# Multiple Hypothesis Semantic Mapping for Robust Data Association

Lukas Bernreiter, Abel Gawel, Hannes Sommer[1], Juan Nieto, Roland Siegwart and Cesar Cadena

***Abstract*—In this paper, we present a semantic mapping approach with multiple hypothesis tracking for data association. As semantic information has the potential to overcome ambiguity in measurements and place recognition, it forms an eminent modality for autonomous systems. This is particularly evident in urban scenarios with several similar looking surroundings. Nevertheless, it requires the handling of a non-Gaussian and discrete random variable coming from object detectors. Previous methods facilitate semantic information for global localization and data association to reduce the instance ambiguity between the landmarks. However, many of these approaches do not deal with the creation of complete globally consistent representations of the environment and typically do not scale well. We utilize multiple hypothesis trees to derive a probabilistic data association for semantic measurements by means of position, instance and class to create a semantic representation. We propose an optimized mapping method and make use of a pose graph to derive a novel semantic SLAM solution. Furthermore, we show that semantic covisibility graphs allow for a precise place recognition in urban environments. We verify our approach using real-world outdoor dataset and demonstrate an average drift reduction of 33 % w.r.t. the raw odometry source. Moreover, our approach produces 55 % less hypotheses on average than a regular multiple hypotheses approach.**

***Index Terms*—SLAM, Semantic Scene Understanding, Probability and Statistical Methods**

## I. Introduction

SEMANTIC data is a reliable and ubiquitous flow of information in structured and non-structured environments. Especially for perception systems, semantically annotated data and higher reasoning about the underlying scene on top of purely geometric approaches have the potential to increase the robustness of the estimation [1], [2]. A reliable mapping is eminently important especially for autonomous, as well as, augmented reality systems since the recognition of the surrounding objects and the localization in a globally unknown environment are crucial factors there.

Manuscript received: February, 24, 2019; Revised May, 16, 2019; Accepted June, 10, 2019.

This paper was recommended for publication by Editor Cyrill Stachniss upon evaluation of the Associate Editor and Reviewers' comments.

This work was supported by the National Center of Competence in Research (NCCR) Robotics through the Swiss National Science Foundation and has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688652 and from the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 15.0284.

All authors are with the Autonomous Systems Lab, ETH Zurich, Zurich 8092, Switzerland, {berlukas, gawela, sommerh, nietoj, rsiegwart, cesarc}@ethz.ch.

[1] Additionally with Sevensense Robotics AG, Zurich.

Digital Object Identifier (DOI): see top of this page.

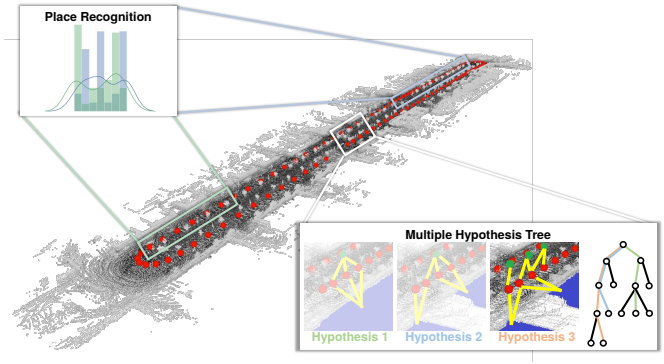


Fig. 1: We propose a semantic SLAM system that maintains multiple hypotheses of the landmark locations structured in a hypothesis tree (bottom right image). Data association is done in a semantic framework to create new branches in the hypothesis tree. Furthermore, we perform a semantic place recognition method utilizing the object class distribution of a submap (top left image).

Traditional approaches for localization often rely on specific low-level visual features such as points and lines which are inherently ambiguous preventing the approach to scale well to large environments. In contrast, semantic information features a promising approach for many robotic applications by allowing more unique local and global descriptors for landmarks as well as potential viewpoint-invariance. Therefore, this constitutes a crucial factor for the measurement association to mapped landmarks and thus influences the quality of the localization. Moreover, semantics are very efficient at dealing with place recognition as they are less affected by seasonal or appearance changes as well as large drifts.

In a conventional SLAM setting, the measurement noise is commonly relaxed to the continuous Gaussian case [1] which however, does not apply to semantic variables. Uncertainties in the object detection such as class labels and object instances typically involve the handling of non-Gaussian discrete variables. How to properly handle such variables is still quite challenging and remains an open research question [3].

Many existing semantic mapping approaches are primarily concerned with the creation of an indoor semantic representation with minor illumination and viewpoint changes [4]. In contrast, realistic outdoor applications often come with severe changes of illumination and viewpoint. This can hamper loop closure detection since drastic view-point changes might render scenes completely different when revisiting.

Additionally, local descriptors for place recognition often

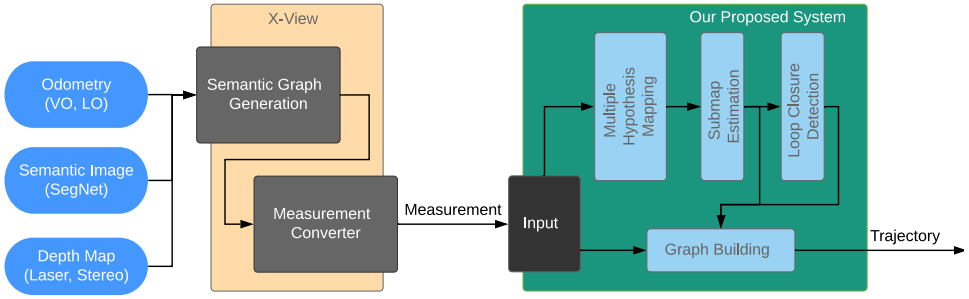


Fig. 2: System overview of our proposed approach. We make use of our previous work X-View [5] by means of the semantic object extraction. Hereby the semantic objects are converted to measurements and used as input for our system. Afterwards we start simultaneously with the creation of the factor graph as well as with the multiple hypothesis mapping of the environment. Loop closures are detected and placed into the factor graph when submaps are completed.

rely on the bag-of-words paradigm [6], [7] which can fail often in environments with repetitive features which commonly occur in urban environments leading to false loop closures.

