2020 IEEE Intelligent Transportation Systems Conference (ITSC)
Rhodes, Greece. September 20-23, 2020 (Virtual)

Online Road Model Generation From Evidential Semantic Grids

Julian Thomas, Julian Tatsch, Alois Knoll and Raúl Rojas

Abstract— **The knowledge about the local environment is of utmost importance for all robotics and autonomous driving applications. Currently this information is often extracted from high definition maps used in combination with a highly-accurate localization restricting the operation to prior mapped areas and making it vulnerable to changes in the environment. On the other hand, occupancy grids, a well known representation of the static environment, provide a common way to model the environment with online sensor measurement data collected during the operation. However, the approach is limited to static occupancy probabilities without further classification or differentiation.**

This paper addresses the topic of estimating the local static environment solely from online sensor measurements by using an evidential semantic grid. Based on the Dempster-Shafer theory and a novel frame of discernment, sensor measurements, such as lane markings, point clouds from image-based semantic segmentation, occupancy grids and observed traffic participants are fused into an evidential grid estimating the *semantic* meaning of each grid cell. Afterwards, an online road model is generated by extraction lane geometries from the evidential grid. Real sensor data from German highways and urban areas is used to show the effectiveness of the proposed approach.

I. Introduction

Modeling and estimating the environment of a vehicle is one of the major challenges for autonomous driving. A common approach for building the environment model is the utilization of offline high definition maps, which contain all static information of the environment. Thus, it is only necessary to estimate the dynamic parts of the environment like information about other traffic participants and traffic light states. However, using offline maps restricts the operational domain of autonomous vehicles to mapped areas and makes them vulnerable to temporary or permanent changes of the environment (e.g. construction sites). Additionally, the information in the map can only be used in combination with a highly accurate self-localization, which in complex dynamic urban environments still proves to be challenging. In contrast to offline high definition maps one can aim for a local environment model built solely using sensor measurements from the autonomous vehicle to overcome the aforementioned issues. Therefore, in this paper, we propose such an local environment model consisting of evidential semantic grids to model the static part (Lane Surface, Lane Markings, Lane Boundaries, Sidewalks) of the vehicle's environment using measurement data from different sensors. The sensors are mounted on an fully autonomous driving prototype vehicle.

The remainder of the paper is structured as follows: In Sec. II related work about grid mapping, semantic segmentation and fusing different sensor modalities in grids is discussed. In Sec. III the proposed evidential grid framework is presented. Generating the road model from the semantic grid is described in Sec. IV and Sec. V shows some results of the evidential grids as well as the generated road model. Sec. VI provides concluding remarks and points out future work.

II. Related Work

Although the overall problem of estimating an environment model based solely on sensor data measurements is not very well studied yet, several contributions solving subproblems exist - mainly focusing on detection and classification of lane markings and road boundaries. In [1] a stereo camera is used to extract 3D lane marking information out of camera images for advanced driver assistant systems (ADAS). In [2] a complete road model is estimated based on lane markings detected by a camera system. But lane markings are sometimes hard to detect (e.g. in construction sites), or in the case of smaller urban streets not present at all.

Regarding the more general problem of building up the static environment model, there exists a broad range of solutions based on classical static occupancy grids. The original idea of storing occupancy probabilities in grid maps was first presented in [3] and has been extended in several directions. For example, [4] and [5] propose a method to distinguish between static and dynamic parts of the environment on a grid-cell level. A particle filter is used to estimate the state of occupancy (occupied by a static or a dynamic obstacle). In [6] the combination of a static occupancy grid and a particle filter is used to cluster individual grid cells, resulting in static and dynamic objects represented by oriented bounding boxes. [7] describe an alternative approach of extending classical static occupancy grids. Here, LiDAR measurements are fused with a-priori knowledge from an offline map in a grid to model the static environment. The differentiation between static objects and other (dynamic) traffic participants is done by analyzing the conflict between the past and the current cell's states (the cell's states change from free to occupied and vice versa). Other approaches based on conflict analysis of cell states can be found in [8] and [9]. In [10] the concept of a static occupancy grid is extended to include more information about the type of occupancy. A grid is generated

J. Thomas and J. Tatsch are with the Autonomous Driving Project at BMW Group AG, Munich, Germany and contributed equally to the work (julian.thomas@bmw.de, julian.tatsch@bmw.de).

A. Knoll is with the Department of Informatics, Technical University of Munich, 85748, Garching b. München, Germany (knoll@in.tum.de).

R. Rojas is with the Institut für Informatik, Freie Universität at Berlin, 14195, Berlin, Germany (raul.rojas@fu-berlin.de).

978-1-7281-4148-0/20/$31.00 ©2020 IEEE

from radar measurements and then enhanced with a cell-wise classification by a neural network trained to classify each occupied cell as *car*, non-car or *unlabeled*.

