Modeling evidential grids using semantic context information for dynamic scene perception

Giovani Bernardes Vitor, Alessandro Corrêa Victorino, Janito Vaqueiro Ferreira

Modeling Evidential Grids using Semantic Context information for Dynamic Scene Perception

Giovani Bernardes Vitor[a,*], Alessandro Corrêa Victorino[b], Janito Vaqueiro Ferreira[c]

[a] *Federal University of Itajubá, Institute of Technological Sciences - ICT/UNIFEI, Laboratory of Robotics, Intelligent and Complex Systems - RobSIC, 35903-087 Itabira - MG, Brazil*

[b] *Federal University of Minas Gerais, School of Engineering - UFMG Department of Mechanical Engineering, 31270-901 Belo Horizonte - MG, Brazil*

[c] *State University of Campinas, School of Mechanical Engineering - FEM/UNICAMP Autonomous Mobility Laboratory - LMA, 13083-860 Campinas - SP, Brazil*

Abstract

Uncertainty about urban environments stems not only from imprecise pose estimation and noisy information in images but also from the lack of semantic information. This article presents an approach to improve the perception capability of intelligent vehicles in complex urban environments. The new method uses the meta-knowledge extracted from semantic context images associated with depth information to model occupancy grids from stereo vision. It uses the evidential formalism of the Dempster-Shafer theory to manage uncertainties involved in grid discretization, partial observation of the environment and also dynamic elements present in the scene. Real experiments carried out in a challenging urban environment using the KITTI benchmark are reported, from which meaningful evaluations can be made to illustrate the validity and applicability of this approach.

Keywords: Dynamic Peception, Semantic Context Image, evidential grids, belief function theory, contextual fusion, intelligent vehicles

*Principal corresponding author

Email addresses: giovanibernardes@unifei.edu.br (Giovani Bernardes Vitor), alessvict@gmail.com (Alessandro Corrêa Victorino), janito@fem.unicamp.br (Janito Vaqueiro Ferreira)

1. Introduction

The development of autonomous vehicles capable of getting around on urban roads can provide important benefits in reducing accidents, increasing life comfort and providing cost savings. For example, intelligent vehicles often base their decisions on observations obtained from various sensors such as LIDAR, GPS and cameras. Attention is currently focusing increasingly on camera sensors because they are inexpensive, easy to use and provide rich data. Inner-city environments represent an interesting but also very challenging scenario in this context, since the road layout may be highly complex, the presence of obstacles such as trees, bicycles and cars may generate partial observations, and also because these observations are often noisy or even missing due to significant obstructions. Thus, the perception process must be able to deal with uncertainties about the world surrounding the car. While autonomous navigation on highways using prior knowledge of the environment has advanced significantly, understanding and navigating in general inner-city scenarios based on little prior knowledge remains an unsolved problem.

In this context, a dynamic local perception system is developed to build the representative model of the environment around the car. The metric representation based on occupancy grid mapping is implemented. This representation uses the evidential formalism proposed by the Dempster-Shafer theory, which has been receiving considerable attention. This formalism allows one to manage uncertainties associated with grid discretization, partial observation of the environment, and also dynamic elements of the scene.

This paper contributes by proposing a new strategy based on the fusion of semantic context information extracted from the machine learning procedure and depth information obtained from epipolar geometry to build a local perception that uses meta-knowledge to influence the formulation of the belief mass in the evidential grid, directly considering semantic, dynamic and uncertainty aspects in the representation.

The article is organized as follows. Section 2 presents an overview of related works pertaining to the scope of this work, positioning the proposed approach vis-à-vis previous ones. The proposed method is outlined in Section 3, which describes the conception of the newly developed system. Section 4 characterizes

the inclusion of semantic context information obtained from imagens into the evidential grid, explaining the inverse sensor model considering meta-knowledge and also the updating of temporal fusion. Section 5 presents the experimental results of the dynamic local perception. Finally, Section 6 presents our conclusions and future prospects.

2. Related Works

This work focuses on modeling the local environment during the vehicles navigation and movement, using visual strategies. It is denoted by mapping in the context of Simultaneous Localization and Mapping (SLAM). Although most of the environmental representations are metric, some approaches also use 2D and 3D topological representations that have produced outstanding results, such as those reported by Meilland et al. [1]. This paper presents the main modes of geometric representation and their use in existing perception systems, which are divided into feature-based and grid-based approaches.

Feature-based approach: This method uses geometric features to represent the environment. The type of feature employed depends on the target application, the environment in question, the required accuracy and the computational power. Birds-eye view modeling is employed in many cases [2]. As explained, this method depends on feature extraction and matching to ensure the consistency of the mapping at each moment of time.

Many feature-based SLAM systems, such as that proposed by Montemerlo et al. [3], use a representation of the environment based on natural features. In this type of approach, the environment is represented by a state vector containing the coordinates of these landmarks. The state vector is filtered over time using a Kalman filter [4] or particle filter [5]. The upgrade process between detections consists of the aforementioned problem of data association, which is processed in different ways [6, 7]. In other systems, this representation pertains to obstacle detection using 2D or 3D shape-models, and depends on the intended application. Petrovskaya and Thrun [8] and Fayad and Cherfaoui [9] are interested not only in vehicle detection using laser sensors, in which objects are modeled as rectangular boxes, but also in performing tracking over time. Therefore, as can be seen, the problem with SLAM is that it considers the key-

points as fixed. To avoid errors caused by the inclusion of moving entities, the SLAMMOT algorithm, as presented by Lin and Wang [10], is proposed as a way to improve the mapping, avoiding the use of these keypoints in the localization process. According to Moras [2], some of the pros and cons of feature-based representation are as follows:

- Advantages:

    - Simple representation;
    - Easy propagation over time;
    - Mobile objects are considered;
    - Low memory consumption.

- Disadvantages:

    - Non-exhaustive representation, and therefore, inadequate for navigation;
    - Very high sensitivity to the results of the matching process;
    - Lack of precision for safe local autonomous navigation tasks.

Grid-based approach: In this approach, the environment is modeled as a grid of cells with no parametric object representation, in which each cell contains information indicating whether or not the given associated portion of the environment is occupied [11]. The occupancy state of each cell is evaluated independently. The update process considers all the modeled cells of the grid. In general, the cells are square, but some works, such as that of Herrmann et al. [12], consider grids with different geometries.

Works on occupancy grids using a 2D grid to build and update the map of the environment were first proposed by Elfes [11, 13, 14]. Initially limited by its computational complexity, this approach has recently been widely used for navigation. In the works of Bourgault et al. [15], Thrun et al. [16], Steux and El Hamzaoui [17], and Levinson and Thrun [18], the authors have modeled a fixed grid that allows the position to be corrected based on each new measurement. Coué et al. [19] propose mobile object tracking using a vehicle-to-grid reference. The update process is performed considering the occupation

and also the speed vector of each cell. The object is then regrouped considering the occupation, speed vector and position of the cells. Gate [20] uses a grid to perform the SLAMMOT. Alternatively, some studies consider the 3D space in which the grid is represented as a cube [21], or else define the grid as a QuadTree [22], aiming to reduce the memory space and the calculation for homogeneous areas. Basically, the grid-based representation approach has the following characteristics [2]:

- Advantages:

    - A comprehensive representation that enables precise local autonomous navigation tasks to be performed;
    - No assumptions about the geometry of the elements in the environment.

- Disadvantages:

    - It has a complex propagation of model, motion and perception uncertainties over time;
    - It is difficult to consider moving objects;
    - Its computational cost and memory use are considerable.

Most of the earlier studies have used the probabilistic model to represent occupancy uncertainties in the grid. According Moras [2], how these uncertainties are represented has important implications on how the information contained in the grids is processed. In this sense, in addition to the probabilistic model, the literature presents two other approaches involving accumulation methods and evidential methods. For completeness, a brief overview of these three methodologies is given here.

The formalism of accumulation is quite simple, based on the principle of voting: the more occupied the cell appears to be the more likely it is to be occupied. Albeit rarely used, it is possible to find original contributions using this formalism. Borenstein and Koren [23] use accumulation grids for the navigation of an experimental indoor robot equipped with sonar. Xie et al. [22] employ accumulation grids to map an external environment using a scanning

laser rangefinder. Online localization is ensured by grid matching using multiresolution grids (QuadTree).

The probabilistic approach is based on Bayes' theory [24] and is the one most widely used in the field of robotics. It was the first formalism of uncertainty management used in occupancy grids. The first author who proposed this scheme was Elfes [11, 13]. This type of approach defines the state of a cell based on two mutually exclusive possibilities, occupied O or free F. Each cell of the occupancy grid contains a probability of occupancy $P(O)$ and/or non-occupancy $P(F)$, and all the cells are assumed to be independent of each other. Different formulations exist, using either a direct sensor model or an inverse sensor model [25], and also considering static or dynamic environments, as explained earlier.

The third method, the evidential approach, derives from the Dempster-Shafer theory and the Transferable Belief Model (TBM) [26, 27, 28], which is a generalization of probabilities. The underlying problem of all grid-based approaches has to do with the conflicts generated by sensing the presence of moving objects in the scene [29, 30]. The approaches proposed by Moras et al. [31], Kurdej et al. [32, 33] have presented satisfactory results using heuristics that combine several sources of information. However, these approaches consider the following hypotheses: (i) They are restricted to places for which prior digital map information of the environment must be available. (ii) The precise pose estimation of the ego-car must also be supplied (using differential GPS) in order to combine and update the evidential grid. (iii) The perception system is not able to accurately distinguish the feasible navigable area in urban scenarios, i.e., the street area.

A reliable perception with the annotation of relevant objects could be used as an alternative source to improve safety in urban scenarios. In general, the vision-based approaches to perception cited in the preceding paragraphs reveal the lack of ability to annotate the environment with semantic information and maintain a satisfactory level of precision. Based on previous works and considering these assumptions, which are required to perform mapping, the methodology proposed in this article uses the evidential method to deal with uncertainties in a grid-based approach. This grid is built online, performing

local ego-centered mapping to avoid drift errors and associating the Semantic Context which associates automatically collected meta-knowledge in the grid, providing the required accuracy for an application in autonomous navigation. This approach enables the uncertainties of different entities in a complex urban scenario to be managed using only a pair of stereo cameras.

3. Overview of the Proposed System

This article, as part of the field of visual perception for car-like robots or so-called intelligent vehicles, pertains to the field of mobile robotics and computer vision. Following these multidisciplinary domains that are merged to reach desired outcomes, it proposes a solution by means of the system depicted in Figure 1. As can be seen, a set of tasks have been designed to accomplish an appropriate perception scheme. The solution approach is divided into two main tasks, defined as (I) Semantic Context and (II) Dynamic Evidential Grid.

The task of the (I) Semantic Context is to understand urban road scenes. In this layer, two modules are reported that produce meta-knowledge from a pair of stereo images. The first module, Disparity Map, is used to obtain the map os disparities from the pair of stereo images captured from cameras in front of the intelligent vehicle.