Other semantic SLAM systems do not directly incorporate the semantic information into the estimation but rather use it to filter out bad classes such as cars or pedestrians beforehand [8].

In this work, we aim to build a globally consistent semantic mapping formulation by improving the incorporation of discrete random variables in the map building and localization processes.

Throughout this paper, observations comprise the semantic class and position from static landmarks as well as the spatial relationship to other static landmarks. We approach the measurement association problem by utilizing the semantic class of an object and deferring the decision on associations until the ambiguity is resolved. In other words, the decision on the association is done at a time when more observations are available or a place is revisited allowing to correctly identify the instance label with a certain assurance. This is motivated by the fact that in many cases the most likely association given only a few measurements does not necessarily need to be the correct one.

Furthermore, we derive a loop closure detection and verification algorithm operating directly on the level of semantic objects. Utilizing the class labels and the spatial relationships between the objects enables a robust recognition of places in urban environments. An overview of our proposed system is given in figure 2.

The main contributions of this work are

Consistent multiple hypothesis mapping using an optimized Multiple Hypothesis Tracking (MHT) approach.
A Dirichlet Process (DP)-based relaxed probabilistic Hungarian algorithm for viewpoint-invariance.
Semantic selection strategy to identify potential submaps for loop closures.
Place recognition based on the semantic classes and the covisibility graphs.
Incorporation of the proposed approach into a graph-based semantic SLAM pipeline and evaluation of the resulting system.

### A. Related Work

In recent years, the advances to deep learning systems led to more reliable as well as practically usable object detectors [9]. Consequently, this allowed SLAM systems to additionally include semantically rich information in order to improve their estimation [10]–[12].

Recently, some research specifically addresses the problem of correctly assigning measurements to already known objects utilizing additional semantic information [13], [14]. These systems, however, do not deal with the estimation of the camera's position, i.e. their application implies a static position of the camera and is often placed indoors. Thus, they are not optimized for viewpoint-invariance, but rather emphasize on the probabilistic data association and the tracking of objects across multiple scenes.

Nevertheless, we make use of the close relationship to robotic mapping since target tracking is a special case of mapping. Elfring et al. [13] presented a semantic anchoring framework using MHTs [15] which defers the data association until the ambiguity between the instances is resolved. Generally, the MHT enables accurate results but is inherently intractable with a large amount of objects and requires frequent optimizations [16]. The work of Wong et al. [14] presents an approach using the DPs which yields estimation results comparable to the MHT but with substantially less computational effort. Nevertheless, their proposed approach is not incremental and therefore, not directly applicable for the mapping of a robot's environment. Similar, in their previous work [17] the authors propose a world modeling approach using dependent DPs to accommodate for dynamic objects. In their proposed framework, the optimal measurement assignment is computed using the Hungarian method operating on negative log-likelihoods for the individual cases. Furthermore, Atanasov et al. [18] emphasizes on a novel derivation of the likelihood of Random Finite Set (RFS) models using the matrix permanent for localizing in a prior semantic map. Their system utilizes a probabilistic approach for data association which considers false positives in the measurements.

There is a vast literature on indoor semantic mapping available which however, does often not directly incorporate semantic information in a SLAM pipeline but rather uses the

information for mapping and scene interpretation [19] [20]. The work that is most similar to ours is the work of Bowman et al. [21] which proposed a semantic system which enables to directly facilitate semantic factors in their optimization framework. Despite using probabilistic formulation for the data association, their approach inherently neglects false positives and false negatives, and lacks including a prior on the assignments. Moreover, they limited the possible classes in their mapping so that only cars were enabled in their outdoor experiments. This greatly reduces the complexity in outdoor scenarios with semantically rich information and further, is not a reliable source for place recognition.

Another direction of research is to represent landmarks as quadrics to capture additional information such as size and orientation [12], [22]. However, they either assume that the measurement association is given [22] or utilize the semantic labels for a hard association using a nearest neighbor search [12]. Thus, their work does not include any probabilistic inference for the association and does not consider false positives.

Our previous work by Gawel et al. [5] represents the environment using semantic graphs and performs global localization by matching query graphs of the current location with a global graph. The query graphs however, are not used in a data association framework and thus, landmarks could potentially be duplicated. This system does neither deal with map management and optimization nor with drift reduction for globally consistent mapping. Our semantic SLAM system does not require any prior of the object shapes and comprises a soft probabilistic data association for semantic measurements. To the best of our knowledge, a complete semantic SLAM system comprising the aforementioned approaches has not been reported in literature before.

In the remaining part of this paper we will start deriving a semantic mapping approach (section II) and a concrete algorithm for localization (section III). The presented work is evaluated in chapter IV. Finally, chapter V concludes this work and gives further research directions.

## II. Semantic Multiple-Hypotheses Mapping

When performing SLAM, measurement noise typically leads to drift and inconsistent maps – in particular when measurements get wrongly associated to landmarks. We approach this problem by introducing locally optimized submaps. Each submap maintains an individual Multiple Hypothesis Tree (MHt) and propagates a first-moment estimate to proximate submaps. Specifically, for each submap we want to maximize the posterior distribution, $f(\mathbf{\Theta}_t | \mathbf{Z}_t)$, of the associations $\mathbf{\Theta}_t$ of all measurements $\mathbf{Z}_t$ received till the time step $t$ which is proportional to[1]

$$f(\underline{z}_t | \mathbf{\Theta}_t, \mathbf{Z}_{t\ 1}) f(\underline{\theta}_t | \mathbf{\Theta}_{t\ 1}, \mathbf{Z}_{t\ 1}) f(\mathbf{\Theta}_{t\ 1} | \mathbf{Z}_{t\ 1}). \quad (1)$$

Here, the set of $N$ measurement associations at time step $t$ is represented by $\underline{\theta}_t = [\theta_t^1 ... \theta_t^N]$. The first factor, $f(\underline{z}_t | \mathbf{\Theta}_t, \mathbf{Z}_{t\ 1})$, in (1) represents the distribution of all measurements at time

[1]Vectors are underlined and matrices are written with bold capital letters.