In situations where lane markings are not available, at least road boundaries can be detected by camera, Radar or LiDAR measurements (e.g. [11]). Most of the existing road boundary estimation approaches [11]–[13] rely on an occupancy grid built from LiDAR or Radar measurements to extract the road boundaries. Unfortunately, road boundaries only delimit *free* areas from *occupied* areas without their semantic meaning. In that case individual lanes cannot be estimated robustly. For example, the *free* area in-between the left and right road boundary can be considered as drivable. However, if the area is divided into several lanes by lane markings painted on the ground, the semantic meaning of the drivable area is neither reflected by the road boundaries nor by the underlying occupancy grid. Another important information missing in occupancy grids is the type of the road boundary. For maneuver- and trajectory-planning the type of the boundary should be taken into account. For example, if the lane is delimited by a sidewalk, the desired behavior of the autonomous vehicle should be more conservative than when driving along a crash barrier where no pedestrians can cross the lane. Other examples for the need of semantic information is the classification of (free) parking space, bus- and bicycle lanes. These areas are often part of the road and drivable in the sense that there is no static obstacle. However, they should not be used for driving by default, because of their semantic meaning. Semantic meaning is usually difficult to extract from LiDAR or Radar measurements alone. However, semantic segmentation methods with convolutional neural networks (CNNs) are well suited for this task. The publication of the Cityscapes dataset [14] for the first time enabled training for a wide variety of classes relevant to autonomous driving, such as road, sidewalk, parking, rail track, vegetation and buildings (e.g. [15]–[18]). To be truly useful for autonomous driving adding spatial information for the segmentation is essential. It can typically measured explicitly by ranging sensors such as Radar, LiDAR, stereo cameras [19], [20] or inferred implicitly in neural network architectures e.g. [21]. In [19] the spatial information is fused with the information from semantic segmentation in an occupancy grid. However, only the states *free*, *occupied*, *unknown* and *conflict* are modeled, thus, missing a more detailed modeling of the environment.

To summarize, existing approaches for estimating the local static environment based on lane marking detection, static/dynamic occupancy grid mapping or road boundary estimation may work reasonably well in some highway scenarios but are unable to correctly model the static environment in urban scenarios. This is mostly because the information about drivable and non-drivable areas is insufficient. Regarding grid-based environment modeling existing approaches focus on either road boundary estimation or detecting/extracting dynamic objects (traffic participants). However, semantic grid-based modeling of the static environment is not well studied yet.

Compared to the existing state of the art, the contribution of this paper is as follows: We propose a novel evidential grid mapping approach to model the static part of the environment including its semantic meanings. Therefore, arbitrary information about the road, individual lanes, unclassified static obstacles, free space and pedestrian sidewalks from various sensors can be fused in an evidential grid-based representation. To the best of our knowledge this is the first time building a consistent environment model representing all relevant static parts of the vehicle's surroundings without using a-priori information. Because of the use of the Dempster-Shafer theory, our approach is easily extensible when a more detailed classification (like the distinction between different types of static obstacles) becomes available. Furthermore, we describe a novel approach for projecting semantic segmentation information into grids taking the position and classification uncertainty into account.

III. Evidential Grid Framework

In contrast to classical occupancy grids used in e.g. [12], [13] this work is based on the theory of belief functions proposed by A. Dempster and reformulated later by G. Shafer [22]–[24]. The Dempster-Shafer theory (DST) allows to calculate the belief in a specific hypothesis taking all available evidence from different sources into account. In the DST the *frame of discernment* (FOD) is defined as a set Ω, which elements represent all possible states/hypotheses θ_i of the system under consideration:

$$\Omega = \{\theta_1, \theta_2, ..., \theta_N\} \tag{1}$$

Note, that the elements θ_i must be mutually exclusive and exhaustive, meaning that at least one hypothesis must be true. A *basic belief assignment function* (BBA) is used to assign belief masses not only to a single hypothesis θ_i but to any subset of Ω. 2^Ω is defined as the set of all subsets of Ω including the empty set $\varnothing$:

$$2^\Omega := \{U \mid U \subseteq \Omega\} \tag{2}$$

The BBA itself is defined as

$$m : 2^\Omega \rightarrow [0, 1] \tag{3}$$

with the following properties

$$m(\varnothing) = 0 \tag{4}$$

$$\sum_{A \in 2^\Omega} m(A) = 1 \tag{5}$$

The belief mass of an element $A \in 2^\Omega$, written as $m(A)$, is the proportion of all available evidence implying exactly A is true, but no particular subset of A. In contrast to that, the belief in A, *Bel(A)*, is defined as the sum of the masses of all subsets of A including A itself:

$$Bel(A) = \sum_{B \subseteq A} m(B) \tag{6}$$

It is the amount of evidence that either the given hypothesis-set A or one of its subsets is true. Belief functions are combined by combining their respective BBAs. This results in a new belief function, which includes the knowledge/evidences of both belief functions. Let Bel_1 and Bel_2 two different belief functions over the same frame of discernment, and m_1 and m_2 their corresponding BBAs, then their BBAs can be combined as follows:

$$m(A) = \frac{1}{1-k} \sum_{A_i \cap B_j = A} m_1(A_i) m_2(B_j), \; A \neq \varnothing \tag{7}$$

$$m(\varnothing) = 0 \tag{8}$$

with

$$k = \sum_{A_i \cap B_j = \varnothing} m_1(A_i) m_2(B_j) \tag{9}$$

Eq. 7 is also called *Dempster's Rule of Combination*. k is a measure for the amount of conflict between m_1 and m_2.

In contrast to all previous work, e.g. [5], [9], [19], our frame of discernment is defined as follows

$$\Omega = \{L, M, S, O\} \tag{10}$$

where L stands for *Lane*, M for *Marking*, S for *Sidewalk* and O for *Obstacle*. We define a lane as the drivable area between the left and right lane boundaries, which are not necessarily the same as the road boundaries. Markings are the white/yellow lane markings painted on the ground. The hypothesis *Sidewalk* includes the boundary between the lane and the sidewalk (curbstone) as well as the area of the sidewalk itself. *Obstacle* represents any static obstacle which is not passable without a collision.

From the power set 2^{Ω} we only use a "reduced" power set

$$\begin{aligned} 2_r^{\Omega} = \{&L, M, S, O, \{S, O\}, \{M, L\}, \{L, M, S\}, \\ &\{M, S, O\}, \{L, O\}, \varnothing, \Omega\} \end{aligned} \tag{11}$$

since in our case the masses for all other sets are not directly measurable with our current sensor setup. However, if some of these masses become measureable, e.g. due to an improved sensor setup, more elements from 2^{Ω} to 2_r^{Ω} can easily be added without changing the overall approach presented here at all. Using the established Dempster-Shafer Theory and the above defined reduced power set, measurement data from different types of sensors can now be fused and accumulated in an evidential semantic grid, which is described in the following section.