The task of the (I) Semantic Context is to understand urban road scenes. In this layer, two modules are reported that produce meta-knowledge from a pair of stereo images. The first module, Disparity Map, is used to obtain the map os disparities from the pair of stereo images captured from cameras in front of the intelligent vehicle. The Machine Learning module receives the pair of stereo images to be classified in an implemented classifier. Using this structure, the principle to obtain meta-knowledge is employed to understand the semantic urban road scene. The output result of the (I) Semantic Context task is illustrated in Figure 1(a). It should be highlighted that the meta-knowledge presented by the color image is not just a superpixel/segments obtained from image processing. Each color represents an estimated object class addressed by the output of an implemented machine learning algorithm.

The purpose of the (II) Dynamic Evidential Grid task is to perform the local perception mapping and characterization of static and moving obstacles, using

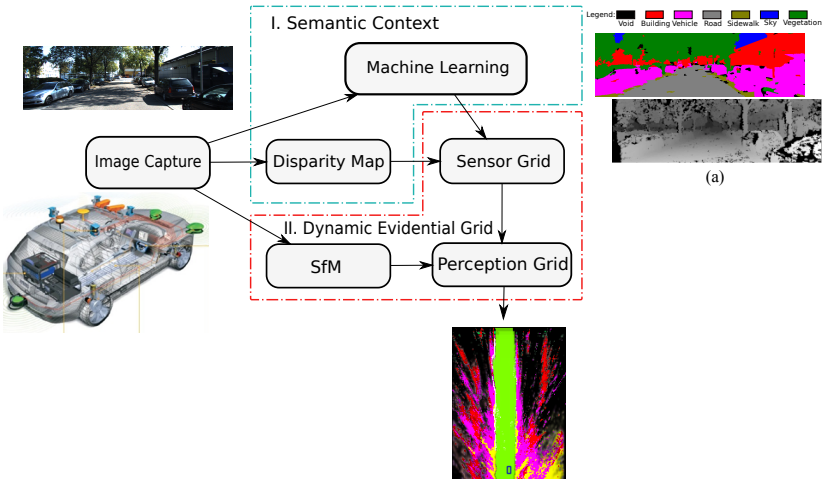

Figure 1: Proposed solution for Dynamic Scene Perception using only a pair of stereo camera sensors.

the output responses elicited by the layer (I). This layer comprises three modules that are employed to model a dynamic local occupancy grid, applying the Dempster-Shafer theory. The Structure from Motion (SfM) module is applied to achieve rigid transformation between two successive pairs of images. The Sensor Grid module builds a novel inverse sensor model that projects 3D points obtained from the disparity map onto a metric grid, taking into account the noise in the stereo measurements and the uncertainty linked with stereo geometry reconstruction, where exponential error is observed with increasing distance. Furthermore, the meta-knowledge extracted from urban road scene understanding is associated with this proposed inverse sensor model, in which it provides a better and reliable representativeness of navigable, infrastructure and obstacles areas. After that, the Perception Grid module performs the temporal fusion and mobile cell detection. The output result of this layer (II) is illustrated in Figure 1(b) and described in detail in Section 4, in which concentrates the main contributions of the paper.

4. Dynamic Local Perception using Evidential Grids

This section describes the approach employed to deal with local perception mapping, relative localization and characterization of static and moving obsta-

cles, using stereo vision cameras.

Unlike other works, this approach does not require prior digital mapping information, pose estimation, or vehicle tracking. As mentioned in Section 3 and illustrated in Figure 1, the main contributions of this section are the proposed technique to build a new sensor model that provides reliable urban environment sensing, notwithstanding uncertainties in distance measurements associated with the epipolar geometry of stereo vision and the combination/update rules for meta-knowledge that characterize the semantic context in evidential grids. Hence, the new sensor model provides a direct and joint representation of semantic, dynamic and uncertainty aspects in the grid.

4.1. Architecture of the system based on the egocentric referential approach

In robotics, two strategies are usually employed to define the spatial position of a robot and the elements that compose this spatial environment. These two strategies are denoted by allocentric and egocentric frames of reference.

In the allocentric frame of reference, all the objects in the environment have a spatial position with reference to a fixed point, and this fixed reference point does not move over time. This kind of strategy is widely used in cartography or in SLAM-based approaches. In the egocentric frame of reference, all the objects in the environment occupy a spatial position with reference to a relative point that moves over time. In this work, the egocentric frame of reference is adopted to avoid restrictions with respect to precise global localization and drifts inherent to this strategy.

Consider a car-like robot that operates within a finite domain D of a world plan. This domain is defined in an Euclidian space that has two dimensions $\mathbb{E}^2$, as can be seen in Figure 2(a). In this case, the Egocentric frame of reference is determined by associating the reference R_M to the fixed point M, which is defined as the center of the car-like robot in the spatial environment. For the sake of simplicity, this robot is hereinafter referred to as an ego-car. Thus, the occupancy grid that is relative to the reference R_M will move together with the ego-car. This approach has the advantage of always covering the same area around the ego-car, without limiting its field of evolution.

To increase the reliability of the perception while the ego-car moves in the environment, a temporal filter can be used on grids to take into account ob-

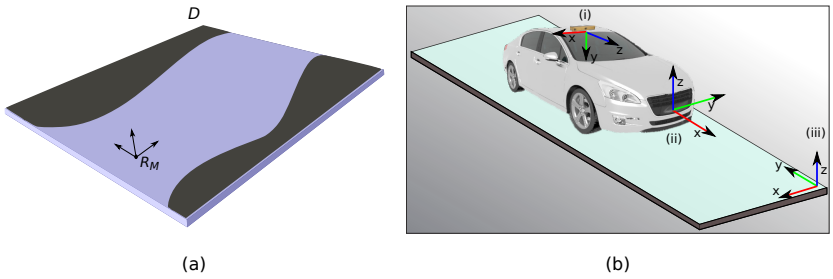

Figure 2: (a)Egocentric frame of reference for car-like robot (ego-car). The irregular surface represents the evolution field of the ego-car embedded in the domain D, having a relative point of reference fixed at the center of the robot. (b)The defined coordinate Systems. (i) Camera coordinate system, (ii) vehicle coordinate system, (iii) grid coordinate system.

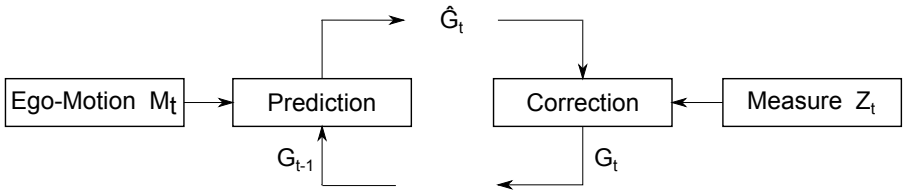

Figure 3: Temporal filter based on predictor-corrector approach.

served redundancies obtained from measurements over time. This approach can render the occupancy grid more robust against noise and with more complete information. This temporal filter uses the predictor-corrector type formalism, as depicted in Figure 3. Thus, this predictor-corrector type formalism is employed on Recursive Bayesian Filters and here is referred to handle uncertainties models using the formalism of the Dempster-Shafer theory (DST).

Based on the predictor-corrector approach, the principle of the system proposes the use of two distinct occupancy grids, called Sensor Grid and Perception Grid, to perform the sequential updating, as depicted in Figure 4. The Sensor Grid (SG) is built from the measured Z_t and merged in the Perception Grid, which is described in subsection 4.4. The Perception Grid (PG) maintains the cells state between the different instants of time, thus performing a Dynamic Local Perception. To merge the SG and the PG, these two grids must be spatially and temporally coherent. At every instant of time when a new SG is available, the current position is estimated and the prediction of PG must be done for the current time. The variable M_t represents the rigid transformation $M = [R|T]$,

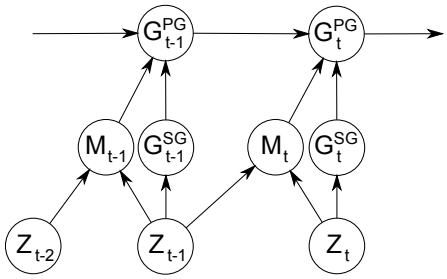

Figure 4: Sequential update of the Occupancy Grid.

where R is the rotation matrix and T is the translation vector. The measure Z_t is obtained by a stereo camera.

Prediction for the grid is performed whenever a new measure is available, and is necessary because of the movement of the ego-car and of the elements of the scene that compose the dynamic environment. An incorrect prediction may cause inconsistency in the PG. Therefore, the vehicles movement (ego-motion) between two sequential sampling measurements must be compensated. This is done by applying the rigid transformation that includes the cars geometry and the position of the camera sensors. Moreover, the bilinear interpolation method is used to better fit the values among the cells of the grid in two instants of the time. In fact, the matrix of the rotation and translation that composes the rigid transformation is obtained by employing a Visual Odometry technique based on Structure from Motion (SfM), which is proposed in the distinguished work of [34] and applied here.

Although the ego-motion has been compensated, the information contained in the preceding grid is no longer completely valid because the scene has changed from one instant to the next. The dynamic of the scene between these two instants of time remains fixed, thus increasing the uncertainty in the grid. An advantage of this action is that when the SG and PG are merged, as detailed in subsection 4.4, there will be a conflict between the values of the same cell. These conflicting cells indicate a possible moving object in the scene. The mechanism that manages uncertainties in the form of evidence is the same as the one that governs this detection procedure.

A brief overview the system architecture was presented. Details about specific methods are now given in the next sections. Subsection 4.2 describes the

formalism based on evidence theory to manage the uncertainties and also to detect the mobile cells. Subsection 4.3 presents the inverse model sensor to model the SG, followed by subsection 4.4, which explains the complete Local Dynamic Perception proposed in this article.

4.2. Fundamentals of Evidential Grid

Occupancy grids are used to estimate the occupation of space with uncertainties. How these uncertainties are represented has important implications on how the information in the grid is handled and how the data is then interpreted. On this basis, the tool to manage uncertainties associated with the responses of the occupancy grid governed by a mathematical theory of evidence [27]. Specifically, the occupancy grid uses the formalism of the Dempster-Shafer theory (DST) to model the uncertainties, which is a generalization of the Bayesian theory of subjective probability [35]. The DST model associated in the occupancy grid is called the Evidential Occupancy grid or the Evidential grid. As previously explained, there are some works which use evidential grids in the context of mobile perception [29, 32] and autonomous vehicles [36, 37]. The reason for this choice is that the approach allows for faster convergence [38], conflict detection, fusion of unreliable sources, etc. [2].