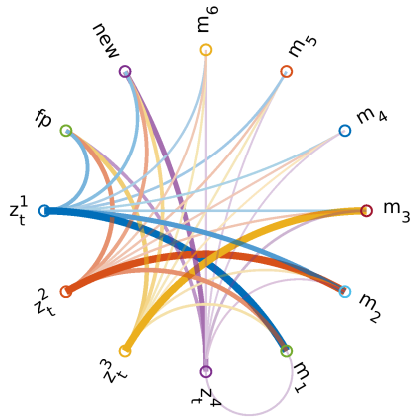


Fig. 3: Likelihood of assigning a measurement to a specific scenario. Each measurement $z_t^i$, at time $t$, could be assigned to any of the existing landmarks $m_{f1..6g}$, represent a new landmark (new) or a false positive (fp). The thickness and opacity denotes how likely the association is given a certain example of measurements and landmarks.

$t$ for which we assume conditional independence of the individual measurements such that it equals to

$$\prod_{i=1}^{n} f(\underline{z}_t^i | \theta_t^i = l, \mathbf{Z}_{t\ 1}) = p_s(c_t^i) \prod_{i=1}^{n} p(c_t^i | \gamma_{t\ 1}^l) f(\underline{p}_t^i | \underline{\pi}_{t\ 1}^l), \quad (2)$$

where $\underline{z}_t^i$ denotes the attributes of the $i$th semantic measurement at time $t$, $\theta_t^i$ is the index of the landmark, $l$, this measurement is associated with, and $\underline{\pi}_{t\ 1}^l$, $\gamma_{t\ 1}^l$ the assigned landmark's position and class estimated at time $t\ \ 1$. Furthermore, a semantic measurement, $\underline{z}_t^i$, is split into its position, $\underline{p}_t^i$, and class, $c_t^i$, component, whose Probability Mass Function (PMF), $p_s$, is a prior assumption based on how well the classes fit into the current environment. For the $i$th class measurement, $c_t^i$, we assume $p(c_t^i | \theta_t^i = l, \gamma_{t\ 1}^l) := \delta_{c_t^i, \gamma_{t-1}^l}$, where $\delta$ denotes the Kronecker delta. For $\underline{p}_t^i$, we assume the following form of a stochastic measurement model

$$f(\underline{p}_t^i | \theta_t^i = l, \underline{\pi}_{t\ 1}^l) = f_{\underline{\nu}}(\underline{p}_t^i \quad \underline{\pi}_{t\ 1}^l),$$

where $f_{\underline{\nu}}$ denotes the Probability Density Function (PDF) of the additive position measurement noise, $\underline{\nu}_t^i$, which we model as a zero-mean Gaussian distribution with covariance $\mathbf{\Sigma}^z$. For practical stability, we use an Unscented Kalman Filter (UKF) [23] for the estimation of $\pi_t^l$. The second factor, $f(\underline{\theta}_t | \mathbf{\Theta}_{t\ 1}, \mathbf{Z}_{t\ 1})$, in (1) is the assignment prior and is calculated using the well-known equation [13], [16]

$$f(\underline{\theta}_t | \mathbf{\Theta}_{t\ 1}, \mathbf{Z}_{t\ 1}) = \frac{N_t^n! N_t^f!}{N_t^m!} p_n(N_t^n) p_f(N_t^f), \quad (3)$$

where $N_t^n$ denotes the number of new measurements, $N_t^f$ the number of false positives identified by the Hungarian method and $N_t^m$ the total number of measurements at time step $t$. The functions $p_n$ and $p_f$ are prior PMFs over the number of new measurements and false positives, respectively. Typically both

are chosen as Poisson PMFs with a specific spatial density $\lambda$ and a volume $V$ [24], e.g.

$$p_n(N) = \exp(\lambda_n V_n)\frac{(\lambda_n V_n)^N}{N!}.$$

Each branch in the MHt comprises a different set of associations $\boldsymbol{\Theta}_t$. Utilizing (2) and (3) together with the previous posterior distribution $f(\boldsymbol{\Theta}_{t-1}|\mathbf{Z}_{t-1})$ we can evaluate these branches using (1).

### A. Probabilistic Measurement Association

Finding the correspondence $\underline{\theta}_t$ between measurements and mapped object can be challenging since the current set of measurements often does not allow deriving a correct assignment. Fortunately, this problem can be considered as a weighted combinatorial assignment problem for which the Hungarian algorithm [17], [25] is well known. Figure 3 illustrates the probabilistic combinatorial assignment problem. To find the most likely assignment we utilize a stochastic association algorithm based on the DP. DPs are a good choice for modeling the probability of seeing new and re-observing already mapped landmarks [26].