A. Grid-based Sensor Fusion

A grid is a multidimensional lattice with equally-sized, quadratic cells, each cell storing stochastic information m inferred from sensor measurements [25]. For computational tractability, it is moreover assumed that all cells are conditionally independent of each other, allowing the parallel computation of all cells. For the sake of readability we define $\mathbf{m}(\cdot)$ as a grid containing for each cell belief masses for all elements in 2_r^{Ω}.

All beliefs in the grid $\mathbf{m}$ can be computed recursively, for details the reader is referred to [26]. $m^i(\cdot)$ denotes the belief masses in the cell with index i, which is omitted in the following for the sake of better readability.

Measurements from detected lane markings, information about dynamic objects (other traffic participants), a point cloud generated from a stereo-vision system combined with image-based semantic segmentation, and occupancy grids are now used to infer belief masses for the hypotheses contained in 2_r^{Ω}. For each information source, a belief grid is created containing - from this specific source - the data from a single time step only. By analogy with the expression LiDAR or radar "scan", which means the process of getting measurement data from a single time step, we name these belief grids "ScanGrid" and define it as $\mathbf{m}_t^{\text{source}}(\cdot)$. The following subsections describe the considerations necessary for projecting the various types of measurements into their corresponding grids.

*1) **Lane Markings**:*

Obviously, the white/yellow lane markings painted on the ground directly provide semantic information about the static environment. They subdivide the free, drivable area into different subregions based on the type of the marking. At least in Germany the lane markings are classified as either *lane dividers* or *road edge markings* or *lane edge markings*. The semantic meaning of lane dividers is to separate individual lanes, road edge lines separate the road (tarred area) from non-road regions like grass stripes. Lane edge lines are mostly found on highways only since they separate the most right lane from the emergency lane.

In addition to the ability to detect the physical appearance of lane markings, state-of-the-art vision-based recognition systems provide information about the type of the marking (e.g. continuous line, dashed line) as well as the width and its color. Especially in construction sites, the color of the markings makes a significant difference regarding the semantic meaning. If there are white and yellow markings present, the yellow ones take precedence over the white ones.

In the present case lane markings are provided by an automotive camera system and continuously delivers a set of detected markings $\mathcal{M}$

$$\mathcal{M} = \{\mathcal{M}_1, \dots, \mathcal{M}_n\} \tag{12}$$

$$\mathcal{M}_i = \left[\{\mathbf{x}_1, \dots, \mathbf{x}_n\}, k, w, c, \sigma, p\right]^T \tag{13}$$

where $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ is a set of points (x, y) describing the marking's geometry in 2D and k the type of the marking. It can be one of the following: (single or double) solid line, (single or double) dashed line, sidewalk, road edge or unknown. w is the width of the marking, c its color, σ a value reflecting the measurement uncertainty and p the existence probability.

To infer belief masses for the hypotheses in 2_r^{Ω} for each grid cell we use a BBA defined as a convolution of a Gaussian distribution with a rectangular function. The BBAs are the equivalent to an inverse sensor model $p(m|\mathbf{z}_t)$ given a new sensor measurement $\mathbf{z}_t$ in Bayesian occupancy probability approach (see e.g. [13]). The BBA is normalized to values in $[0;1]$ to fulfill Eq. (3). A grid cell containing a belief mass of 0 for a specific hypothesis A means that there is no evidence, that the cell's state is A, whereas 1.0 indicates that the state is truly A.

Depending on the hypothesis and the type k of the marking the BBA is used in different ways to derive the masses to be written into the grid cells. In the case of a marking classified as solid or dashed line the mass for the hypothesis *Marking*, $m(M)$ can be inferred. For this, the mean of the Gaussian distribution is positioned in the middle of the marking at each position $\mathbf{x}_i$ orthogonal to the orientation of the marking. The σ of the Gaussian is directly taken from $\mathcal{M}_i$. The Gaussian is convolved with a rectangular function to take the area of the marking on the ground into account. The width of the rectangular function equals the width w of the marking from $\mathcal{M}_i$. The same procedure is applied when the type of the marking is sidewalk or road edge. In this case, the mass is either allocated to $\{S\}$ or $\{O\}$ in case of a road edge. If the type of the marking is unknown, the mass is given to $\{M, S, O\}$. Another deduction can be drawn using the fact that the frame of discernment is per definition complete - meaning that the frame has to be defined in a way that at least one hypothesis in Ω is true. Therefore, for all regions, where no lane markings or sidewalks were detected, the belief of $m(\neg(M \vee S)) = m(\{L, O\})$ can be inferred, as long as they are covered by the field of view of the camera.

By taking the semantic meaning of the marking's type into account a more optimistic estimate can be used, too. Since markings painted on the ground split the drivable area into individual lanes, they implicitly define the existence of lanes. Although a dashed line is a lane boundary, crossing it is allowed. Thus, there must exist a lane on both sides of the marking. In contrast, for solid lines, the existence of a lane can only safely be assumed on the vehicle-facing side of the marking. Eventually, to calculate the belief mass for lane, $m(L)$, for each marking $\mathcal{M}_i$ the mean of the Gaussian is shifted to the left/right by half of a typical lane width depending on the type of the marking. The width of the rectangular function corresponds to the lane width used for the Gaussian.

Depending on the existence probability p of $\mathcal{M}_i$, the more conservative estimate (assigning belief mass to $m(\{L, O\})$) or the optimistic one (assigning it to $m(L)$) is chosen. This way, the decision about the conservative or optimistic estimate is automatically deduced from the reliability of the measured lane marking leading to a more precise estimate ($m(\{L\})$) in case of reliably detected lane markings. Currently, this decision is made based on a predefined threshold value for p.

