Sven Richter*, Sascha Wirges, Hendrik Königshof und Christoph Stiller

Fusion of range measurements and semantic estimates in an evidential framework

Fusion von Distanzmessungen und semantischen Größen im Rahmen der Evidenztheorie

DOI 10.1515/teme-2019-0052

Abstract: A detailed reconstruction of the environment is a crucial component of mobile robotic systems and enables higher level scene understanding. To achieve information redundancy heterogeneous sensors need to be used with each sensor having specific strengths and weaknesses. Therefore, the goal of this work is to fuse information from multiple lidars, radars, a stereo camera and semantic camera information into one common scene representation. In contrast to past publications, we focus on the combination of distance measurements and semantic estimates in the image domain in one common evidential framework. Grid maps are used as common fusion structure which enable efficient data processing. The approach is validated on an automated driving platform in real traffic scenarios. Experiments show that the scene reconstruction precision increases while still retaining the real-time capability.

Keywords: Sensor data fusion, evidence theory, grid maps.

Zusammenfassung: Eine präzise Rekonstruktion der Umgebung ist ein notwendiger Bestandteil mobiler automatisierter Systeme und ermöglicht abstrakteres Szenenverstehen. Um Informationsredundanz zu erhalten, müssen heterogene Sensoren verwendet werden, wobei jeder Sensor spezifische Stärken und Schwächen besitzt. Ziel dieser Arbeit ist es daher, Information von mehreren Lidar-, Radar- und Stereokamerasensoren sowie semantische Kamerainformation in einer gemeinsamen Szenenrepräsentation zu fusionieren. Im Gegensatz zu bisherigen Arbeiten wird hier das Hauptaugenmerk auf die Kombination von Distanzmessungen und semantischen Informationen aus dem

Bildbereich in einem gemeinsamen evidenztheoretischen Rahmen gelegt. Dabei werden Rasterkarten verwendet, welche eine effiziente Datenverarbeitung ermöglichen. Der Ansatz wird mit einem automatisierten Versuchsträger anhand eines echten Verkehrsszenarios validiert. Experimente zeigen, dass die Genauigkeit mit Hinzunahme von Sensoren erhöht werden kann ohne die Echtzeitfähigkeit zu verlieren.

Schlüsselwörter: Sensordatenfusion, Evidenztheorie, Rasterkarten.

1 Introduction

A detailed reconstruction of the environment is a crucial component of mobile robotic systems such as automated vehicles. As different sensing principles such as camera, lidar or radar have their own strengths and weaknesses a heterogeneous sensor setup is used to ensure information redundancy. However, in order to reduce computational and design complexity for subsequent higher-level automation modules (e.g. path planning) it is desirable to fuse all sensor data into one common representation. In this context, environment reconstruction may not only include the estimation of areas occupied by traffic participants or static obstacles but also the identification of contradicting sensor measurements or occluded areas. Based on this information an automated system can decide whether it should transit into a fail-safe state or hand-over to a human driver.

1.1 Related work

As in most of the traffic scenarios relevant traffic participants move on a common ground surface, top-view grid maps are often used as common representation for range sensor data fusion. Occupancy grid maps were first introduced by Elfes et al. [3] and enable efficient convolutional operations or GPU-based raycasting [4] due to their dense grid structure, at the cost of higher memory consumption

*Korrespondenzautor: Sven Richter, Institut für Mess- und Regelungstechnik (MRT), Karlsruher Institut für Technologie (KIT), Engler-Bunte-Ring 21, 76131 Karlsruhe, E-Mail: sven.richter@kit.edu

Sascha Wirges, Hendrik Königshof, FZI Forschungszentrum Informatik, Haid-und-Neu-Straße 10-14, 76131 Karlsruhe, E-Mail: wirges@fzi.de, koenigshof@fzi.de