In the evidential grid, uncertainties are modeled as a belief function. The proposition in question is defined by Free and Occupied, having a set composed of $\Omega = \{Occupied(O), Free(F)\}$. The frame of discernment (FOD) of Ω is the set of all possible subsets of Ω and is denoted by $2^{\Omega} = \{Occupied(O), Free(F), unknown(\Omega), conflict(\emptyset)\}$. There are various forms to represent the belief function, such as mass, belief, plausibility and communality, and all these representations are equivalent [2]. In this approach, the mass function m^{Ω} is used, and has the following property given by Equation (1):

$$\begin{gathered} m : 2^{\Omega} \rightarrow [0...1] \\ \sum_{A \in 2^{\Omega}} m(A) = 1 \end{gathered} \tag{1}$$

The mass function of all the cells of the evidential grid is a vector containing four masses, defined by I^{O}, whose values represent the belief for each element

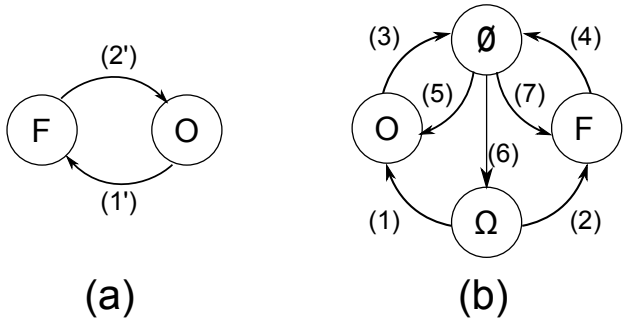

Figure 5: Comparison of belief transition in Bayes' theorem and DST. (a) The belief transition in the probabilistic approach. (b) In the DST approach, the transitions are: fusion (1,2), conflict generation (3,4) and conflict normalization (5,6,7). Source: adapted from [2].

of index i and j in 2^{Ω}:

$$I_{ij}^{O} = \{\ m(F)\ m(O)\ m(\Omega)\ m(\emptyset)\ \} \tag{2}$$

In Equation (2), the mass of each element corresponds to a belief level that the cell of the grid (G_{ij}) is in a given state. All the cells of the grid are initialized with a mass function called Basic Belief Assignment (BBA).

Before explaining the mechanism of decision and fusion in the evidential grid, it is opportune to make a brief comparison between DST and Bayes' theorem (BT). In Bayes' theorem, the probabilistic method describes the occupation of a cell using only one probability value per cell, and the belief function requires the computation of three masses (the fourth mass is obtained by the condition of equation 1) [39]. In this case, the computation cost required by the evidential approach is higher, considering memory use and processing time. However, in the probabilistic approach, the belief transition is possible only between the two states and is symmetrically restricted. In the evidential approach, the belief can be transferred among the four states, in which each of these transitions has a different meaning, dynamic and importance. Figure 5 compared the belief transition of the Bayes' theorem and the DST.

As a simple example extracted from [40], considering that one cell contains 3D points from obstacles, according to Bayes' theorem, $P(O)$ would be somewhat greater than 0.5. Let us assume that $P(O) = 0.6$. According to the DST, it has a belief mass of $m(O) = 0.6$. The fewer 3D points from obstacles one

cell contains the lower the certainty that the cell is occupied. This uncertainty can be represented by $m(\Omega)$. Since there is no evidence detected that a cell is free, it has $m(\Omega) = 1.0 - m(O) = 0.4$. According to Bayes' theorem, it has $P(F) = 1.0 - P(O) = 0.4$. This means that the uncertainty is automatically represented as free, which is not quite correct. Thus, these comparisons illustrate the applicability and relevance of the DST approach. It should be highlighted that both strategies (DST and BT) have their pros and cons, but, in the context of this paper, DST is employed mainly to distinguish between uncertainty caused by different phenomena like missing or conflicting information. Others techniques could also be applied to this end [46].

The updating procedure is formalized using Dempster's rule of combination. This fusion operator allows for the merging of two independent mass functions defined in the same FOD. Furthermore, it assumes that all the sources are reliable and its result leads to a more informative mass function than the two previous sources [35]. In this sense, two reliable sources can be merged in two steps: the conjunctive combination rule followed by the normalization of the conflicting mass function $m(\emptyset)$. In Equation 3, the result is denoted by $m_1 \otimes m_2(A)$, taking m_1 and m_2 as mass functions of two reliable sources and applying the conjunctive rule $\otimes$.

$$\begin{cases} m_{1,2}(\emptyset) & = 0 \\ m_{1,2}(A) & = (m_1 \otimes m_2)(A) = \frac{1}{1-(m_1 \otimes m_2)(\emptyset)} \sum_{B \cap C = A \neq \emptyset} m_1(B).m_2(C) \end{cases} \tag{3}$$

where

$$(m_1 \otimes m_2)(\emptyset) = \sum_{B \cap C = \emptyset} m_1(B).m_2(C) \tag{4}$$

It is observed that the merging process using Dempster's rule of combination only works with independent mass functions defined in the same FOD. To include the meta-knowledge extracted from Urban Road Scene Understanding, it is essential to extend the FOD to incorporate this information. A problem arises when the set $\Omega = \{F, O\}$ shifts to represent $\Omega = \{F, O, ..., C_n\}$, since, in this case, the FOD increases exponentially to 2^{Ω}, depending on the number of propositions C_n added to the set. This phenomenon has a direct impact on the

time involved in processing any calculation on the grid. A proposed alternative to avoid this problem is to maintain the propositions of Free and Occupied, and then generate a refinement in the Occupied proposition, as described below.

The refinement consists in expanding a subset of propositions with respect to the Occupied proposition. This subset is defined by $r(O) = \{V, B, T, S\}$, which denotes, respectively, (V) vehicle, (B) building, (T) vegetation and (S) sidewalk. It is assumed that each proposition in the occupied refinement subset is represented by $\{or_i | \forall i \in r(O)\}$. It should be noted that the set $r(O)$ is not the same as the meta-knowledge delivered by the Semantic Context task, i.e., the set $r(O)$ considers only meta-knowledge defined as obstacle, in this case disregarding the classes sky, road and void. The combination rule for the occupied refinement proposition, denoted by $Prop(O)$, is obtained by the ro argument of the highest mass function between the two items of evidence, conditioned to the fact that $argmax[m_1 \otimes m_2(A)] = O$. The fusion rule for proposition refinement is given by (5):

$$Prop_1 \otimes Prop_2(A) = \left\{ or \;\middle|\; \begin{array}{c} argmax[m_1(A), m_2(A)] \text{ and} \\ argmax[m_1 \otimes m_2(A)] = O \end{array} \right\} \tag{5}$$

After demonstrating the procedure to update two grids by merging its items of evidence, it is now possible to introduce the conflict analysis used by [29] and [32] to detect mobile cells. In fact, the conflict is determined when two items of information are merged. In DST, the conflict is represented explicitly by $m(\emptyset)$. If the mass function resulting from the fusion of Dempster's rule (before normalization) is $m(\emptyset) \neq 0$, this means that the merged information is at least partially contradictory. According to [2], the conflict may be caused by different factors: a difference in expert opinion or an incorrect system modeling. In the present case, the source of conflict arises from two principal errors:

- The assumption that the grid is static, since the observed scene contains dynamic elements.

- The geometric approximation due to discretization in the sensor model and during the grid propagation.

Considering these errors, [2] proposes breaking down the term $(m_{t-1}^{PG} \otimes$

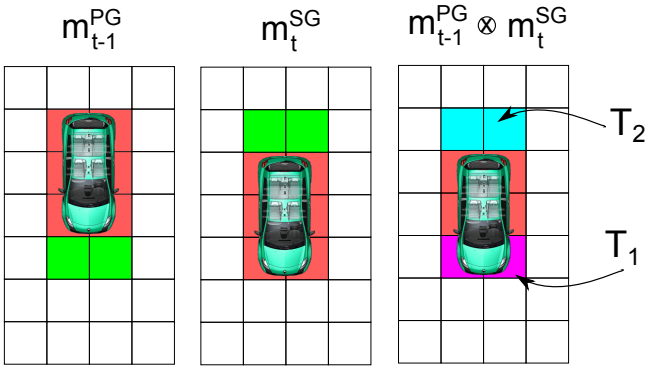

Figure 6: Example of the analysis of a conflict generated by a mobile object. Red cells represent the occupied area, green cells represent free area. Cyan cells represent the conflict depicted by T_2 and pink cells represent the conflict explained by T_1.

$m_t^{SG})(\emptyset)$ into two other terms (Equation 6):

$$(m_{t-1}^{PG} \otimes m_t^{SG})(\emptyset) = \underbrace{m_{t-1}^{PG}(F).m_t^{SG}(O)}_{T_1} + \underbrace{m_{t-1}^{PG}(O).m_t^{SG}(F)}_{T_2} \tag{6}$$

The first term T_1 corresponds to a cell that was previously free with a certain confidence level $m_{t-1}^{PG}(F)$ and, at the current time t, it is observed to be occupied with a confidence level of $m_t^{SG}(O)$. If one considers that the conflict arises from a moving object in the scene, the term T_1 means that a free cell becomes occupied, and therefore, that an object is entering the space represented by the cell. Likewise, the term T_2 means that an occupied cell becomes free and consequently, that an object is leaving the space represented by the cell.

Taking into account the conflict generated in the cells due to a moving object, the terms T_1 and T_2 can be analyzed to provide insights not only about the conflict itself, but also to determine the direction of a moving object, as illustrated in Figure 6.

4.3. Sensor Grid Model

This subsection describes the method to build the Sensor Grid (SG). The SG is built in every instant when the sensor provides a new measurement. It transforms the acquired data to its representation in the evidential grid. Therefore, in some way it implements the sensor model used in the algorithm.

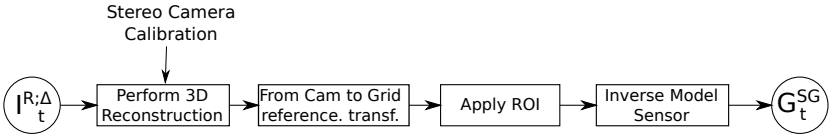

Figure 7: Architecture of the Sensor Grid Model.

The general conception is depicted in the diagram of Figure 7. The disparity map I^{Δ} is obtained by applying a standard algorithm to a pair of rectified stereo images. Based on the Epipolar Geometry and camera calibration, 3D reconstruction is performed obtaining the 3D points referenced in the camera. After that, an affine transformation using homogeneous coordinates is calculated to represent the points in the reference of the grid R_G. Because of restrictions in camera position and grid dimensions, only 3D points that fall within this region of interest (ROI) are considered. An improvement is done in the ROI to consider also specific points that have an associated meta-knowledge, which is observed in the Semantic Urban Road Scene Understanding and denoted by I^R. Finally, the Inverse Sensor Model is computed, using a Gaussian to represent the uncertainties associated with the points.

The principle of 3D reconstruction is to recover metric points from associated pixels of a rectified pair of stereo images, and also to incorporate the meta-knowledge linked to the Semantic Context. The 3D reconstruction applies the methodology explained by [41] to obtain the points in the 3D Cartesian space relative to the camera on the left, with homogeneous coordinates $[X^c, Y^c, Z^c, W]^T$. Incorporating the meta-knowledge associated with the semantic context, a 5-tuples denoted by P^c, where c represents the reference of the camera R_c, is defined containing the homogeneous 3D point and the information of the occupied refinement proposition denoted by ro. This transformation is shown in Equation (7):

$$P^c = \begin{bmatrix} X^c \\ Y^c \\ Z^c \\ W \\ or \end{bmatrix} = \begin{bmatrix} \frac{u.Z^c}{f_x} \\ \frac{v.Z^c}{f_y} \\ \frac{f.b}{d} \\ 1 \\ I^R_{ij} \end{bmatrix} \tag{7}$$

where

$$\begin{aligned} u &= i - c_x \\ v &= j - c_y \end{aligned} \tag{8}$$

In Equation 7, f, f_x, f_y represent the focal lengths in pixels and are obtained by the off-line calibration process. b represents the baseline of the stereo cameras (in meters). d represents the value of disparity obtained from I^{Δ}. I^R_{ij} represents the value of the semantic context at index position i, j. Finally, in equation (8), c_x and c_y are the coordinates of the optical axis in the image plane.