The likelihood of the associations of new measurements $\underline{z}_t$ at time $t$ with landmarks, $\underline{\theta}_t$, is expressed by $f(\underline{\theta}_t|\underline{z}_t, \boldsymbol{\Theta}_{t-1}, \mathbf{Z}_{t-1})$. We assume that at each time step a landmark in the scene can at most generate one observation. Inspired by the dependent DP formulation in [17], we differentiate four cases: (i) landmarks that have already been seen in the current submap, (ii) landmarks seen in previous submaps, (iii) new landmarks and (iv) false positives. The likelihood for the association of a measurement $\underline{z}_t^i$ with an existing landmark $k$ of the same class in the current submap is modeled as

$$f(\theta_t^i = k|\underline{z}_t^i, \boldsymbol{\Theta}_{t-1}, \mathbf{Z}_{t-1}) = \exp(N_t^k) f_{\underline{\nu}}(\underline{p}_t^i \;\; \underline{\pi}_t^k). \tag{4}$$

Here, the scalar $N_t^k$ denotes the number of assignments to the landmark $k$. Despite the fact that we only deal with static objects, a landmark, $l$, which was seen in a previous submap at time, $\tau$, is modeled using a transitional density, $T$, i.e.

$$f(\theta_t^i = l|\underline{z}_t^i, \boldsymbol{\Theta}_{t-1}, \mathbf{Z}_{t-1}) = \int f_{\underline{\nu}}(\underline{p}_t^i \;\; \underline{x}) T(\underline{x}, \underline{\pi}_\tau^l)\, d\underline{x}\,. \tag{5}$$

The transitional density, $T$, depends on the semantic class of the object and is used to accommodate for the unknown shape of the landmarks. Since we take the centroid of the segmented objects as input, we employ two approaches for the choice of $T$ to compensate for large measurement noise. Objects such as poles and trees are modeled using a Dirac-$\delta$ distribution: $T(\underline{x}|\underline{\pi}_\tau^l) = \delta(\underline{x} \;\; \underline{\pi}_\tau^l)$, reducing the right hand side of (5) to $f_{\underline{\nu}}(\underline{p}_t^i \;\; \underline{\pi}_\tau^l)$, the measurement distribution of $\underline{z}_t^i$ given the last seen position of $l$. The transition of objects like buildings and fences is modeled using a Gaussian distribution with covariance $\boldsymbol{\Sigma}^\alpha$ resulting in

$$\begin{aligned}\int f_{\underline{\nu}}(\underline{x} \;\; \underline{p}_t^i)\mathcal{N}(\underline{x}; \underline{\pi}_\tau^l, \boldsymbol{\Sigma}^\alpha)\, d\underline{x} \\ = \mathcal{F}^{-1}\{\mathcal{F}\{\mathcal{N}(0, \boldsymbol{\Sigma}^z)\} \cdot \mathcal{F}\{\mathcal{N}(\underline{\pi}_\tau^l, \boldsymbol{\Sigma}^\alpha)\}(\underline{p}_t^i)\} \\ = \mathcal{N}(\underline{p}_t^i; \underline{\pi}_\tau^l, \boldsymbol{\Sigma}^z + \boldsymbol{\Sigma}^\alpha),\end{aligned}$$

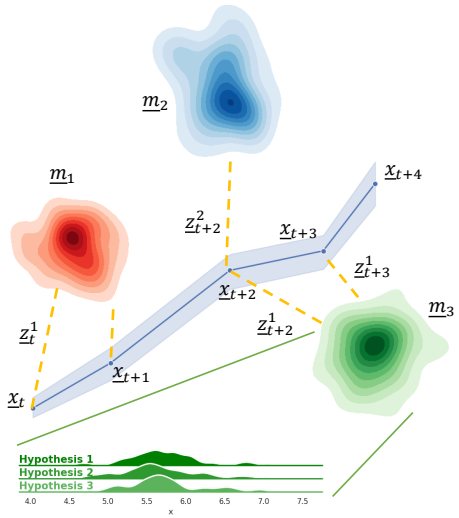


Fig. 4: Illustration of the Gaussian mixture landmarks created by the fusion of the weighted hypotheses.

where $\mathcal{F}$ is the Fourier transform and $\boldsymbol{\Sigma}^\alpha$ is a transitional covariance depending on the object's class $c$. The likelihood of assigning a new landmark, $l$, to the $i$th observation at time $t$, is approximated by the uniform distribution in $\underline{z}_t^i$ over the volume of the map [17], $\mathcal{M}$, i.e.

$$f(\theta_t^i = l|\underline{z}_t^i, \boldsymbol{\Theta}_{t-1}, \mathbf{Z}_{t-1}) = \alpha \int f_{\underline{\nu}}(\underline{p}_t^i \;\; \underline{x}) H_{\mathrm{DP}}(\underline{x})\, d\underline{x} \approx \alpha U_{j\mathcal{M}j},$$

where $H_{\mathrm{DP}}$ is the base distribution of the DP. Generally, false positives occur due to clutter in the images and are essentially detected objects which do not physically exist in the environment. Having such cases in the map may result in improper assignments of future measurements. False positives have the likelihood $f(\theta_t^i = 0|\underline{z}_t^i, \boldsymbol{\Theta}_{t-1}, \mathbf{Z}_{k-1})$, i.e. the likelihood of the measurement $i$ being an observation of the *false* landmark, 0. False positives, are assumed to occur at a fixed rate $\rho$, i.e.

$$f(\theta_t^i = 0|\underline{z}_t^i, \boldsymbol{\Theta}_{t-1}, \mathbf{Z}_{t-1}) \propto \left(\prod_j f(\underline{z}_t^i|\theta_t^i = j, \gamma_{t-1}^j, \underline{\pi}_{t-1}^j)\right)^{-1} \begin{cases}\rho N_t^0 & N_t^0 > 0 \\ \rho\alpha & N_t^0 = 0\end{cases},$$

where $N_t^0$ denotes the number of occurred false positives until time step $t$, and $\alpha$ is the concentration parameter of the DP. The aforementioned four cases will be used as an input to the Hungarian algorithm yielding an optimal assignment for each measurement as well as landmark. Based on this initial assignment, the optimal branch of the MHt will be formed. In case the assignment is not distinct enough, new branches in the MHt are generated by re-running the Hungarian algorithm without the previous optimal assignment. Since small MHts are generally better for computational performance, we only create branches for associations that are reasonable.