Since the lane marking sensor measurements are a snapshot of a single time step t, we denote the grid containing the inferred masses as $\mathbf{m}_t^{\text{LM}}(\cdot)$. The upper index stands for the source of the information (lane marking).

*2) **Dynamic Objects**:*

The paths of other road users are a valuable source of information for environment model estimation, too. Under the assumption that most vehicles drive on valid lanes, evidence for lanes and lane boundaries can be derived by means of appropriate BBAs.

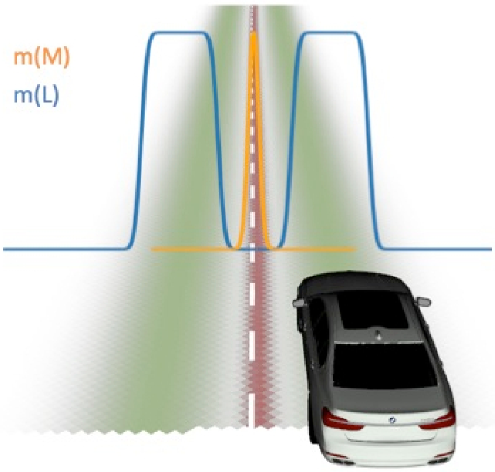


Fig. 1. Situation with a single marking measured by the lane marking detection system. A dashed marking on the left of the ego vehicle was detected. On top, the BBAs for *Marking*, $m(M)$ (orange) and for *Lane*, $m(L)$ (blue), are shown. Below, the resulting grid is displayed, where the color denotes the hypothesis ($\{M\}$: red; $\{L\}$: green). The belief mass is encoded in the alpha channel of the grid (transparency).

Here, dynamic objects are extracted from evidential dynamic occupancy grids and tracked over time. The approach is described in [27] and provides a set of detected dynamic objects $\mathcal{O}$

$$\mathcal{O} = \{\mathcal{O}_1, \ldots, \mathcal{O}_n\} \tag{14}$$

$$\mathcal{O}_i = \left[x, y, \phi, l, w, k\right]^T \tag{15}$$

where each object is modeled as a rectangular, oriented bounding box with length l and width w. (x, y) is the 2D position of the box and ϕ the orientation. k is the type of the dynamic object.

A convolution of a Gaussian distribution with a rectangular function is applied in the same manner as in the case of lane markings to infer the belief masses for the grid cells. The width of the object's bounding box, w, is used as width for the rectangular function. In case of the belief mass for a lane, $m(L)$, for each object $\mathcal{O}_i$ the mean of the Gaussian is placed at the object's box position (x, y), orthogonal to the orientation ϕ of the box and the resulting masses are written into the underlying grid cells. Following the above mentioned assumption that most vehicles drive on valid lanes it is possible to infer lane boundaries on the left and right side of each object. The corresponding mass is assigned to $m(\{M, S, O\})$ since there is no evidence what boundary it is exactly. The resulting grid containing the belief masses is denoted by $\mathbf{m}_t^{\text{Obj}}(\cdot)$ from now on, since it only holds measurement data from a single time step t.

*3) **Semantic Segmentation**:*

Because of the huge progress made in recent years in the field of (fully) convolutional neural networks, semantic segmentation is nowadays widely used for environment perception and scene understanding in robotics and autonomous driving applications. Semantic segmentation aims to annotate every pixel in a camera image with a certain class label. For the approach presented in this paper, we use a re-implementation of PSPNet [17].

By using a stereo-vision system, combining semantic segmentation and disparity calculation (e.g. via Semi-Global Matching [28]) a semantic point cloud containing semantic and spatial information can be generated. Our semantic segmentation system continuously delivers a point cloud $\mathcal{P}$ with points $\mathcal{P}_i$

$$\mathcal{P} = \{\mathcal{P}_1, \dots, \mathcal{P}_n\} \tag{16}$$

$$\mathcal{P}_i = \left[x, y, z, k, \sigma_k, \sigma_z\right]^T \tag{17}$$

where (x, y, z) is the 3D position of the point, k the class label, σ_k the class confidence and σ_z the depth confidence, which is directly obtained from the stereo disparity calculation. We merge classes like building, vegetation, wall and guard rail and treat them as static obstacles, because they have the same semantic meaning with respect to the drivability. In Fig. 2 such a semantic point cloud overlaid on top of a high definition offline map (ground truth) is shown.

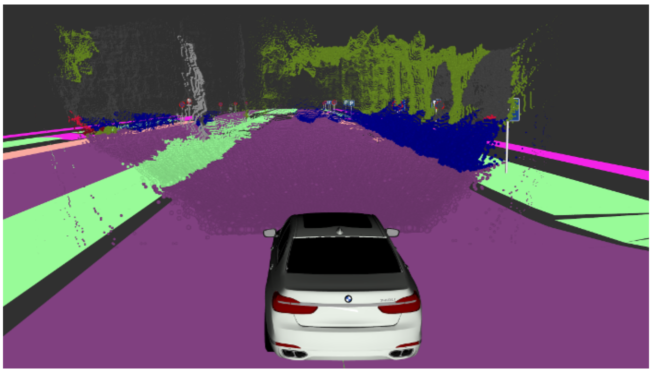


Fig. 2. A real-world example of the semantic point cloud overlaid over the HD offline map. The color of each pixel indicates its class affiliation. The color coding of the HD map is adjusted to match the color coding of the semantic segmentation

To infer belief masses from the semantic point cloud and project them into our grid representation, the classes are mapped to their corresponding hypotheses groups from 2_r^Ω:

$\{M, L\} \leftarrow$ road

$\{S\} \leftarrow$ sidewalk

$\{O\} \leftarrow$ wall, fence, building, vegetation, ...