Christoph Stiller, Institut für Mess- und Regelungstechnik (MRT), Karlsruher Institut für Technologie (KIT), Engler-Bunte-Ring 21, 76131 Karlsruhe, E-Mail: stiller@kit.edu

compared to other representations such as search trees. Assuming sensor data was measured from a single view-point, polar grid maps can be used which simplifies the modeling of sensor noise and enables parallel raycasting. In [12], Yu et al. present a model which efficiently maps lidar range measurements from polar coordinates into an evidential cartesian grid. In contrast to modeling cell occupancies using Bayesian probability theory, evidence theory (see Sec. 1.3) makes it possible to resolve contradicting measurements. Assuming proper classification into ground and non-ground detection points the authors also develop a method to assign evidences with a backward free space propagation from polar into cartesian coordinates. Nuss et al. [5] combine lidar range measurements and monocular camera images into a common grid map representation using an evidential approach. Their method incorporates spatial uncertainty due to the flat-world assumption in the image projection and is able to detect obstacles that violate the ground assumption. Capellier et al. [1] estimate semantic image labels using a neural network and fuse them into an evidential grid to assess drivability in areas of the grid map. The authors classify range measurements into ground and non-ground, estimate an evidential grid according to [12] and warp segmented camera images into this grid using the estimated ground plane.

1.2 Goals of this work

As it is desired to combine as many information as possible into a common framework, the goal of this work is to estimate occupancy, free-space and semantic information by fusing data from multiple lidars, radars, a stereo camera and semantically labeled camera images into an evidential top-view grid map. In contrast to other approaches presented in Sec. 1.1, model observability along the height dimension is explicitly modeled during raycasting. After a brief summary of evidence theory in Sec. 1.3, a new frame of discernment containing relevant semantic hypotheses and our sensor models, which are interpreted as functions from the measurement into the belief space, are introduced in Sec. 2. To retain sensor redundancy, the belief mass assignments are calculated for each sensor individually. In Sec. 3 it is explained how evidential support from different sensors is fused into a common representation by applying evidential combination rules. Finally, an experimental validation of the approach is presented in Sec. 4 followed by a brief outlook to future work in Sec. 5.

1.3 Evidence theory

Evidence theory (e.g. [8]) can be interpreted as an extension of Bayesian probability theory and allows to estimate belief, plausibility and uncertainty of hypotheses from different sensor sources. Let Θ be the set including elementary hypotheses of interest. Then, a belief assignment

$$m : 2^\Theta \rightarrow [0, 1] , \quad \sum_{\mathcal{H} \in 2^\Theta} m(\mathcal{H}) = 1 \quad (1)$$

defines a mapping from the power set 2^Θ , thus also assigning belief masses to multiple hypotheses. Given a particular hypothesis $\mathcal{H}$, the uncertainty $\text{unc}(\mathcal{H}) = \text{pl}(\mathcal{H}) - \text{bel}(\mathcal{H})$ can be expressed by the belief $\text{bel}(\mathcal{H})$ which describes the support only for $\mathcal{H}$ and the plausibility $\text{pl}(\mathcal{H})$ modeling support for not disagreeing with $\mathcal{H}$. Belief assignments from multiple sources can be combined using different rules such as Dempster's rule [8] which ignores conflicting belief or Yager's rule [11] which interprets conflicts as uncertainty.

2 Sensor models

Sensor specific tasks as a pixel-wise semantic labeling in the image domain or a lidar-based ground surface estimation and segmentation of ground detections are assumed to be solved and provided to the presented workflow.

In this work, the set

$$\Theta = \{c, cy, p, om, nm, s, sw, t\},$$

which consists of the elementary hypotheses depicted in Tab. 1, is used. The set Θ can be divided into the presence and the absence of potential obstacles, represented by the sets $\mathcal{O}$ and $\mathcal{G}$, respectively (cf. Tab. 1), summarizing Θ to the classical occupancy frame. Typically for range sensors, the absence of obstacles is measured by a ray casting between sensor origin and detection. Subsequently, those measurements are referred to as negative measurements

Tab. 1: Elementary and combined hypotheses used in this work.

