, M. Geppert, S. Hu, B. Kuan, P. Liu, R. Nguyen, Y. C. Yeo, A. Geiger *et al.*, “Project autovision: Localization and 3d scene perception for an autonomous vehicle with a multi-camera system,” in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4695–4702.

[18] Y. Park, S. Yun, C. S. Won, K. Cho, K. Um, and S. Sim, “Calibration between Color Camera and 3D LIDAR Instruments with a Polygonal Planar Board,” *Sensors*, vol. 14, no. 3, pp. 5333–5353, 2014. [Online]. Available: https://www.mdpi.com/1424-8220/14/3/5333

[19] R. Ishikawa, T. Oishi, and K. Ikeuchi, “LiDAR and Camera Calibration Using Motions Estimated by Sensor Fusion Odometry,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 7342–7349.

[20] V. De Silva, J. Roche, and A. Kondoz, “Robust Fusion of LiDAR and Wide-Angle Camera Data for Autonomous Mobile Robots,” *Sensors*, vol. 18, no. 8, 2018. [Online]. Available: https://www.mdpi.com/1424-8220/18/8/2730

[21] S. Baldoni, F. Battisti, M. Brizzi, and A. Neri, “A Hybrid Position Estimation Framework based on GNSS and Visual Sensor Fusion,” in *2020 IEEE/ION Position, Location and Navigation Symposium (PLANS)*, 2020, pp. 979–986.

[22] “Livox Horizon LIDAR,” [Available at: https://www.livoxtech.com/horizon - accessed 24-November-2021].

[23] “Neuvition Titan M1-Pro,” [Available at: https://www.neuvition.com/products/titan-m1-pro.html - accessed 24-November-2021].

[24] “Nerian Scarlet 3D depth camera,” [Available at: https://nerian.com/products/scarlet-3d-depth-camera/ - accessed 24-November-2021].

[25] “Nerian online calculator,” [Available at: https://nerian.com/support/calculator/ - accessed 24-November-2021].