Each point $\mathcal{P}_i$ resembles a pixel P in the segmented image. Knowing the pixel location, the camera intrinsics and extrinsics each pixel can be projected to a *projected pixel area* A_{px} on the ground plane. A simplified picture of the situation is depicted in Fig. 3.

with

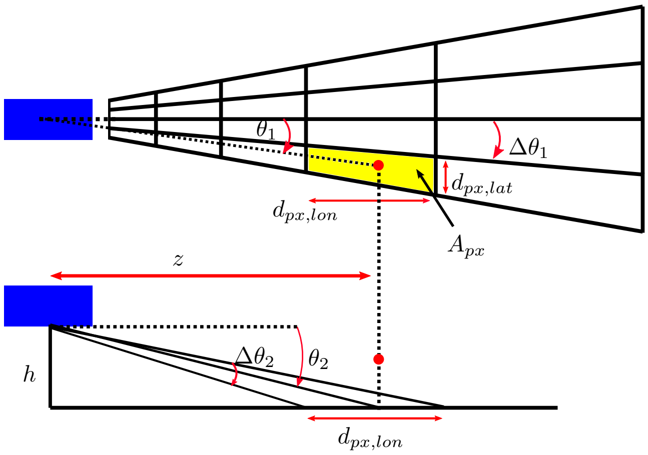


Fig. 3. Projection of a pixel P (red dot) onto the ground plane. Top view (top) and side view (bottom).

θ_1: lateral pixel location, directly calculated from pixel coordinate and focal length. Because the location of P is already available, it can be calculated using $atan(z/x)$.

θ_2: vertical ground projection angle.

z: horizontal distance of P to the camera.

h: height of the camera above ground plane.

$\Delta\theta_{1,2}$: angular pixel size, causing the projected area A_{px} with dimensions $d_{px,lon}$ and $d_{px,lat}$.

Hence, in the ground plane, we approximate the projected pixel area by:

$$A_{px} \approx d_{px,lon} \cdot d_{px,lat} \tag{18}$$

$$d_{px,lon} = \frac{z^2}{h \cdot f_y} \tag{19}$$

$$d_{px,lat} = \frac{z}{f_x} \tag{20}$$

where z is the distance (z-component of $\mathcal{P}_i$), h the height of the camera above ground plane, f_x, f_y the focal lengths in horizontal and vertical direction respectively.

Spatial Uncertainty:
In this context, we define spatial uncertainty (or location probability) as the uncertainty with respect to the projected location of the pixel P to the ground plane. The better part of the spatial uncertainty is caused by the uncertainty of the depth measurement (z). Here, it is modeled as a bi-variate Gaussian distribution around the projected point P on the ground plane. The bi-variate Gaussian distribution is defined with standard deviations in two directions: The major axis tangential to the pixel-ray (depth-direction), the minor axis orthogonal to that in the ground plane (see Fig. 4).

Then, the location probability of point P at each location in the ground plane is calculated as:

$$p_P(u, v) = \frac{1}{2\pi \cdot \sigma_u(P)\sigma_v(P)} \cdot exp\left(-\frac{1}{2}\left(\frac{u^2}{\sigma_u^2(P)} + \frac{v^2}{\sigma_v^2(P)}\right)\right) \tag{21}$$

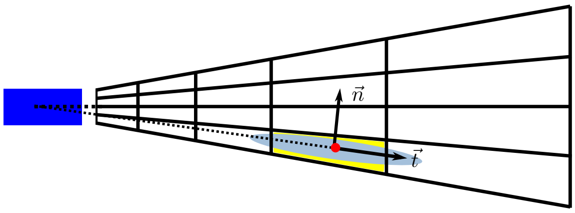


Fig. 4. Spatial uncertainty modeled as bi-variate Gaussian distribution (blue ellipse) with major axis ($\vec{t}$) along the pixel-ray for a Point P (red dot). $\vec{n}$ points into the normal direction.

with u, v the distance to the projected point P along the tangential and normal axis, respectively, and $\sigma_u(P)$, $\sigma_v(P)$ the standard deviation in those directions.

However, the probability density p_P does not denote the probability density of the projected pixel area to be located here. If there were no spatial uncertainty, then the probability density of the projected pixel area would be an uniform distribution with magnitude 1 exactly covering the projected pixel area. The final probability density of the pixel projection with spatial uncertainty is a convolution of both distributions.

The convolution is approximated by

$$p_{pixel,P}(u,v) = \frac{1}{2\pi \cdot \sigma_u \sigma_v} \cdot exp\left(- \frac{1}{2}\left(\frac{u^2}{\sigma_u^2} + \frac{v^2}{\sigma_v^2} \right) \right) \tag{22}$$

$$\sigma_u = \sigma_u(P) + d_{px,lon} \tag{23}$$

$$\sigma_v = \sigma_v(P) + d_{px,lat} \tag{24}$$

Depending on the specific class label, points from the point cloud will add belief mass to one or more of the following hypotheses: $\{M, L\}$, $\{S\}$ and $\{O\}$. The calculation of the belief mass is the same for each hypothesis, but each time, only those points in the point cloud that have the correct class labels are considered for the calculations.

Estimated class confidence at grid cell:
Each point $\mathcal{P}_i$ in the point cloud has an associated class confidence σ_k, so the total class confidence of a grid cell is a fusion of the class confidences of nearby projected pixels. Therefore, a class confidence for each grid cell is calculated and updated by using a weighted average with the location probability as weight:

$$p_{class}(q)_i = \frac{\sum_{P \in P^q} \sigma_k \cdot p_{pixel,P}(i)}{\sum_{P \in P^q} p_{pixel,P}(i)} \tag{25}$$

with q the hypothesis group, i the grid cell index, σ_k the class confidence of point $\mathcal{P}_i$ and P^q all nearby projected pixels for which the corresponding point $\mathcal{P}_i$ has a class label of q.