The set points P^c should be expressed in coordinates of the grid to compute the subsequent steps. To do this, two relations should be defined, from the reference of the camera to the reference of the vehicle and from the reference of the vehicle to the reference of the grid. The coordinate systems of these three references are defined as illustrated in Figure 2(b), i.e.:

- **Camera**: x = right, y = down, z = forward
- **Vehicle**: x = forward, y = left, z = up
- **grid**: x = right, y = backward, z = up

The first transformation is obtained by defining the position of the camera with respect to the center of the ego-car. Let us assume that the left camera is fixed at point ${}^0P^m$ in the vehicle reference R_M. The affine transformation is built by applying the translation of the point ${}^0P^m$ followed by two rotations, $\beta = -90$ degrees on the Y axis and $\alpha = 90$ degrees on the X axis. Therefore, the affine transformation from camera to vehicle is obtained by Equation (9):

$$M_{cam\_car} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & cos(-\alpha) & -sin(-\alpha) & 0 \\ 0 & sin(-\alpha) & cos(-\alpha) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{\text{Rx}} * \underbrace{\begin{bmatrix} cos(-\beta) & 0 & sin(-\beta) & 0 \\ 0 & 1 & 0 & 0 \\ -sin(-\beta) & 0 & cos(-\beta) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{\text{Ry}} * \underbrace{\begin{bmatrix} 1 & 0 & 0 & -{}^0P^m_x \\ 0 & 1 & 0 & -{}^0P^m_y \\ 0 & 0 & 1 & -{}^0P^m_z \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{\text{T}} \tag{9}$$

The second transformation is obtained by defining the position of the ego-car with respect to the origin of the grid. Let us assume that the ego-car is fixed at point ${}^0P^g$ in the grid reference R_G. The affine transformation is built by applying the translation of the point ${}^0P^g$ followed by a rotation of $\theta = -90$

degrees on the Z axis. The Y axis should then be inverted, and finally, a scale factor should be employed considering the discretization (Δ_x, Δ_y) in the grid. Therefore, the affine transformation from vehicle to grid is obtained by Equation (10):

$$M_{car\_grid} = \underbrace{\begin{bmatrix} 1/\Delta_x & 0 & 0 & 0 \\ 0 & 1/\Delta_y & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{\text{resolution factor}} * \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{\text{invert Y axis}} * \underbrace{\begin{bmatrix} cos(-\theta) & -sin(-\theta) & 0 & 0 \\ sin(-\theta) & cos(-\theta) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{\text{Rz}} * \underbrace{\begin{bmatrix} 1 & 0 & 0 & -{}^0P_x^g \\ 0 & 1 & 0 & -{}^0P_y^g \\ 0 & 0 & 1 & -{}^0P_z^g \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{\text{T}} \tag{10}$$

To finish the affine transformation between the reference of the camera and the reference of the grid, the final transformation simply consists in multiplying these two previous matrices. Thus, the set point P^c is represented in the reference of the grid as P^g and is obtained by Equation (11):

$$P^g = M_{car\_grid} * M_{cam\_car} * P^c \tag{11}$$

Due to the restrictions in camera position and grid dimensions, the set composed of all the reconstructed points P_g under filtering molded by a ROI. The ROI is defined considering the following restrictions:

- Assuming that the plane formed by the optical axis (Z axis) with the horizontal axis (X axis) is parallel to the road surface, a value of height from the road surface is defined, at which the points that exceed this threshold are not considered;
- Observing the grid dimensions, all the points outside of this condition, $0 \leq P^g \leq Grid_{size}$, are also discarded;
- For the remaining points inside the ROI, only the ones whose semantic context is associated with obstacles are considered.

Therefore, taking into account the defined ROI, the selected set of points can be projected onto the grid. This selected set, denoted P^s, is defined by (12):

$$P^s = \{P^g \mid \forall P^g \subseteq ROI \text{ and } or \in r(O)\} \tag{12}$$

To project the P^s set onto the grid, an inverse sensor model is described considering the noise in stereo measurements and also the uncertainty linked

with epipolar geometry reconstruction, where exponential error is observed as the distance increases. This method approximates the uncertainties using a Gaussian distribution, as shown in Figure 8. The inverse sensor model defined by $\psi_O^{prob}(G^{SG}, P^s)$ has the *prob* index representing the probability distribution, and O index representing the Occupied proposition. The function can be described by Equation (13):

$$\psi_O^{prob}(G^{SG}, P^s) = \left\{ min(\sum_{G_{ij}^{SG} \cap AG} k.exp^{\Upsilon}, \vartheta_O) \,|\forall P_n^s \in \{P^s\} \right\} \tag{13}$$

where

$$\Upsilon = (-\alpha.Dx^2 + 2\beta.Dx.Dy + \gamma.Dy^2) \tag{14}$$

and

$$\alpha = \frac{cos^2\theta}{2\sigma_x^2} + \frac{sin^2\theta}{2\sigma_y^2} \tag{15}$$

$$\beta = -\frac{sin2\theta}{4\sigma_x^2} + \frac{sin2\theta}{4\sigma_y^2} \tag{16}$$

$$\gamma = \frac{sin^2\theta}{2\sigma_x^2} + \frac{cos^2\theta}{2\sigma_y^2} \tag{17}$$

In Equation (13), κ is a constant representing the percentage that a single 3-D point could contribute to the occupancy level of a cell G_{ij}. ϑ_O is a parameter that belongs to $[0, 1]$ and reflects the confidence in the measurement (1 if confident). This confidence is linked to the principle of measurement (false alarm or missed detection). Dx and Dy represent the difference between the coordinates of P_n^s and C_{ij}, i.e., $Dx = C_{ij}.x - P_n^S.x$ and $Dy = C_{ij}.y - P_n^S.y$. The index $G_{ij}^{SG} \cap AG$ at the sum in Equation (13) represents the area of the Gaussian AG, whose distribution overlaps the cells of the grid G^{SG}. The parameters σ and θ model the dispersion of the distribution as a function of the distance and orientation relative to the ego-car. The dispersion of the Gaussian considering the distance Z of the camera is modeled considering each value of disparity $\{d_i \in \{d\}\}$, as depicted in Equation (18):

$$\begin{aligned} \sigma_y &= \left\{ \frac{\sqrt{\sum_{n \in N(d)} \left[\left(\frac{f.b}{d_i}\right) - \left(\frac{f.b}{d_n}\right) \right]^2}}{card(N(d)) - 1} \,|\forall d_i \in \{d\} \right\} \\ \sigma_x &= \left\{ \frac{\sigma_y}{d_i} |\forall d_i \in \{d\} \right\} \end{aligned} \tag{18}$$

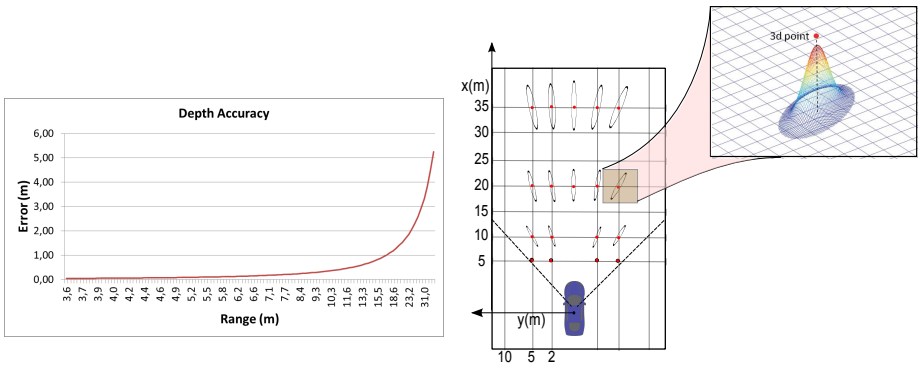

Figure 8: Model of uncertainty associated with epipolar geometry reconstruction and noise in stereo measurements.

In Equation (18), $N(d)$ represents the neighborhoods of the disparity and $card(N(d))$ represents the cardinality of the set $N(d)$. In the case of parameter θ, the disparity is modeled by Equation (19):

$$\theta = \left| arctan\left(\frac{{}^{0}P^{g}.x - P^{s}_{x}}{{}^{0}P^{g}.y - P^{s}_{y}} \right) \right| \tag{19}$$

The inverse sensor model proposed to project the meta-knowledge information onto the grid is based on the principle of voting. This method assumes that the semantic information that best represents the cell is defined by the sum of votes that a given meta-knowledge has received from the points belonging to the cell. Therefore, the occupied refinement subset is modeled with proposition $\{or \in r(O)\}$, in this case denoted as $\psi^{prop}_{O}(G^{SG}, P^{s})$, as described by (20):

$$\psi^{prop}_{O}(G^{SG}, P^{s}) = \left\{ \underset{ro}{\text{argmax}}(\omega(G^{SG}_{ij}, P^{s})) | \forall G_{ij} \in G^{SG} \right\} \tag{20}$$

where

$$\omega(G^{SG}_{ij}, P^{s}) = \left\{ \sum_{P^{s}_{n} \subseteq S^{SG}_{ij}} \delta(P^{s}_{n}.or, or_{l}) | \forall or_{l} \in r(O) \right\} \tag{21}$$

$$\delta(or, or_{l}) = \begin{cases} 1 & \text{, if } or = or_{l} \\ 0 & \text{, otherwise} \end{cases} \tag{22}$$

In Equation (21), the index $P_n^s \subseteq S_{ij}^{SG}$ represents all the 3D points $P_n^s \in \{P^s\}$ contained on the surface of the cell S_{ij}^{SG}.

Up to this point, the solution is able to manage the probability of occupied areas using 3D points that correspond to obstacles. It is also able to perform the occupied refinement, which determines the proposition that best represents those occupied areas. Based on the principle that the obstacles have already been processed, a simple but effective method is employed to model the free areas. If a light ray from the camera sensor reaches a detected obstacle point, the purpose of this method is that it can be stated, within a given probability, that every cell that lies along this line is free. Thus, the solution for free areas, defined by $\psi_F^{prob}(G^{SG}, FL)$, where $\{FL\}$ denotes the set of Free Lines, is modeled by a function that attributes the free probability to all the cells that intercept the line generated from the camera sensor position to all the first obstacles detected. This technique, which is performed using the Bresenham algorithm [42], is described by Equation (23):

$$\psi_F^{prob}(G^{SG}, FL) = \left\{ \max_{fl \cap G_{ij}^{SG}} (1, 0 - \vartheta_F)) | \forall fl \in \{FL\} \right\} \tag{23}$$

In Equation (23), ϑ_F is a parameter that belongs to $[0, 1]$ and reflects the confidence in the measurement of the Free area (0 if confident). As previously explained, this confidence is linked to the principle of measurement (false alarm or missed detection).