### B. Optimized Resampling of Hypotheses

Each hypothesis is weighted using their measurement likelihood (2) and assignment prior term (3). At each time step, the

Fig. 5: Illustration of the trajectories of the KITTI sequence 05, sequence 00, sequence 07 together with a laser map. Regions in blue denote the estimated trajectory, orange regions are submaps which were checked for loop closures and green areas show performed loop closures. We gain additional efficiency by only checking a subset of the submaps for loop closure.

existing hypotheses are reweighted and eventually resampled by a systematic resampling technique [27]. In general, this fuses the current knowledge in the hypothesis set and eliminates the hypotheses which have a low weight and preserves hypotheses with a good weight.

A crucial factor is when to decide that resampling should be performed on the hypothesis tree. In this case, it is common to use selective resampling [28] based on the calculation of the effective sample size which essentially captures the diversity of the hypothesis set. Consequently, resampling is only performed when the effective sample size exceeds a certain threshold. Furthermore, many particle filter implementations only consider a fixed particle size. However, it is desired that the number of particles is high for a high state uncertainty, and low when the uncertainty is low. Fox [29] introduced a variable sampling algorithm based on the KLD distance for particle filters. During each iteration of the resampling procedure, the number of hypotheses is dynamically bounded by $n$ using

$$n = \frac{k \quad 1}{2\epsilon}\left(1 \quad \frac{2}{9(k \quad 1)} + \sqrt{\frac{2}{9(k \quad 1)}} z_{1 \quad \delta}\right), \qquad (6)$$

where $k$ is the current number of resampled hypotheses and $z_{1 \quad \delta}$ is the upper $1 \quad \delta$ quantile of a normal distribution which models how probable the approximation of the true sample size is [29]. The value $n$ is dynamically calculated at each step of the resampling until the number of resampled hypotheses is greater than $n$. Nevertheless, we bound the maximum number of resampled hypotheses to avoid drastic changes.

## III. Semantic Localization

Every time a submap is completed, the resulting map as well as the odometry measurements are used to compute a trajectory estimate. The weighted hypotheses allow for the creation of weighted mixtures of probability distributions resulting in a weighted fusion which considers the uncertainty of each hypothesis (cf. figure 4). The result of the fusion is formulated as a relative constraint and incorporated into a nonlinear factor graph as semantic landmarks.

### A. Semantic Evaluation of Submaps

Loop closures are identified by first evaluating the quality of the submap in terms of the occurred landmarks. This is motivated by the fact that in many cases (e.g. highways) it is not necessary to check for loop closures. The examination whether a submap is good enough for loop closure detection is based on a decision tree. We train a decision tree by comparing the trace of the state covariance before and after incorporating a specific region in the factor graph. The trained decision tree is specific to an urban environment and furthermore, to the length of the submap. Thus for other environments, a retraining of the decision tree or online learning approaches are required. A submap is considered as good either when it lowers the size of the bounding box or by having loop closures in it.

Evaluating a submap requires extracting descriptive attributes from it and we argue that semantic information is a crucial factor for this. In more detail, we first approximate the Shannon entropy $H$ of the mixture distribution using an approximate single multivariate Gaussian distribution over the submap, with covariance $\mathbf{\Sigma}$, i.e.

$$H = \frac{1}{2}\log\left((2\pi e)^3 \det(\mathbf{\Sigma})\right).$$

On a level of semantic classes we then calculate a term frequency-inverse document frequency (tf-idf) score, i.e.

$$S^i_{tf\text{-}idf} = \sum_c \frac{n^i_c}{n^i} \log\left(\frac{N}{n_c}\right),$$

where $n^i_c$ denotes the number of occurrences of class $c$ in submap $i$, $n^i$ the total number of classes in $i$. Furthermore, $N$ denotes the total number of submaps processed so far and $n_c$ represents the number of scenes within the submaps which included an object of type $c$. This is efficiently compared and updated with the previous submaps. As a final score, we make use of the number of landmarks within the submap.

As shown in figure 5, the loop closure detection is triggered once the decision tree predicts that a submap is potentially good in terms of its mapped objects.

### B. Semantic Loop Closure Detection

Loop closures are found in multiple steps. First, we find similar submaps using an incremental kd-tree [30] of the submap's normalized class histograms while employing the Jensen-Shannon divergence (JSD) [31] as the distance measure. For each similar submap the individual scene candidates