Estimated location probability at grid cell:
Each point in the point cloud with the class belonging to the correct group will increase the probability of that cell really belonging to one of the projected pixels in the image. The fusion of probabilities due to multiple points at each grid cell is a normal product of inverse probabilities (it is the inverse if the probability that the grid cell does **not** belong to any of the projected areas):

$$p_{pixel}(q)_i = 1 - \Pi_{P \in P^q}(1 - p_{pixel,P}(i)) \tag{26}$$

Final estimated mass at grid cell:
The mass for hypothesis group q at each grid cell i is simply the multiplication of the class confidence and the projected pixel probability

$$m(q)_i = p_{class}(q)_i \cdot p_{pixel}(q)_i \tag{27}$$

As the semantic segmentation is done for every image frame independently, the resulting grid reflects a single time step only. The grids are therefore denoted as $m_t^{\mathrm{SS}}(\cdot)$.

*4) **Occupancy Grids**:*
For the creation of the occupancy grids the approach described in [5] is used. The grids are formulated in the Dempster-Shafer Theory leading to a seamless integration into our semantic grids. The output of the the occupancy grid mapping from [5] consists of a grid $\mathbf{m}_t^{\mathrm{OG}}(\cdot)$ containing per cell the evidences for the two hypotheses *static* and *free*.

To infer belief masses for the semantic grid the two evidences are mapped to their corresponding hypotheses groups from 2_r^{Ω}:

$$\{L, M, S\} \leftarrow \textit{free}$$
$$\{S, O\} \leftarrow \textit{static}$$

B. Grid Fusion

In order to obtain belief masses containing the data from all sources (lane markings, dynamic objects, semantic segmentation, occupancy grids) the grids described in the previous section have to be fused. Additionally, the grids have to be accumulated over time to include all information from the past up to the current time t. Here, the grid-based modeling shows its full potential, since the grids already provide a spatial and temporal connection of the data implicitly. Formally, the fusion of the ScanGrids can be described as

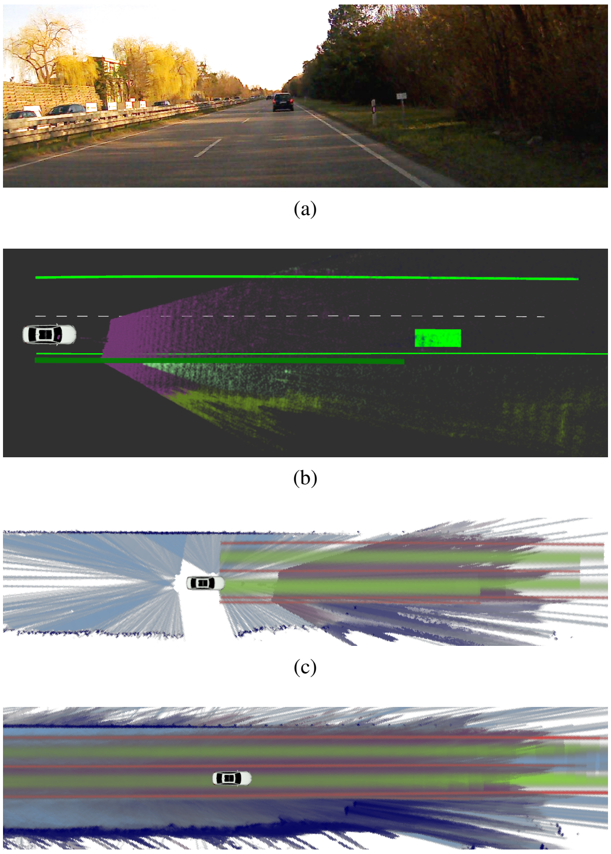
$$\mathbf{m}_t^{\mathrm{all}}(\cdot) = \mathbf{m}_t^{\mathrm{LM}}(\cdot) \oplus^{\mathrm{S}} \mathbf{m}_t^{\mathrm{Obj}}(\cdot) \oplus^{\mathrm{S}} \mathbf{m}_t^{\mathrm{SS}}(\cdot) \oplus^{\mathrm{S}} \mathbf{m}_t^{\mathrm{OG}}(\cdot) \tag{28}$$

This way, the fused ScanGrid $\mathbf{m}_t^{\mathrm{all}}(\cdot)$ contains the information available from all sensors at a single time step t.

For the temporal fusion, the accumulated grid from the previous time step is fused with the current grid $\mathbf{m}_t^{\mathrm{all}}(\cdot)$:

$$\mathbf{m}_{1:t}^{\mathrm{all}}(\cdot) = \mathbf{m}_{1:t-1}^{\mathrm{all}}(\cdot) \oplus^{\mathrm{T}} \mathbf{m}_t^{\mathrm{all}}(\cdot) \tag{29}$$

For the fusion operator $\oplus^{\mathrm{S}}$ Eq. 7 is used and the cumulative fusion rule from [29] for $\oplus^{\mathrm{T}}$. The resulting grid $\mathbf{m}_{1:t}^{\mathrm{all}}(\cdot)$ is denoted as "DSTMap" in the following. Fig. 5 shows the different processing steps for a road with two parallel lanes.



(a)

(b)

(c)

(d)

Fig. 5. (a) Camera image of the situation. (b) Visualization of the sensor data: Solid (green) and dashed lane markings, the detected dynamic object in front of the ego (green box) and the semantic point cloud. (c) Fused ScanGrid $(\mathbf{m}_t^{\text{all}}(\cdot))$. The color indicates the hypothesis (green: $\{L\}$, red: $\{M\}$, dark blue: $\{O\}$, light blue: $\{L, M, S\}$, purple: $\{M, L\}$). The amount of belief (belief mass) is encoded in the transparency value of the pixel. (d) Resulting DSTMap after temporal fusion $(\mathbf{m}_{1:t}^{\text{all}}(\cdot))$.