To conclude the inverse sensor model for all the states, the function that models the unknown state (Ω) should respect the property presented in Equation (1), and is defined by Equation (24):

$$\psi_\Omega^{prob}(G^{SG}) = 1.0 - \psi_O^{prob}(G^{SG}, P^s) - \psi_F^{prob}(G^{SG}, FL) \tag{24}$$

Therefore, the resulting evidential grid, which is modeled to represent the sensor grid at each instant of a measurement, has its BBA defined as Equa-

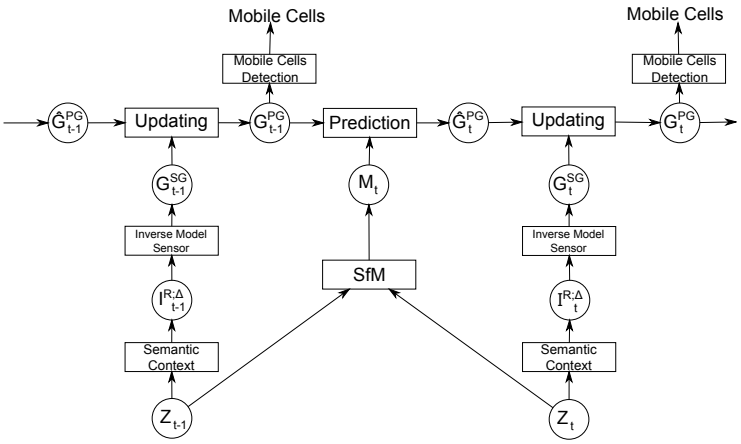

Figure 9: Detailed system architecture proposed for Dynamic Local Perception.

tion (25) :

$$\begin{aligned} m^{SG}(O) &= \psi_O^{prob}(G^{SG}, P^s) \\ m^{SG}(F) &= \psi_F^{prob}(G^{SG}, FL) \\ m^{SG}(\Omega) &= \psi_\Omega^{prob}(G^{SG}) \\ m^{SG}(\emptyset) &= 0 \\ Prop^{SG}(O) &= \psi_O^{prop}(G^{SG}, P^s) \end{aligned} \tag{25}$$

4.4. Dynamic Local Perception Grid

As introduced in previous subsections, the *Perception grid* (PG) is responsible for building the final representation of the environment over time. The complete Dynamic Local Perception (DLP) proposed in this work is a new conception involving the Semantic Urban Road Scene Understanding with Occupancy grids. Its architecture is detailed in Figure 9. At each instant t that a measurement is acquired, the meta-knowledge I_t^R that represents the Urban Road Scene Understanding is stored. For the sale of simplicity, this process is dubbed Semantic Context. The I_t^R and the Disparity Image I_t^Δ jointly supply the necessary information to construct the G_t^{SG}, performing the novel Inverse Sensor Model. The prediction of the $\hat{G}_t^{PG}$ is then estimated, after which the G_t^{SG} is updated with this estimation, using the DST methodology to manage the system's uncertainties. In the updating procedure, it is possible to detect the mobile cells based on G_t^{PG}.

Having described the concept of the DLP, the details of the system architecture are presented. The Semantic Context is mentioned in Section 3, and the Inverse Sensor Model is described in subsection 4.3.

The prediction process estimates the grid G^{PG} to $\hat{G}^{PG}$ as a function of the displacement generated by the ego-car at instant $t-1$ to t. The rigid transformation ($M_t = [R|T]$) that represents this displacement is performed in two consecutive images, as explained in subsection 4.1. Thus, $G^{PG}_{t-1} \rightarrow G^{PG}_{t}$ uses the affine transformation function $f(G^{PG}_{t-1}, M_t)$ to update the cell information, given by Equation (26).

$$\begin{array}{rcl} \hat{m}^{PG}_{t} & = & B(f(m^{PG}_{t-1}, M_t)) \\ \hat{Prop}^{PG}_{t} & = & \varpi(f(Prop^{PG}_{t-1}, M_t)) \end{array} \tag{26}$$

where

$$\varpi(f) = \{max(\sum_{x=N_{cell}} \delta(or_j, or_x)) \mid \forall j \in r(O)\} \tag{27}$$

In Equation (26), the function $B(.)$ applies the bilinear interpolation to the mass function, and the function $\varpi(.)$ performs the same process as $B(.)$, but in the occupied refinement proposition. Following Equation (27), N_{cell} stands for all the neighbors of the cell in the grid, and $\delta(.)$ is defined in Equation (22). In this process, some cells disappear and other cells appear within the scope of the new grid. These new cells are initialized with the unknown mass function ($\Omega = 1.0$).

After $\hat{G}^{PG}_{t}$ is computed, the fusion process with G^{SG}_{t} can be performed. Each cell refers to an occupancy mass function defined on 2^{Ω} plus the refinement $r(O)$ shown in Equations (3) and (5) of subsection 4.2. The values of the mass function m^{PG}_{t} at time $t = 0$ represent no prior information (28):

$$\begin{array}{rcl} m^{PG}_{t}(O) & = & 0.0 \\ m^{PG}_{t}(F) & = & 0.0 \\ m^{PG}_{t}(\Omega) & = & 1.0 \\ m^{PG}_{t}(\emptyset) & = & 0.0 \\ Prop^{PG}_{t}(O) & = & (\{\}) \end{array} \tag{28}$$

However, the updating mechanism is achieved in two steps in order to keep

the conflicting information and also to combine the proposition refinement. The first step, the fusion process, uses Equation (3) without the normalization factor to merge the mass function. Equation (5) is applied to obtain the associated proposition, as demonstrated in Equation (29).

$$\begin{array}{rcl} m'^{PG}_t & = & \hat{m}^{PG}_t \otimes m^{SG}_t \\ Prop^{PG}_t & = & \hat{Prop}^{PG}_t \otimes Prop^{SG}_t \end{array} \tag{29}$$

In Equation (29), m'^{PG}_t is the conjunctive fusion, i.e., Dempster's rule without the normalization factor. The second step, the updating process, is performed by normalizing the mass function by the conflict mass, as shown in Equation (30). It should be noted that the conflict mass $m'^{PG}_t(\emptyset)$ is stored for mobile detection analysis.

$$\begin{cases} m^{PG}_t(A) = \frac{m'^{PG}_t(A)}{1-m'^{PG}_t(\emptyset)} & A \neq \emptyset \\ m^{PG}_t(\emptyset) = 0 & A = \emptyset \end{cases} \tag{30}$$

A mobile object is detected by analyzing the conflict mass m'^{PG}_t. If G^{PG}_{t-1} and G^{SG}_t are contradictory, this indicates the occurrence of a conflict, which can be analyzed based on the Equation (6). As previously explained, the first term T_1 detects the conflict generated when a moving object leaves the cell, while the second term T_2 detects the conflict generated when a moving object appears in the cell. Due to noise and imprecise measurements arising from data acquisition, poor displacement estimation, etc., many false-positive detections may appear. In this case, using the meta-knowledge associated in the G^{PG}_t, the detection of mobile obstacles can be improved by implementing a restriction that allows only the $r(O) \supset V$ to generate such a conflict.

$$m'^{PGr}_t(\emptyset) = \{m'^{PG}_{ij} | Prop^{PG}_t(O) \subseteq \{V\}\} \tag{31}$$

5. Experimental Results

Experiments were carried out in real-life conditions using the common Kitti benchmark[1] [43]. The experiments were conducted and aimed at applications in

[1] `http://www.cvlibs.net/datasets/kitti/` accessed on 06 Dec 2018

Advanced Driver Assistance Systems (ADAS) and any future driving maneuver or vehicle control for autonomous navigation.

The validation platform is implemented in C++ and the experiments were executed on a computer equipped with an Intel I7-7700HQ processor with 2.8Ghz and with 24Gb DDR3, running version 16.04 of the Linux Ubuntu.

The dataset used is composed of 446 images acquired in an inner-city scene, having a sequence image of 0:45min. The images include common objects such as cars, trees, and buildings at a resolution of 1392 x 512 pixels. In this dataset, the sensor camera is characterized by two Point Grey Flea2 FL2-14S3C-C color cameras with a focal length of 4 mm and a ∼90 degree horizontal opening angle. The baseline of the stereo camera rigs is approximately 54 cm. The principal point of the left calibrated camera is in $c_x = 609.5593$ px. and $c_y = 172.8540$ px., and its focal length is $f_x = f_y = 721.5377$ px. The mounting position of the sensors with respect to the vehicle body, which is illustrated in Figure 10(a), is taken from the work of [43].

For this dataset, the grids are defined to cover an area of 39.9m x 53.1m with a resolution of 0.3m x 0.3m. The center of the ego-car was positioned on the grids with coordinates (20m, 50m), keeping in mind that the reference is fixed on the left upper side of the grid. The transformations between references were modeled considering the car geometry and positions of sensor cameras described earlier herein.

The validation and performance analysis of this type of perception system was carried out comparing the classic evidential grid with the semantic evidential grid defined by DLP. The classic evidential grid does not take into account the semantic information. The generation of the classic evidential SG and PG grids are similar to those described earlier in sections 4.3 and 4.4. The main differences are how the 3D points are selected to be projected onto the Sensor Grid. Instead of using Equation 12 to determine the P^s set, the procedure was changed as follows. The P^s set is modeled considering as restriction all 3D points that have only a determined height value from the ground, which defines those points belong to an obstacle or not. After that, the filtered P^s set is then projected onto the SG as explained, without considering those Equations related to meta-knowledge information. In the same way, the classic evidential PG grid

handles the prediction and fusion processes normally as explained, highlighting that the classic evidential grid does not have the meta-knowledge information to be managed.

For these experiments, the height parameter of the classic evidential SG grid was fixed to 0.30m. So, all 3D points with a height higher than 0.30m are projected in the grid. Therefore, the impact of the semantic information introduced in the evidential grid can be verified by the temporal analysis of a given cell's mass of the grid that remains in a given state and with a specific proposition, as well its qualitative result of the observed scene. This validation is observed in both grids, sensor grid (SG) and perception grid (PG).

A simple method is created to better visualize the various items of information contained in the DLP. This method proposes to give the HSV color space a different meaning. Figure 10(c) presents the HSV color space with its axis representing the arranged visualization, as follows. The idea is to represent the four states of the evidential grid considering also the semantic context information, such as building, sidewalk, road, vegetation and vehicle. To do this, the colors in the Hue axis, which range from 0 to 360 degrees, are changed to represent the semantic context information. Consequently, a fixed degree is defined for each class. Thus, 0, 60, 90, 120 and 300 represent, respectively, building, sidewalk, road, vegetation and vehicle. The set of degrees $\{0, 60, 120, 300\}$ belongs to the occupied state (O) and the 90 degree belongs to the free state (F). It should be mentioned that the 210 degree (blue) was inserted just for differentiating the obstacle detection from the classic evidential grid in comparison to the semantic evidential grid. The Saturation axis, which ranges from 0 to 100%, is then changed to represent the conflict state ($\emptyset$), but the values of the axis are inverted, i.e., the value of Saturation $S = 1 - m(\emptyset)$. This means that the greater the conflict the closer it will be for the white color. To conclude the proposed conception of visualization, the unknown state (Ω) is presented by the Value axis, modeled in the same way as the conflict state. This means that complete ignorance is represented by the color black. In this regard, the variations of the colors are proportional to the mass value of these respective states.