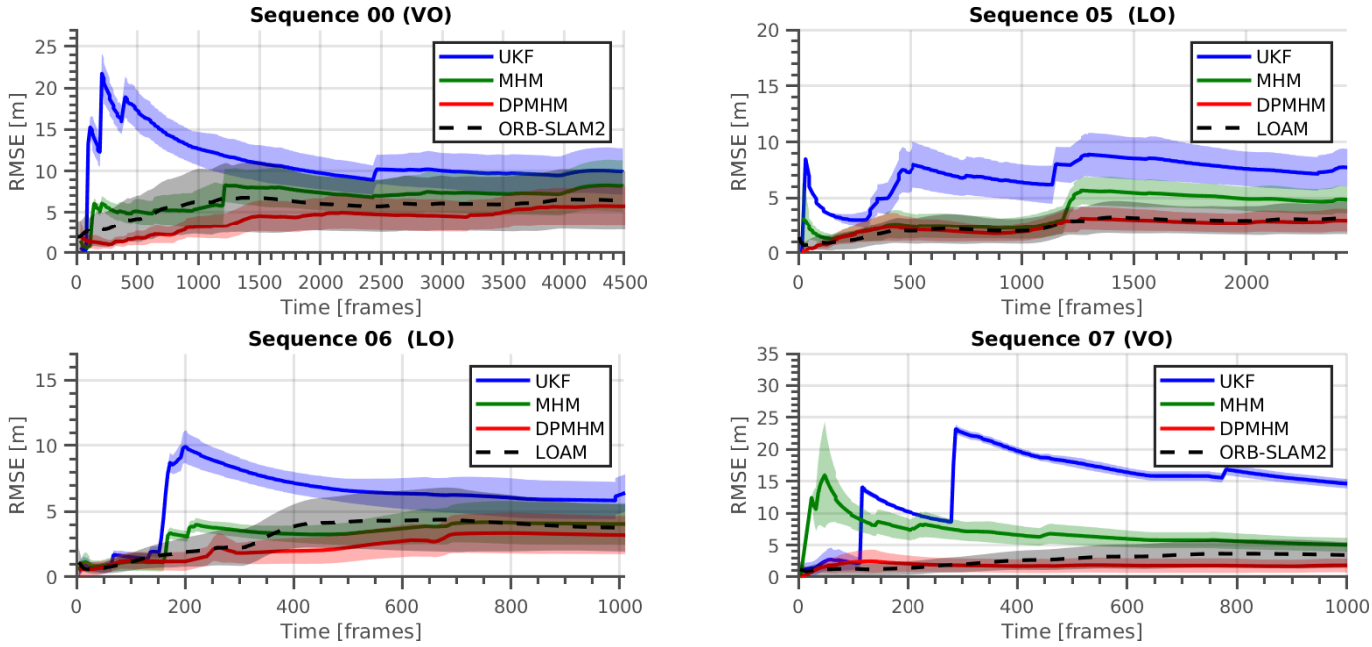


Fig. 6: Comparison of several KITTI sequences by means of the RMSE plotted as a function of the frames. The two methods, UKF and MHM, show several jumps in the error due to wrong data associations. Each wrong association pulls the factor graph towards a wrong direction which results in the jumps of the RMSE. Our method performs a more correct data association and keeps therefore the least error and does not include any sudden changes.

are identified with another kd-tree of the scene's normalized class histograms using the $L_2$-norm for faster retrieval. Additional efficiency can be achieved with a tuning parameter that restricts the search space of the kd-tree in terms of the distance.

Good loop closure candidates found by the second kd-tree are verified and further filtered with a discrete Bayes filter. We define a Markov chain between the events for loop closure and no loop closure. The transitional probabilities are chosen to be similar to [7].

The verification process calculates two scores for how similar the candidate scene and the current scene are. First, the topology of a scene is represented by the Laplacian matrix which is calculated based on the spatial relationship between the semantic classes as well as their degrees in the scene. We compare the topologies of two scenes based on a normalized cross correlation (NCC) [32] score, $S_{NCC}$. Second, another score, $S_{scene}$, expresses the overall similarity of the landmarks in the two scenes. For this the landmarks get associated with the Hungarian algorithm on the estimated landmark positions of each scene and their Euclidean distances. A pair of matched landmarks $i$ and $j$ contribute to $S_{scene}$ through

$$s^{i,j}_{match} := 1 \quad \frac{\mathbf{H}_{i,j}}{2}, \qquad s^{i,j}_{class} := (1 \quad \delta_{c_i,c_j}) \quad p,$$

where $\mathbf{H}$ is the Hungarian cost matrix (output of the Hungarian algorithm), $p$ denoting a penalty factor, $c_i$, $c_j$ being the label of the landmarks $i, j$, and $\delta$ denoting the Kronecker delta. The two scores $s_{match}$ and $s_{class}$ are combined using all matched landmark pairs, as follows

$$S_{scene} := \sum_{i,j} 1 \quad s^{i,j}_{match} s^{i,j}_{class}.$$

The sum of both scores, $S_{NCC}$ as well as $S_{scene}$ has to be larger than a threshold (tuning parameter) to verify the match of the two scenes. This binary decision serves as input to the discrete Bayes filter which finally gets to decide whether to use the scene pair as a loop closure candidate for the next step. As the last step, the set of all loop closures candidates undergoes a final geometric consistency check based on RANSAC before the actual loop closure constraints are inserted into the factor graph. Both, the Hungarian algorithm and RANSAC can be computationally expensive. Therefore, we filter most invalid candidates beforehand using the kd-trees which can be performed in logarithmic time. For additional robustness, we use m-estimators with Cauchy functions [33] in the optimization of the factor graph.

## IV. Evaluation

We evaluate our system on the KITTI dataset sequences 00, 05, 06 and 07 [34] where we use SegNet [35] to derive the semantic classes of the individual scenes. For each image, the semantic objects are extracted and projected into the world frame [5] using the Velodyne scans. In general, our approach is not limited to the use outdoors but rather depends on the object detector. Additionally, one might need to adapt the decision tree and $p_s$ in equation (2).

Since, to our knowledge, no appropriate approach for comparison is publicly available, we could not compare our proposed approach to another semantic SLAM system.

Therefore, we evaluate our proposed system to two other semantic solutions as well as two non-semantic approaches. As a baseline we compare to LeGo-LOAM [36] and ORB-SLAM2 stereo [37]. For a semantic baseline, we utilize a single UKF estimator with a Hungarian algorithm based on the $L_2$ norm for data association. This essentially performs a nearest neighbor data association with a single hypothesis. We also added a multiple hypothesis mapping (MHM) using a maximum likelihood approach and a MHt. Similar to our main approach, frequent optimizations of the MHt are needed. Hence, we threshold the likelihood if the MHt reaches a certain size (see equation (2)) and keep only the best third of all.

Our proposed system is agnostic to the source of odometry which we show by making use of two different ones for all sequences. More specifically, we utilized the tracking of ORB features [37] as well as LiDAR surface and corner features [36] to get an odometry estimate. We have used the provided camera calibration parameters from Geiger et al. [34] for both, Visual Odometry (VO) and ORB-SLAM2. Thus, the results of ORB-SLAM2 are different than the results reported in the work of Mur-Artal et al. [37] where they used different parameters per sequence.

### A. Results

We demonstrate the performance of our proposed approach by means of calculating the RMSE of the estimated trajectory location to the GPS ground truth provided by the KITTI dataset using the VO and Laser Odometry (LO) sources. Both, VO and LO, accumulate an error and hence, are subject to drift over time. Using our DP-based multiple hypothesis mapping approach together with our place recognition (cf. figure 5) we can reduce the drift up to 50 % for several sequences.