IV. Road Model Generation

For fulfilling the autonomous driving task a consistent road model including the geometry of lane boundaries and lane center lines is required. In this paper, potential lanes are extracted by using the previously described evidential semantic grid as a cost map for a path planning based approach searching for drivable paths and potential lane boundary points. Since the goal state is unknown a predefined fixed path length is used as goal predicate and a combination of A* and RRT (see [30]) is used for the path planning

Iterative Path Planning: As the fixed path planning length is used as goal predicate, resulting goal paths will only be found if the grid exploration extends beyond this distance. Beyond the explored area in the grid, there is no information about the presence of a lane and therefore it is not possible to assume any path planned in such regions are drivable, and in effect, a goal pose cannot be determined. For this reason, the path planning length should not be set too high. To still enable paths to be found as far as the evidence for the hypothesis *Lane* allows, path planning is performed in an iterative manner, by clustering the goal paths and using the end points of the clustered paths as new starting points for the next path planner iteration. Compared to a direct path planning approach with the same total length, this will also improve the computational efficiency, as the iterative path planner only needs to expand on the end nodes of the clustered paths of the previous iteration. In Figure 6b an example of the iterative path planner is depicted.

Road Model Extraction: Due to the non-deterministic nature of RRT and the limited information in the evidential grid, the clustered paths from the path planner should not be directly used as lane center hypothesis. However, accurate estimates of the right and left boundary points along the clustered paths *can* be obtained, by finding all points in the evidential grid closest to the clustered path that have a high belief for the hypothesis group $\{M, S, O\}$, see e.g. Fig. 6b. Using least squares estimation with smoothness penalties, smooth splines are fitted to the boundary points. Moreover, smooth estimations of lane center line and lane width are determined in Fig. 6c. Additional logic is used to correctly section lanes at branching and merging points and to connect neighboring lanes by association of shared boundaries.

V. Results

The approach was tested and evaluated with real measurement data from urban scenarios in Munich's inner city and on highway-like roads. The autonomous driving test vehicle is equipped with five LiDARs, a lane marking recognition system and a stereo camera. The implementation runs - except the re-implemented PSPNet for image-based semantic segmentation - in real-time on an NVIDIA GTX 1080 GPU.

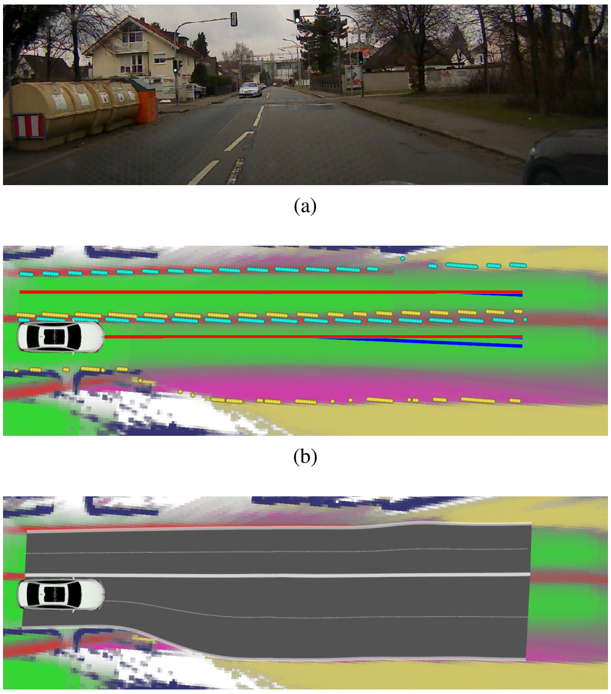
Fig. 6 shows the results for a challenging situation in an urban environment with sidewalks, stationary objects and parked cars. Next to the ego vehicle on the right side the lane is limited by a car parked on the lane. Ahead, the lane is limited by its ”real” boundary (sidewalk) resulting in a non-constant lane width.

VI. Conclusion and Future Work

We presented an evidential framework to generate a consistent road model from online sensor data alone. It consists of a grid-based sensor fusion approach, which is solely based on online sensor data and thus completely independent of any offline map or localization. The fusion approach uses a refined frame of discernment containing all relevant elements of the environment and allows to elegantly fuse evidence and beliefs for subsets of hypotheses in a consistent and coherent manner. In contrast to competing state of the art approaches, our approach has been shown to also work well for more complicated urban scenarios. Additionally, it is easily extensible towards new sensor types.

In future the image-based semantic segmentation will be optimized in order to let the complete system run in real-time. Additionally, an extensive evaluation of all hypotheses on longer tracks in urban scenarios is planned.

The resulting road model is already being used for map validation and online motion planning.



(a)

(b)

(c)

Fig. 6. Traffic situation in an urban environment. (a) Camera image of the situation. In (b) the DSTMap (green: $\{L\}$, red: $\{M\}$, dark blue: $\{O\}$, light blue: $\{L, M, S\}$ and the result of the path planning are visualized. For each lane the goal- (blue) and cluster paths (red) discovered by the path planner are shown. The planner was parameterized to plan two iterations of paths with each 15m length. Extracted points on the left/right lane boundary are depicted as turquoise/yellow dots. (c) Generated road model including the center line (white dots), left/right lane boundaries per lane (white lines) and the lane surface (grey area). The ego lane is limited by a car parked next to the ego car. The boundaries of the extracted lane are correctly identified leading to a more narrow lane at the beginning and a wider lane ahead with a smooth continuous left lane boundary.