To illustrate the qualitative results of both methods, the following figures 11, 13 and 16 show on the left the original image, the disparity map contextualiz-

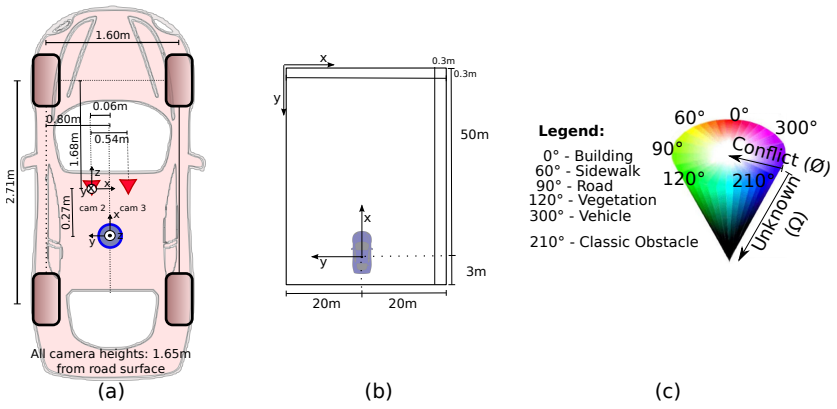

Figure 10: Sensor Setup. This figure illustrates the dimensions and mounting positions of the sensors (red) with respect to the vehicle body. Heights above ground are measured in relation to the road surface, using (a) the KITTI Benchmark [43] and (b) the Local Perception Grid setup. (c) The proposed visualization method using the HSV color space.

ing the metric information and the semantic context result upon which the meta-knowledge is based. These results are related to the task of the Semantic Context, regarding the proposed solution diagram (Fig. 1). Considering the Disparity Map module, it uses the SGBM algorithm [45] to obtain the resulting disparity map and applies the ProbBoost algorithm [44] to achieve the semantic information, at the Machine Learning module. The motivation behind these choices lies in the fact that these algorithms are naive ones and are passive of inaccuracies, leading to uncertainties inserted into the proposed system. Therefore, it is a way to analyze if the proposed system is able to cope with uncertainties come from different sources such as the inner-city environment as well as the system itself. For comparison purpose, on the right are presented either the classic with the semantic sensor grid (SG) or the classic with the semantic perception grid (PG).

The first example in Figure 11 illustrates the robustness of the road detection, which is the main element to perform autonomous navigation. As can be seen, this figure shows a typical scene of an urban area. The yellow circle highlights a problem that should be managed by a perception system using a camera sensor. The presence of a shadowy area combined with the stronger influence of the sunlight represents a challenging task to be dealt with. Considering the temporal analysis depicted in figure 12, upon which the masses were taken up

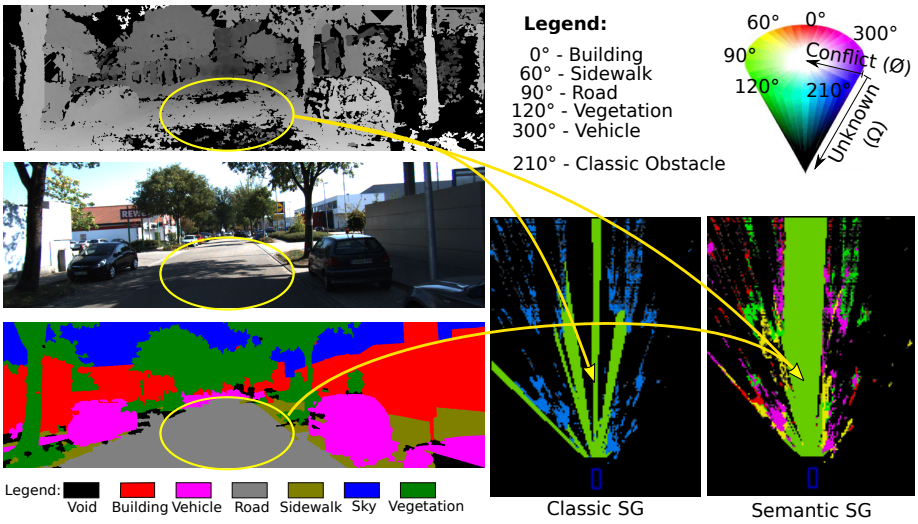

Figure 11: DLP result highlighting the road detection. The yellow circle highlights the presence of a shadowy area jointly with higher influence of the sun, and the sidewalk, which is quite similar to the road.

from the cell at position addressed by the yellow arrow, it is possible to see along the time how better become the road detection using semantic information to compose the sensor grid generation, compared with the classic sensor grid. Thanks to these factors, the DLP system is able to maintain a high level of confidence about the free space without using any other sensor or a prior digital map to build this perception.

The example in figure 13 demonstrates two cases, the conflicting cells that are able to detect a mobile vehicle and the robustness of the detection of the sidewalk. In the first case, represented by number 1, a moving car is passing by the ego-car, which is also moving. Due to inaccuracies in the estimation of the ego-motion and errors produced by the phenomenon of discretization and transformation between grids, several conflicting cells appear, as observed in both perception grid (Classic and Semantic). In this case, however, using the semantic context information to improve the detection (as presented by Equation 31), it is possible to distinguish these type of conflict by considering that only cells recognized as vehicles could be in movement. The same conclusion is verified in the temporal analysis in Figure 14. Observing the cell at the center

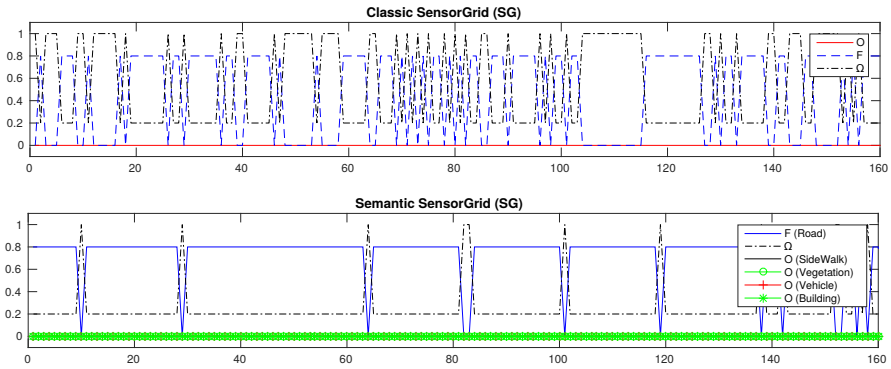

Figure 12: Temporal analysis of belief masses from SensorGrid (SG), considering the comparison between the classic evidential approach and the semantic evidential approach.

of the circle 1, the belief masses from the classic perception grid show the conflict mass that results from Free to Occupied transition and then from the Occupied to Free transition. For security reasons, it is not prudent to assume that this cell is moving at that specific interval observed, arising from the issues previously mentioned. Differently, the semantic perception grid highlights that the cells considered occupied were due to the fact that they belong to the specific vehicle class.

The second case, represented by number 2, clearly outline the influences of the semantic context information into the perception grid. As can be seen (Figure 13), the classic evidential approach is not able to detect precisely the sidewalk, since that sidewalk is quite similar to the road and the 3D points recovered from the disparity map are not distinguishable. On the other hand, introducing the semantic context information has been possible to accurately differentiate Free and Occupied areas, improving in this sense the exact road and sidewalk regions by where an autonomous car should navigate. Figure 15 depicts the temporal dynamic of the cell taken up from the center position of the ellipse 2. As can be seen, from the semantic evidential approach, the Occupied belief mass maintains at 1.0 all-time within the analyzed interval, changing among sidewalk, vegetation and vehicle. Taking a look at the classic evidential approach, one can observe that the regions of the sidewalk are not

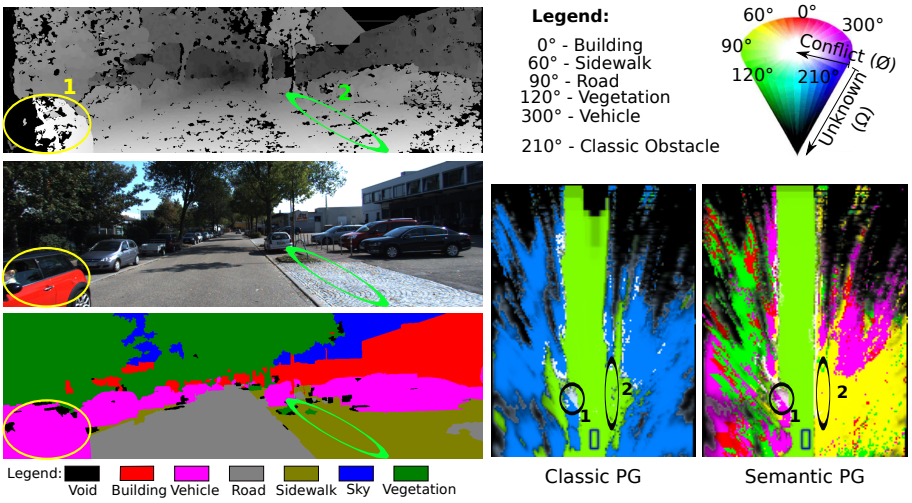

Figure 13: DLP result highlighting the distinction between mobile and static cells. The yellow ellipse highlights the presence of a mobile object, and the green ellipse shows the conflicting cells that can be filter out considering the meta-knowledge of the scene.

satisfactorily represented. Verifying these regions, intervals, they are modeled as belief mass of the Free proposition, that it is not true regarding the scene from the original image. The resulting output of the DLP presents the new conception to substantially improve the representation and understanding of dynamic urban environments, by the semantic context information usage.