Semantic class	Set	Superset
Car	{c}	Obstacle $\mathcal{O} = \{c, cy, p, om, nm\}$
Cyclist	{cy}	
Pedestrian	{p}	
Other movable	{om}	
Non movable	{nm}	
Street	{s}	Ground $\mathcal{G} = \{s, sw, t\}$
Sidewalk	{sw}	
Terrain	{t}	

and assigned to the hypothesis $\mathcal{G}$. Semantic measurements or reflections by range sensors are referred to as positive measurements and assigned to either the corresponding elementary hypothesis or $\mathcal{O}$.

Similar as in [12], statistical independence of measurement errors when computing belief masses is assumed. Let p_{FP} and p_{FN} be the sensor specific probabilities for a false positive and a false negative measurement, respectively. Then, assuming n positive and m negative measurements, the total error probabilities are given as p_{FP}^n and p_{FN}^m .

Both point detection from range sensors and the pixel labels from vision are first mapped into a sensor grid map in polar coordinates. This enables the explicit modeling of sensor noise by applying a Gaussian filter to the mass layers in angular and range direction separately. The polar grid maps are then remapped into cartesian coordinates utilizing a bilinear interpolation algorithm.

2.1 Semantic estimation

For pixel-wise semantic segmentation, a convolutional neural network with an encoder based on ResNet-38 [10] is used. Its output is decoded to a semantic segmentation map with the original image resolution. The detailed architecture of this network is described in [7]. This neural network was trained on the Cityscapes dataset [2] by using both the 5k precisely and 20k coarsely annotated training images. To estimate disparities from the stereo camera setup the block matching algorithm presented in [6] is used. By assuming slanted planes this approach yields competitive results while being fast due to an efficient GPU implementation.

Based on the pixel-wise labeled image, belief masses for the hypotheses $\{s\}$, $\{sw\}$ and $\{t\}$ are computed by projecting the pixel coordinates to the corresponding grid cell based on the estimated ground surface (Inverse perspective mapping). The labels that correspond to obstacles, i.e. $\{c\}$, $\{cy\}$, $\{p\}$, $\{om\}$ and $\{nm\}$ are considered only if additional range information from stereo matching is available. From the image domain, positive measurements are deduced only. Consequently, the basic belief assignment is defined as

$$m_v(\theta) = \begin{cases} p_{FP}^m(1 - p_{FP}^n), & \text{if } \theta \in \Theta, \\ 1 - \sum_{\xi \neq \Theta} m(\xi), & \text{if } \theta = \Theta, \end{cases}$$

where n denotes the number of pixels with label θ assigned to the current cell and m the number of pixels with a different label. Note that all conflicting hypotheses $\theta \in \Theta$ with $|\theta| > 1$ are neglected in the definition.

Abb. 1: Illustration of the ratios r_{x_i} and r_{z_i} used in the computation of the false negative probability.

2.2 Occupancy estimation

For pure range sensors as radars and lidars, masses are assigned to the supersets $\mathcal{O}$ and $\mathcal{G}$ only due to the lack of semantic information.

The basic belief assignment reads

$$m_r(\theta) = \begin{cases} p_{FN}^m(1 - p_{FP}^n), & \text{if } \theta = \mathcal{O}, \\ p_{FP}^n(1 - p_{FN}^m), & \text{if } \theta = \mathcal{G}, \\ 1 - \sum_{\xi \neq \Theta} m(\xi), & \text{if } \theta = \Theta, \end{cases}$$

where the error probabilities p_{FP} and p_{FN} are modeled for each sensor. In contrast to p_{FP} which is assigned a constant value, p_{FN} depends on the ray geometry and the distance to the sensor origin. The false negative probability

$$p_{FN}(\mathbf{x}_i) = 1 - (1 - r_{x_i}) r_{z_i} (1 - p_{FN, \max}),$$

is computed based on the distance ratio

$$r_{x_i} = \frac{\|\mathbf{x}_i\|}{\|\mathbf{x}_{\max}\|},$$

the height ratio

$$r_{z_i} = \frac{\Delta z_i}{\Delta z}$$

and a maximum false negative rate $p_{FN, \max}$. The two ratios r_{x_i} and r_{z_i} are illustrated in Fig. 1 which exemplary shows a ray geometry for a laser scanner with five layers pointing to a fixed angle.