The simple UKF and MHM estimation approaches are strongly affected by wrong measurement associations resulting in bad constraints in the pose graph. These wrong assignments can be observed in figure 6 as sudden jumps in the RMSE. Consequently, the RMSE will have an increased total error which is even worse than the raw odometry source for a few sequences. Our approach is less perturbed with wrong associations and thus, maintains a more robust RMSE over time.

Table I show the mean and standard deviation of the RMSE for each sequence and estimator. Our approach does particularly well on on the longer sequences (00, 05) which results from the correct data association together with the semantic place recognition. Regardless of the odometry source, our proposed system yields results comparable to the state-of-the-art SLAM approaches in VO and LO and compared to the MHM, maintains less hypotheses about the environment as shown in figure 7. Due to the fact that the total number of hypotheses of the environment is only increased when the state uncertainty is high, we gain additional efficiency for our proposed system.

Figure 8 evaluates the performance of our algorithm when the semantic classes are removed as well as for the restriction to a single hypothesis. The single hypothesis, no semantic solution then still performs a probabilistic Hungarian method and achieves a mean RMSE of $5.45\,\mathrm{m}$ $2.94\,\mathrm{m}$.

| Sequence | 00 | 05 | 06 | 07 |
|---|---|---|---|---|
| VO | 8.41±2.51 | 6.42±3.9 | 3.8±1.4 | 6.23±2.4 |
| UKF | 11.14±2.9 | 8.9±4.2 | 5.96±2.4 | 14.55±5.1 |
| MHM | 6.84±1.3 | 5.6±2.94 | 3.17±1.07 | 6.93±2.02 |
| DPMHM | **4.54**±1.58 | **4.4**±2.3 | 2.3±0.74 | 2.9±4.5 |
| ORB-SLAM2 | 5.7±**1.0** | 4.51±**1.3** | **2.1**±**0.6** | **2.71**±**0.9** |
| LO | 7.33±2.5 | 2.96±1.3 | 3.3±1.2 | 6.65±2.85 |
| UKF | 6.96±1.7 | 6.96±1.5 | 5.47±0.75 | 10.4±2.8 |
| MHM | 5.3±2.2 | 3.8±1.4 | 2.67±**0.35** | 11.3±3.9 |
| DPMHM | **3.94**±**1.17** | **2.42**±**0.66** | 2.66±**0.35** | 5.5±2.4 |
| LeGo-LOAM | 5.8±2.2 | 2.54±0.72 | **2.15**±0.52 | **1.0**±**0.16** |

TABLE I: Comparison of the mean RMSE and standard deviation in meters achieved with VO and LO as the underlying odometry source.

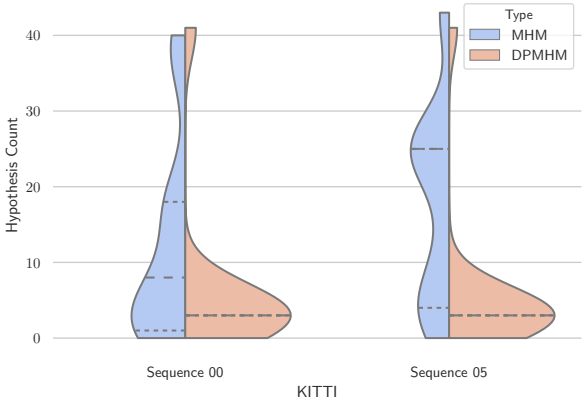


Fig. 7: Evaluation of the two multiple hypotheses-based implementations. The naive likelihood thresholding approach has an average of 12 (sequence 00) and 18 (sequence 05) hypotheses, respectively, whereas our proposed resampling approach has an average of 7 (sequence 00) and 6 (sequence 05) hypotheses.

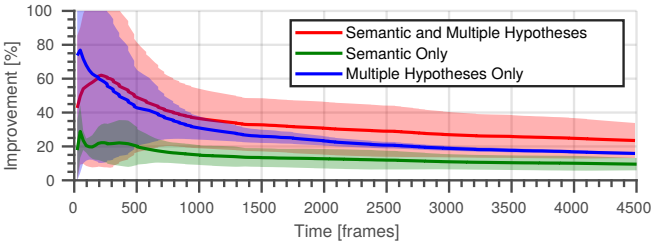


Fig. 8: Reduction of the RMSE for a single hypothesis, non-semantic DPMHM. Including both modalities, we achieve an average reduction of 34 %, with only multiple hypotheses 27 % and 13 % with a pure semantic DPMHM.

## V. Conclusion and Future Work

In this work, we presented a novel semantic SLAM system based on factor graphs and a MHt mapping approach aiming to deal with ambiguities in data association in semantic-based SLAM. We showed that our resampling method for optimizing the hypothesis tree yields a more robust estimation

and requires substantially less hypotheses. Moreover, we gain additional efficiency by preselecting submaps for loop closure detection.

As further research, we intend to remove the assumption that each object can generate at most one measurement per time-step since a bad detector or viewpoint angle might easily violate this assumption. Additionally, this work could potentially also be extended towards utilizing an instance-based detection. Instance information could possibly give a prior on how to associate the measurements at the cost of an additional non-Gaussian discrete random variable.

## REFERENCES


[1] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309–1332, 2016.

[2] A. Ess, T. Müller, H. Grabner, and L. J. Van Gool, "Segmentation-Based Urban Traffic Scene Understanding." in *BMVC*, vol. 1. Citeseer, 2009, p. 2.

[3] H. Blum, A. Gawel, R. Siegwart, and C. Cadena, "Modular sensor fusion for semantic segmentation," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 3670–3677.

[4] I. Kostavelis and A. Gasteratos, "Semantic mapping for mobile robotics tasks: A survey," *Robotics and Autonomous Systems*, vol. 66, pp. 86–103, 2015.