The example illustrated in Figure 16 demonstrates the multi-detection of mobile objects in a challenging and complex scene. In this environment, the DLP system presents an outstanding approach, detecting all the vehicles in the scene. As can be seen in cases 1, 2 and 4, they are correctly detected as moving vehicles. However, two incorrect cases should be mentioned. Analyzing the third case, we find that the semantic context result correctly detected the occluded vehicle, but was unable to detect the vertical signs in front of the car. At the same time, examining the result of the disparity map, one can see that the distance from the vertical signs to the stereo camera was calculated correctly. Consequently, the projection onto the perception grid correctly maps the position of the object but associates an incorrect meta-knowledge, thus impairing the process of filtering out the conflicting noise cells. Regarding the case represented by number 5,

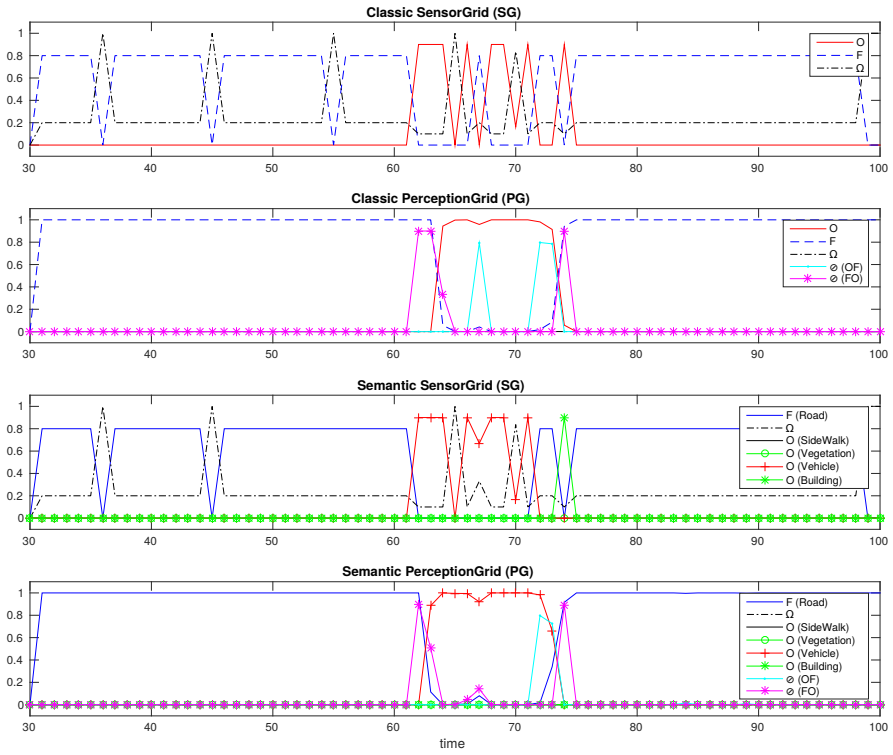

Figure 14: Temporal analysis of belief masses from both approaches, regarding the influences of the semantic context information to differentiate mobile vehicle and inaccuracies in the estimation of the ego-motion, errors produced by the phenomenon of discretization and transformation between grids, regarding the belief masses of the Conflict propositions.

it demonstrates an important risk factor that was not recognized. So far, the proposed system is not yet able to deal with pedestrian recognition. In this case, a pedestrian is identified as a poorly classified obstacle, which occupies an area in the grid that is inadequate to represent this kind of obstacle (as discretized), because each cell represents a 0.3 x 0.3m space, and its projection is represented by only 2 or 3 cells, making it difficult to distinguish considering noise.

To conclude these analysis, Figure 17 presents a phenomenon that happens at the fusion and upgrade processes, concerning the refined propositions management. Verifying the cell taken up from the center position of the circle 4 (Figure 16), the semantic sensor grid observes a car around the interval time 370, followed by the belief mass of the Unknown proposition. Looking the se-

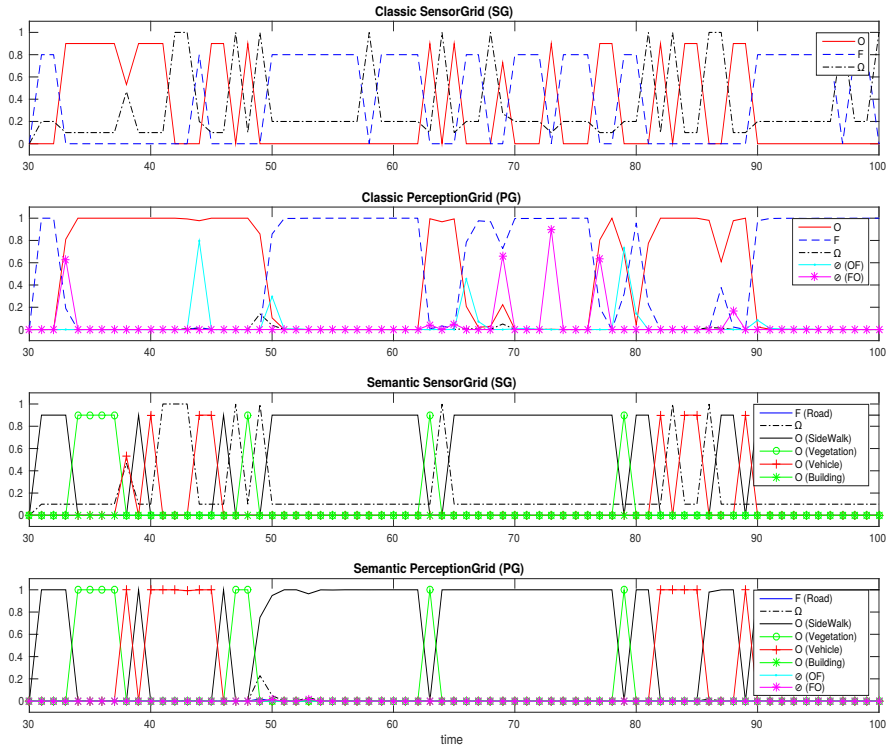

Figure 15: Temporal analysis of belief masses from both approaches, regarding the influences of the semantic context information to improve the representation and understanding of dynamic urban environments, mainly in non-trivial classes as road and sidewalk.

mantic perception grid at the same interval time, one can note that the belief mass of an Occupied area is correct, however, at interval time 380 the vehicle proposition changes to vegetation proposition, despite no evidence addressed in the semantic sensor grid at the same interval time. Thus, this result is related to the update process handled by the voting-based approach along the time. As the ego-car is moving, the past propositions information are propagated from neighborhood cells to the observed cell, even if in the observed cell there are not pieces of evidence about. This phenomenon will be handled in future works by using some temporal discounting and changing the voting-based approach to another one more sophisticated.

To finish, the computational load considering the aforementioned validation platform and the naive algorithms implemented to perform the proposed system,

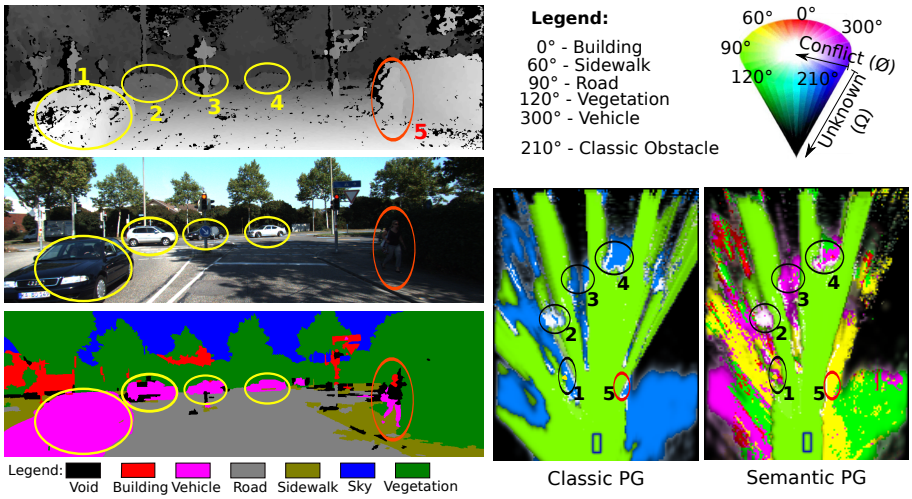

Figure 16: DLP result highlighting the multi-detection of mobile objects. The yellow ellipse highlights the presence of mobile vehicles, and the red ellipse show the wrong case to pedestrian recognition.

the Dynamic Evidential Grid task, including those three modules, takes around 350ms to process each data input. The video containing the complete result of the DLP for this experiment is publicly available, and can be found at [47] [2].

6. Conclusions

A new perception scheme based on dynamic mapping and relative localization using only a pair of stereo cameras has been introduced and applied to autonomous robotic vehicle navigation. The advantage of using stereo cameras, the possibility of measuring distances and the availability of image information. Therefore, the proposed approach, called Dynamic Local Perception, combines the evidential occupancy grid with meta-knowledge acquired by machine learning to characterize the uncertainties of occupied areas, while simultaneously incorporating the semantic context associated with these areas to improve the representation of dynamic urban environments over time.

In summary, this work contributes to this line of research by offering a novel

[2] `http://youtu.be/H_zJjX8uMtI` Accessed on 06 Dec 2018

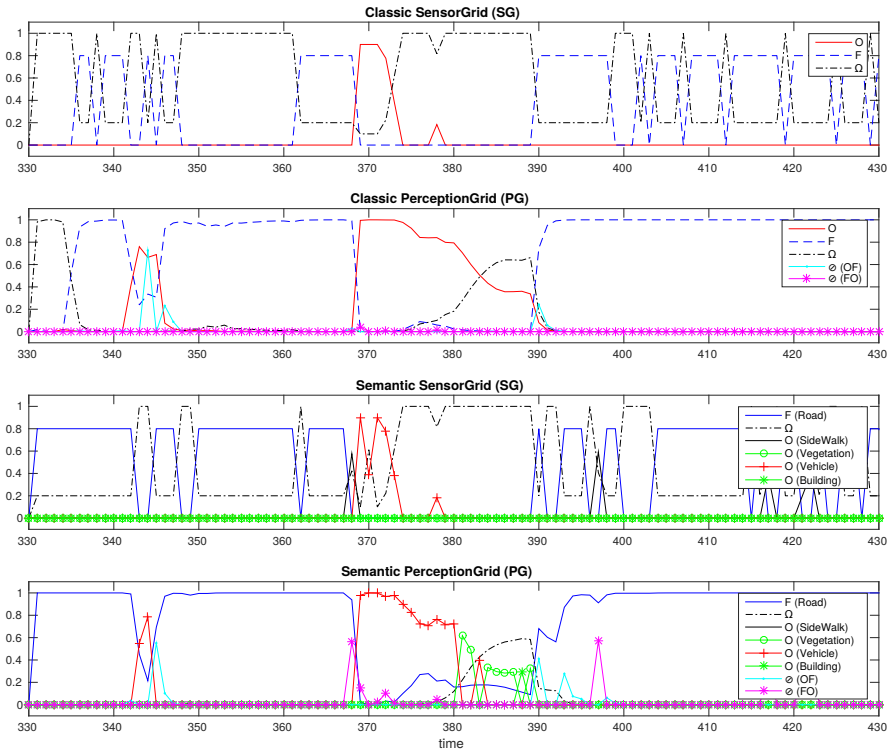

Figure 17: Temporal analysis of belief masses regarding the influence of the phenomenon that happens at the fusion and upgrade processes, concerning the refined propositions management by the voting-based approach usage.

technique that does not require inertial sensors, laser sensors or prior digital mapping to implement a robust perception system. Moreover, the new inverse sensor model considers uncertainties in distance measurements and improves the occupancy grid with associated meta-knowledge. Finally, using the Dempster-Shafer Theory, the prediction and updating processes are modeled to combine semantic context in order to discriminate static and mobile objects in the scene, making this solution a promising approach for urban scene understanding.

With regard to the prospects for the dynamic local perception system, some issues were observed. The first has to do with the formalism employed to manage the meta-knowledge associated with belief masses. Currently, the proposed method uses a voting-based principle, which is not entirely suitable for this purpose. An improvement might be to use a probabilistic or evidence formalism

to upgrade and merge this information. The second issue has to do with the temporal information propagation in the grid. The mechanism of contextual discounting may be used in order to represent the variation in information on the lifetime of objects present in the environment.