3 Fusion

Sensor grid maps $\mathcal{M}_1$ and $\mathcal{M}_2$ with belief assignments m_1 and m_2 are combined by intersecting the corresponding hypothesis sets. Non-conflicting hypotheses are combined by the conjunctive rule of combination [9]

$$m(\theta | \mathcal{M}_1 \oplus \mathcal{M}_2) = \sum_{\theta_1 \cap \theta_2 = \theta} m_1(\theta_1) m_2(\theta_2)$$

for $\theta \in 2^\Theta$.

Conflict masses, i.e. $m(\theta_1)m(\theta_2)$ for disjoint sets $\theta_1, \theta_2 \in 2^\Theta$ are assigned individually. The following conflict masses occur in $\mathcal{G}_1 \oplus \mathcal{G}_2$:

$$\begin{aligned}\zeta_1 &= m_1(\mathcal{O})m_2(\mathcal{G}) + m_2(\mathcal{O})m_1(\mathcal{G}), \\ \zeta_2 &= \sum_{\theta \in \mathcal{G}} m_1(\theta)m_2(\mathcal{O}) + m_2(\theta)m_1(\mathcal{O}), \\ \zeta_3 &= \sum_{\theta \in \mathcal{O}} m_1(\theta)m_2(\mathcal{G}) + m_2(\theta)m_1(\mathcal{G}).\end{aligned}$$

The former two conflict masses ζ_1 and ζ_2 are assigned to $\mathcal{O}$ whereas ζ_3 is assigned to $\mathcal{G}$. That is, conflict masses are assigned to the more conservative hypothesis or the hypothesis that stems from the range sensor due to the more accurate distance estimation. Therefore, the basic belief assignment is defined as

$$m_f(\theta) = \begin{cases} m(\theta \mid \mathcal{M}_1 \oplus \mathcal{M}_2), & \text{if } \theta \in \Theta, \\ m(\theta \mid \mathcal{M}_1 \oplus \mathcal{M}_2) + \zeta_1 + \zeta_2, & \text{if } \theta = \mathcal{O}, \\ m(\theta \mid \mathcal{M}_1 \oplus \mathcal{M}_2) + \zeta_3, & \text{if } \theta = \mathcal{G}. \end{cases} \quad (2)$$

for the fused grid map combining the sensor grid maps $\mathcal{M}_1$ and $\mathcal{M}_2$.

4 Evaluation

The presented approach is validated on real sensor data recorded from a test vehicle in the urban traffic scenario depicted in Fig. 2. In particular, sensor grid maps are estimated and fused for four Velodyne VLP16, one four layer Ibeo LUX lidar, one Continental ARS408 long range radar and a stereo camera setup. To reduce rolling shutter effects, measurements are processed at 20 Hz and then fused within a sliding window of 50 ms. Due to the GPU implementation, the presented procedure runs in real-time on our test vehicle with an update rate of 20 Hz.

Qualitative results of the sensor grid mapping are shown in Fig. 3. Note that the mass assignments are increasingly blurred in angular direction at higher distances as the measurements are first mapped in polar coordinates. The strengths and weaknesses of the used sensors are clearly visible in the corresponding sensor grid maps. The two laser scanners provide accurate range measurements with errors smaller than 3 cm, a dense detection pattern and a small false detection rate, however, lacking a high detection range. The long range radar provides very sparse detections at a higher false detection rate, however, detects obstacles in far distances. The stereo camera has a high detection range, a dense detection pattern, but a high uncertainty in the distance estimation. The high

(a) Camera image of the urban scene.

(b) Pixel-wise labeled image.

Abb. 2: Images of an urban traffic scene which the proposed method is evaluated on. Semantic labels in are colored according to the Cityscapes scheme (see [2])

false positive rate of the radar and the camera is explicitly modeled by assigning significantly higher values to the false positive rate p_{FP} .