[5] A. Gawel, C. Del Don, R. Siegwart, J. Nieto, and C. Cadena, "X-View: Graph-Based Semantic Multi-View Localization," 2017.

[6] M. Labbe and F. Michaud, "Appearance-based loop closure detection for online large-scale and long-term operation," *IEEE Transactions on Robotics*, vol. 29, no. 3, pp. 734–745, 2013.

[7] A. Angeli, D. Filliat, J.-a. Meyer, A. Angeli, D. Filliat, J.-a. M. A. Fast, A. Angeli, D. Filliat, and J.-a. Meyer, "A Fast and Incremental Method for Loop-Closure Detection Using Bags of Visual Words," *IEEE Transactions on Robotics*, vol. 24, 2008.

[8] W. Chen, M. Fang, Y.-H. Liu, and L. Li, "Monocular semantic SLAM in dynamic street scene based on multiple object tracking," in *Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM), 2017 IEEE International Conference on*. IEEE, 2017, pp. 599–604.

[9] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," *arXiv preprint arXiv:1704.06857*, 2017.

[10] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. Kelly, and A. J. Davison, "SLAM++: Simultaneous localisation and mapping at the level of objects," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1352–1359, 2013.

[11] J. Civera, D. Galvez-Lopez, L. Riazuelo, J. D. Tardos, and J. M. M. Montiel, "Towards semantic SLAM using a monocular camera," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011, pp. 1277–1284.

[12] M. Hosseinzadeh, Y. Latif, T. Pham, N. Suenderhauf, and I. Reid, "Structure Aware SLAM using Quadrics and Planes," *arXiv preprint arXiv:1804.09111*, 2018.

[13] J. Elfring, S. Van Den Dries, M. J. Van De Molengraft, and M. Steinbuch, "Semantic world modeling using probabilistic multiple hypothesis anchoring," *Robotics and Autonomous Systems*, vol. 61, no. 2, pp. 95–105, 2013.

[14] L. L. Wong, L. P. Kaelbling, and T. Lozano-Pérez, "Data association for semantic world modeling from partial views," *Springer Tracts in Advanced Robotics*, vol. 114, pp. 431–448, 2016.

[15] S. S. Blackman, "Multiple hypothesis tracking for multiple target tracking," *IEEE Aerospace and Electronic Systems Magazine*, vol. 19, no. 1, pp. 5–18, 2004.

[17] L. L. S. Wong, T. Kurutach, L. P. Kaelbling, and T. Lozano-Pérez, "Object-based World Modeling in Semi-Static Environments with Dependent Dirichlet-Process Mixtures," pp. 1–23, 2015.

[16] I. Cox and S. Hingorani, "An Efficient Implementation and Evaluation of Reid's Multiple Hypothesis Tracking Algorithm for Visual Tracking," in *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, no. 1, 1994, pp. 437–442.

[18] N. Atanasov, M. Zhu, K. Daniilidis, and G. J. Pappas, "Localization from semantic observations via the matrix permanent," *International Journal of Robotics Research*, vol. 35, no. 1-3, pp. 73–99, 2016.

[19] A. Nüchter and J. Hertzberg, "Towards semantic maps for mobile robots," *Robotics and Autonomous Systems*, vol. 56, no. 11, pp. 915–926, 2008.

[20] A. Pronobis and P. Jensfelt, "Large-scale semantic mapping and reasoning with heterogeneous modalities," in *2012 IEEE International Conference on Robotics and Automation*. IEEE, 2012, pp. 3515–3522.

[21] S. L. Bowman, N. Atanasov, K. Daniilidis, and G. J. Pappas, "Probabilistic data association for semantic SLAM," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 1722–1729, 2017.

[22] L. Nicholson, M. Milford, and N. Sünderhauf, "QuadricSLAM: Dual Quadrics as SLAM Landmarks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 313–314.

[23] E. A. Wan and R. Van Der Merwe, "The unscented Kalman filter for nonlinear estimation," in *IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium, AS-SPCC 2000*, vol. 7, 2000, pp. 153–158.

[24] Y. Bar-Shalom, S. S. Blackman, and R. J. Fitzgerald, "Dimensionless score function for multiple hypothesis tracking," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 43, no. 1, pp. 392–400, 2007.

[25] R. Jonker and T. Volgenant, "Improving the Hungarian assignment algorithm," *Operations Research Letters*, vol. 5, no. 4, pp. 171–175, 1986.

[26] A. Ranganathan and F. Dellaert, "A rao-blackwellized particle filter for topological mapping," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2006, no. May, pp. 810–817, 2006.

[27] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman Filter: Particle Filters for Tracking Applications*, 2004.

[28] G. Grisetti, C. Stachniss, and W. Burgard, "Improving Grid Based SLAM with Rao Blackwellized Particle Filters by Adaptive Proposals and Selective Resampling," *International Conference on Robotics and Automation*, no. April, pp. 2443–2448, 2005.

[29] D. Fox, "Adapting the sample size in particle filters through KLD Sampling," *Intl Jour of Robotics Research*, vol. 22, no. 12, pp. 985–1004, 2003.

[30] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509–517, 1975.

[31] J. Lin, "Divergence Measures Based on the Shannon Entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.

[32] S. Cascianelli, G. Costante, E. Bellocchio, P. Valigi, M. L. Fravolini, and T. A. Ciarfuglia, "Robust visual semi-semantic loop closure detection by a covisibility graph and CNN features," *Robotics and Autonomous Systems*, vol. 92, pp. 53–65, 2017.

[33] G. H. Lee, F. Fraundorfer, and M. Pollefeys, "Robust pose-graph loop-closures with expectation-maximization," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*. IEEE, 2013, pp. 556–563.

[34] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361, 2012.

[35] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[36] T. Shan and B. Englot, "LeGO-LOAM: Lightweight and ground-optimized lidar odometry and mapping on variable terrain," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 4758–4765.

[37] R. Mur-Artal and J. D. Tardos, "ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.