References

[1] H. Loose, "Dreidimensionale straßenmodelle für fahrerassistenzsysteme auf landstraßen," Ph.D. dissertation, Institut für Mess- und Regelungstechnik, Karlsruher Institut für Technologie, 2013.

[2] A. Abramov, C. Bayer, C. Heller, and C. Loy, "A flexible modeling approach for robust multi-lane road estimation," in *Intelligent Vehicles Symposium (IV), 2017 IEEE*. IEEE, 2017, pp. 1386–1392.

[3] A. Elfes, "A sonar-based mapping and navigation system," in *Proceedings of the 1986 IEEE International Conference on Robotics and Automation*, vol. 3, 1986, pp. 1151 – 1156.

[4] G. Tanzmeister, J. Thomas, D. Wollherr, and M. Buss, "Grid-based mapping and tracking in dynamic environments using a uniform evidential environment representation," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, May 2014, pp. 6090–6095.

[5] S. Steyer, G. Tanzmeister, and D. Wollherr, "Grid-based environment estimation using evidential mapping and particle tracking," *IEEE Transactions on Intelligent Vehicles*, vol. 3, no. 3, pp. 384–396, Sep. 2018.

[6] R. G. Danescu, "Obstacle detection using dynamic particle-based occupancy grids," in *2011 International Conference on Digital Image Computing: Techniques and Applications*, Dec 2011, pp. 585–590.

[7] M. Kurdej, J. Moras, V. Cherfaoui, and P. Bonnifait, "Map-aided evidential grids for driving scene understanding," *IEEE Intelligent Transportation Systems Magazine*, vol. 7, no. 1, pp. 30–41, Spring 2015.

[8] J. Moras, V. Cherfaoui, and P. Bonnifait, "Credibilist occupancy grids for vehicle perception in dynamic environments," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, may 2011, pp. 84 –89.

[9] M. Kurdej, J. Moras, V. Cherfaoui, and P. Bonnifait, "Map-aided fusion using evidential grids for mobile perception in urban environment," in *Proceedings of the 2nd International Conference on Belief Functions*, ser. Advances in Intelligent and Soft Computing, M.-H. Denoeux, Thierry; Masson, Ed., vol. 164. Compiègne, France: Springer, May 2012, pp. 343–350. [Online]. Available: http://hal.archives-ouvertes.fr/hal-00714465

[10] J. Lombacher, K. Laudt, M. Hahn, J. Dickmann, and C. Wöhler, "Semantic radar grids," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, June 2017, pp. 1170–1175.

[11] G. Tanzmeister, M. Friedl, A. Lawitzky, D. Wollherr, and M. Buss, "Road course estimation in unknown, structured environments," in *Intelligent Vehicles Symposium (IV), 2013 IEEE*, June 2013, pp. 630–635.

[12] M. Konrad, M. Szczot, and K. Dietmayer, "Road course estimation in occupancy grids," in *Intelligent Vehicles Symposium (IV), 2010 IEEE*, june 2010, pp. 412 –417.

[13] F. Homm, N. Kaempchen, J. Ota, and D. Burschka, "Efficient occupancy grid computation on the gpu with lidar and radar for road boundary detection," in *Intelligent Vehicles Symposium (IV), 2010 IEEE*, june 2010, pp. 1006 –1013.

[14] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, 2016.

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[16] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *CoRR*, vol. abs/1606.02147, 2016. [Online]. Available: http://arxiv.org/abs/1606.02147

[17] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2881–2890.

[18] E. Romera, J. M. Álvarez, L. M. Bergasa, and R. Arroyo, "Efficient convnet for real-time semantic segmentation," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, June 2017, pp. 1789–1794.

[19] B. V. Giovani, A. C. Victorino, and J. V. Ferreira, "Stereo vision for dynamic urban environment perception using semantic context in evidential grid," in *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, Sept 2015, pp. 2471–2476.

[20] L. Schneider, M. Cordts, T. Rehfeld, D. Pfeiffer, M. Enzweiler, U. Franke, M. Pollefeys, and S. Roth, "Semantic stixels: Depth is not enough," in *Intelligent Vehicles Symposium (IV), 2016 IEEE*. IEEE, 2016, pp. 110–117.

[21] J. Uhrig, M. Cordts, U. Franke, and T. Brox, "Pixel-level encoding and depth layering for instance-level semantic segmentation," in *German Conference on Pattern Recognition (GCPR)*, 2016. [Online]. Available: http://lmb.informatik.uni-freiburg.de//Publications/2016/BU16

[22] A. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *The Annals of Mathematical Statistics*, vol. 38, no. 2, pp. 325–339, 1967.

[23] ——, "A generalization of bayesian inference," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 205–247, 1968.

[24] G. Shafer, *A mathematical theory of evidence*. Princeton University Press, 1976.

[25] A. Elfes, "Using occupancy grids for mobile robot perception and navigation," *Computer*, vol. 22, no. 6, pp. 46 –57, 1989.

[26] J. Thomas, J. Tatsch, W. Ekeren, R. Rojas, and A. Knoll, "Semantic grid-based road model estimation for autonomous driving," 06 2019, pp. 2329–2336.

[27] S. Steyer, G. Tanzmeister, and D. Wollherr, "Object tracking based on evidential dynamic occupancy grids in urban environments," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, June 2017, pp. 1064–1070.

[28] H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2. IEEE, 2005, pp. 807–814.

[29] A. Jøsang, J. Diaz, and M. Rifqi, "Cumulative and averaging fusion of beliefs," *Information Fusion*, vol. 11, no. 2, pp. 192–200, 2010.

[30] G. Tanzmeister, M. Friedl, D. Wollherr, and M. Buss, "Path planning on grid maps with unknown goal poses," in *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, Oct 2013, pp. 430–435.