The fusion of the sensor grid maps using (2), provides a joined mass assignment for the frame Θ . Fig. 4 shows the assignments for the supersets $\mathcal{G}$ and $\mathcal{O}$, the singleton hypotheses $\theta \in \Theta$ and the classical occupancy grid map. Note that the representation in Fig. 4 (a) only contains the masses that are directly assigned to $\mathcal{G}$ and $\mathcal{O}$. The occupancy grid map in Fig. 4 (c) shows that the sensor grid maps from Fig. 3 were successfully fused into a common representation. The semantic grid map (Fig. 4 (b)) has slightly reduced mass assignments. This is due to the fact that conflicts can not be assigned to semantic labels, cf. (2). The joined representation can be seen as a plausibility check of the semantic grid mapping by range sensors. This is error prone in particular for small obstacles as the pedestrian on the right sidewalk. Due to the asynchronicity of the sensors or an inaccurate calibration those labels might overlap with free space. Those obstacles are then only represented by range detections that are assigned to $\mathcal{O}$.

5 Conclusion

In this work, an evidential sensor data fusion approach on grid maps was presented. In contrast to other publications, a detailed vision-based semantic labeling was included in the representation and fused consistently estimating both occupancy and semantics jointly in a newly introduced frame of discernment. A validation based on real sensor measurements shows that this approach is able to estimate

Abb. 3: Resulting sensor grid maps for the introduced vehicle setup. Belief mass for $\mathcal{G}$ are depicted in green, red the belief mass for $\mathcal{O}$ ((a) - (c)). Semantic labels in (d) are colored similarly as in Fig. 2.

Abb. 4: Fused grid map.

a consistent model of the environment by proving real-time-capability. Future work will focus on improving the inverse perspective mapping from the camera image to the grid map domain. Furthermore, an evidential filtering approach will be explored in order to infer dynamics and obtain a more dense and precise semantic classification.

References

- [1] E. Capellier, F. Davoine, V. Fremont, J. Ibanez-Guzman und Y. Li. Evidential Grid Mapping, from Asynchronous LIDAR Scans and RGB Images, for Autonomous Driving. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, S. 2595–2602. IEEE, 2018.
- [2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth und B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [3] A. Elfes. Using Occupancy Grids for Mobile Robot Perception and Navigation. *Computer*, 22(6):46–57, 1989.
- [4] F. Homm, N. Kaempchen, J. Ota und D. Burschka. Efficient Occupancy Grid Computation on the GPU with Lidar and Radar for Road Boundary Detection. In *IEEE Intelligent Vehicles Symposium (IV)*, S. 1006–1013. IEEE, 2010. ISBN 9781424478682.
- [5] D. Nuss, M. Thom, A. Danzer und K. Dietmayer. Fusion of Laser and Monocular Camera Data in Object Grid Maps for Vehicle Environment Perception. In *17th International Conference on Information Fusion (FUSION)*, S. 1–8. IEEE, 2014.
- [6] B. Ranft und T. Strauß. Modeling arbitrarily oriented slanted planes for efficient stereo vision based on block matching. In *Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on*, S. 1941–1947. IEEE, 2014.
- [7] N. O. Salscheider. Simultaneous object detection and semantic segmentation. *arXiv preprint arXiv:1905.02285*, 2019.
- [8] G. Shafer. *A Mathematical Theory of Evidence*, Band 42. Princeton University Press, 1976.
- [9] P. Smets. The combination of evidence in the transferable belief model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):447–458, May 1990. ISSN 0162-8828. 10.1109/34.55104.
- [10] Z. Wu, C. Shen und A. Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019.
- [11] R. R. Yager. On the dempster-shafer framework and new combination rules. *Inf. Sci.*, 41:93–137, 1987.
- [12] C. Yu, V. Cherfaoui und P. Bonnifait. An Evidential Sensor Model for Velodyne Scan Grids. In *2014 13th International Conference on Control Automation Robotics & Vision (ICARCV)*, S. 583–588. IEEE, 2014.