Another issue concerns a meaningful evaluation comparing the responses obtained by the Dynamic Local Perception using the DST and BT strategies. For future works, it is intended to verify the standard occupancy grids with respect to the standard evidential grids as well as the applicability of the meta-knowledge information in these two kinds of occupancy grids. To ends, a deep study will be conducted to understand and develop an approach that preserves the meta-knowledge clusters on the occupancy grid and exploits these clusters in higher levels of fusion as for example the object and situation assessment.

Acknowledgements

The authors would like to acknowledge the support granted by CAPES and CNPq - processes PDSE:9129/12-0 and SWE:209656/2013-1.

References

[1] Meilland M, Comport Andrew I, Rives P. Dense omnidirectional RGB-D mapping of large scale outdoor environments for real-time localisation and autonomous navigation. Journal of Field Robotics 2014;URL: `http://hal.inria.fr/hal-01010429`.

[2] Moras J. Grilles de perception évidentielles pour la navigation robotique en milieu urbain. These; Université de Technologie de Compiègne; 2013. URL: `http://tel.archives-ouvertes.fr/tel-00866300`.

[3] Montemerlo M, Thrun S, Koller D, Wegbreit B. FastSLAM: A factored solution to the simultaneous localization and mapping problem. In: Proceedings of the AAAI National Conference on Artificial Intelligence. Edmonton, Canada: AAAI; 2002,.

[4] Kalman RE. A new approach to linear filtering and prediction problems. Transactions of the ASME, Journal of Basic Engineering 1960;82:35–45.

[5] Julier S, Uhlmann J. Unscented filtering and nonlinear estimation. Proceedings of the IEEE 2004;92(3):401–22. doi:10.1109/JPROC.2003.823141.

[6] Shalom Y, Blair W, University of California LAUE. Multitarget/Multisensor Tracking: Applications and Advances. Multitarget-multisensor Tracking; Artech House, Incorporated; 2000. ISBN 9781580530910. URL: http://books.google.fr/books?id=-QB0RQAACAAJ.

[7] Hähnel D, Burgard W, Wegbreit B, Thrun S. Towards lazy data association in SLAM. In: Proceedings of the 11th International Symposium of Robotics Research (ISRR'03). Sienna, Italy: Springer; 2003,.

[8] Petrovskaya A, Thrun S. Model based vehicle detection and tracking for autonomous urban driving. Autonomous Robots 2009;26(2-3):123–39. URL: http://dx.doi.org/10.1007/s10514-009-9115-1. doi:10.1007/s10514-009-9115-1.

[9] Fayad F, Cherfaoui V. Tracking objects using a laser scanner in driving situation based on modeling target shape. In: Intelligent Vehicles Symposium, 2007 IEEE. 2007, p. 44–9. doi:10.1109/IVS.2007.4290089.

[10] Lin KH, Wang CC. Stereo-based simultaneous localization, mapping and moving object tracking. In: Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on. 2010, p. 3975–80. doi:10.1109/IROS.2010.5649653.

[11] Elfes A. Using occupancy grids for mobile robot perception and navigation. Computer 1989;22(6):46–57. doi:http://dx.doi.org/10.1109/2.30720.

[12] Herrmann D, Kamphans T, Langetepe E. Exploring simple triangular and hexagonal grid polygons online. CoRR 2010;abs/1012.5253.

[13] Elfes A. A tesselated probabilistic representation for spatial robot perception and navigation. In: In Proceedings of the NASA Conference on Space Telerobotics; vol. 2. JPL, California Inst. of Tech; 1989, p. 341–50.

[14] Elfes A. Occupancy grids: A stochastic spatial representation for active robot perception. In: Iyengar SS, Elfes A, editors. Autonomous Mobile

Robots: Perception, Mapping, and Navigation (Vol. 1). Los Alamitos, CA: IEEE Computer Society Press; 1991, p. 60–70.

[15] Bourgault F, Makarenko A, Williams S, Grocholsky B, Durrant-Whyte H. Information based adaptive robotic exploration. In: Intelligent Robots and Systems, 2002. IEEE/RSJ International Conference on; vol. 1. 2002, p. 540–545 vol.1. doi:10.1109/IRDS.2002.1041446.

[16] Thrun S, Burgard W, Fox D. A probabilistic approach to concurrent mapping and localization for mobile robots. Autonomous Robots 1998;5(3-4):253–71. URL: http://dx.doi.org/10.1023/A%3A1008806205438. doi:10.1023/A:1008806205438.

[17] Steux B, El Hamzaoui O. tinyslam: A slam algorithm in less than 200 lines c-language program. In: Control Automation Robotics Vision (ICARCV), 2010 11th International Conference on. 2010, p. 1975–9. doi:10.1109/ICARCV.2010.5707402.

[18] Levinson J, Thrun S. Robust vehicle localization in urban environments using probabilistic maps. In: Robotics and Automation (ICRA), 2010 IEEE International Conference on. 2010, p. 4372–8. doi:10.1109/ROBOT.2010.5509700.

[19] Coué C, Pradalier C, Laugier C, Fraichard T, Bessiere P. Bayesian Occupancy Filtering for Multitarget Tracking: an Automotive Application. International Journal of Robotics Research 2006;25(1):19–30. URL: http://hal.inria.fr/inria-00182004; voir basilic : http://emotion.inrialpes.fr/bibemotion/2006/CPLFB06/.

[20] Gate G. Reliable perception of highly changing environments: Implementations for car-to-pedestrian collision avoidance systems. These; Ecole Nationale Supérieure des Mines de Paris; 2009. URL: http://tel.archives-ouvertes.fr/docs/00/50/14/59/PDF/These_Gate.pdf.

[21] Miyasaka T, Ohama Y, Ninomiya Y. Ego-motion estimation and moving object tracking using multi-layer lidar. In: Intelligent Vehicles Symposium, 2009 IEEE. 2009, p. 151–6. doi:10.1109/IVS.2009.5164269.

[22] Xie J, Nashashibi F, Parent M, Favrot O. A real-time robust global localization for autonomous mobile robots in large environments. In: Control Automation Robotics Vision (ICARCV), 2010 11th International Conference on. 2010, p. 1397–402. doi:10.1109/ICARCV.2010.5707329.

[23] Borenstein J, Koren Y. Histogramic In-Motion Mapping for Mobile Robot Obstacle Avoidance. IEEE Transactions on Robotics and Automation 1991;7(4):535–9. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.8.1317.

[24] Bayes M, Price M. An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, f. r. s. communicated by mr. price, in a letter to john canton, a. m. f. r. s. Philosophical Transactions (1683-1775) 1763;53:pp. 370–418. URL: http://www.jstor.org/stable/105741.

[25] Thrun S, Burgard W, Fox D. Probabilistic Robotics. The MIT Press; 2005.

[26] Dempster AP. A generalization of bayesian inference. Journal of the Royal Statistical Society 1986;30(1):205–47.

[27] Shafer G. A Mathematical Theory of Evidence. Limited paperback editions; Princeton University Press; 1976.

[28] Smets P, Kennes R. The transferable belief model. Artif Intell 1994;66(2):191–234. URL: http://dblp.uni-trier.de/db/journals/ai/ai66.html#SmetsK94.

[29] Moras J, Cherfaoui V, Bonnifait P. Credibilist occupancy grids for vehicle perception in dynamic environments. In: Robotics and Automation (ICRA), 2011 IEEE International Conference on. 2011, p. 84–9.

[30] Moras J, Cherfaoui V, Bonnifait P. Moving objects detection by conflict analysis in evidential grids. In: Intelligent Vehicles Symposium (IV), 2011 IEEE. 2011, p. 1122–7.

[31] Moras J, Rodriguez F, Drevelle V, Dherbomez G, Cherfaoui V, Bonnifait P. Drivable space characterization using automotive lidar and georeferenced map information. In: Intelligent Vehicles Symposium (IV), 2012 IEEE. 2012, p. 778 –83. doi:10.1109/IVS.2012.6232252.

[32] Kurdej M, Moras J, Cherfaoui V, Bonnifait P. Controlling Remanence in Evidential Grids Using Geodata for Dynamic Scene Perception. International Journal of Approximate Reasoning 2014;55(1):355–75.

[33] Kurdej M, Moras J, Cherfaoui V, Bonnifait P. Map-aided evidential grids for driving scene understanding. IEEE Intell Transport Syst Mag 2015;7(1):30–41. URL: http://dx.doi.org/10.1109/MITS.2014.2352371. doi:10.1109/MITS.2014.2352371.

[34] Geiger A, Ziegler J, Stiller C. Stereoscan: Dense 3d reconstruction in real-time. In: IEEE Intelligent Vehicles Symposium. Baden-Baden, Germany; 2011,.

[35] Dempster AP. A generalization of bayesian inference. Journal of the Royal Statistical Society 1968;30(B):205–47.

[36] Pagac D, Nebot E, Durrant-Whyte H. An evidential approach to map-building for autonomous vehicles. Robotics and Automation, IEEE Transactions on 1998;14(4):623–9.

[37] Yang T, Aitken V. Evidential mapping for mobile robots with range sensors. Instrumentation and Measurement, IEEE Transactions on 2006;55(4):1422–9. doi:10.1109/TIM.2006.876399.

[38] Canas J, Matelln V. Dynamic gridmaps: comparing building techniques. Mathware and Soft Computing 2006;XIII(1):5–22. URL: http://gsyc.es/jmplaza/papers/gridmaps2006.pdf.

[39] Yager RR, Kacprzyk J, Fedrizzi M, editors. Advances in the Dempster-Shafer Theory of Evidence. John Wiley & Sons, Inc.; 1994. ISBN 0-471-55248-8.

[40] Nguyen TN, Michaelis B, Al-Hamadi A, Tornow M, Meinecke M. Stereo-camera-based urban environment perception using occupancy grid and object tracking. Intelligent Transportation Systems, IEEE Transactions on 2012;13(1):154–65. doi:10.1109/TITS.2011.2165705.

[41] Faugeras O. Three-dimensional computer vision: A geometric view point. Cambridge: MIT Press; 1993.

[42] Bresenham JE. Algorithm for computer control of a digital plotter. IBM Systems Journal 1965;4(1):25–30. doi:10.1147/sj.41.0025.

[43] Geiger A, Lenz P, Stiller C, Urtasun R. Vision meets robotics: The kitti dataset. International Journal of Robotics Research (IJRR) 2013;.

[44] Giovani BV, Victorino AC, Ferreira JV. Stereo vision for dynamic urban environment perception using semantic context in evidential grid. In: 2015 IEEE 18th International Conference on Intelligent Transportation Systems. 2015, p. 2471–6. doi:10.1109/ITSC.2015.398.

[45] Hirschmuller H. Stereo processing by semiglobal matching and mutual information. IEEE Transactions on Pattern Analysis and Machine Intelligence 2008;30(2):328–41. doi:10.1109/TPAMI.2007.1166.

[46] Clemens J, Kluth T, Reineking T. β-slam: Simultaneous localization and grid mapping with beta distributions. Information Fusion 2019;52:62 – 75. URL: http://www.sciencedirect.com/science/article/pii/S1566253516301579. doi:https://doi.org/10.1016/j.inffus.2018.11.005.

[47] Vitor GB. Dynamic evidential grid using semantic context - kitti database. 2014. URL: http://youtu.be/H_zJjX8uMtI; accessed on: 06/12/2018.