2019 International Conference on Robotics and Automation (ICRA)
Palais des congres de Montreal, Montreal, Canada, May 20-24, 2019

Multimodal Semantic SLAM with Probabilistic Data Association

Kevin Doherty, Dehann Fourie, and John Leonard

***Abstract*— The recent success of object detection systems motivates object-based representations for robot navigation; i.e. semantic simultaneous localization and mapping (SLAM). The semantic SLAM problem can be decomposed into a discrete inference problem: determining object class labels and measurement-landmark correspondences (the data association problem), and a continuous inference problem: obtaining the set of robot poses and object locations in the environment. A solution to the semantic SLAM problem necessarily addresses this joint inference, but under ambiguous data associations this is in general a non-Gaussian inference problem, while the majority of previous work focuses on Gaussian inference. Previous solutions to data association either produce solutions *between* potential hypotheses or maintain multiple explicit hypotheses for each association. We propose a solution that represents hypotheses as multiple modes of an equivalent non-Gaussian sensor model. We then solve the resulting non-Gaussian inference problem using nonparametric belief propagation. We validate our approach in a simulated hallway environment under a variety of sensor noise characteristics, as well as using real data from the KITTI dataset, demonstrating improved robustness to perceptual aliasing and odometry uncertainty.**

I. Introduction

As object detectors continue to improve, there has been growing interest in their use in conjunction with a suite of inertial and geometric sensors to improve robot navigation, allowing autonomous robots to build more accurate, descriptive maps [1], [2]. However, the addition of discrete categorical sensor measurements from an object detector with continuous sensing modalities poses a challenge in simultaneous localization and mapping (SLAM), where traditional methods assume that all measurement likelihoods are Gaussian. Addressing the combined discrete-continuous problem is necessary for any semantic SLAM system that incorporates discrete object categories. This paper presents a novel solution to the problem of jointly inferring landmark positions and classes, robot poses, and data associations.

In this work, we explore the representation of uncertainty due to data association and landmark class ambiguity in the semantic SLAM problem. We are specifically concerned with full posterior inference of all robot poses and landmarks, relaxing the Gaussian assumption of typical SLAM frameworks, and incorporating discrete measurements from an object detector as probabilistic data associations, which introduce multiple modes in otherwise Gaussian measurement models. We aim to approximate the non-Gaussian posterior, explicitly marginalizing out discrete variables. To perform non-Gaussian inference, we make use of multimodal incremental smoothing and mapping (mm-iSAM) [3] which performs nonparametric belief propagation [4].

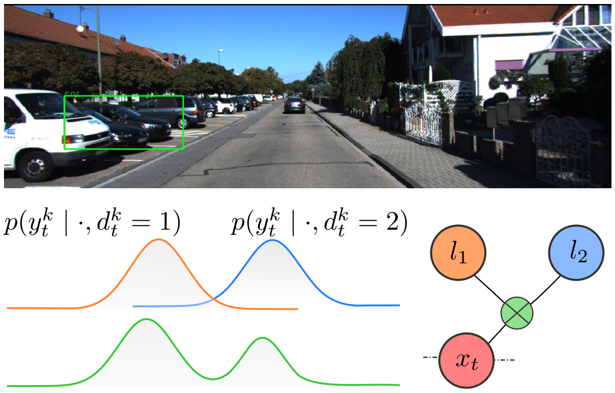

$p(y_t^k \mid \cdot, d_t^k = 1)$
$p(y_t^k \mid \cdot, d_t^k = 2)$
l_1
l_2
x_t

Fig. 1: Multiple modes arise from data association ambiguity between two landmarks (1 and 2), in which $p(d_t^k = 1) > p(d_t^k = 2)$. *Top:* Ambiguity in an object detection results from occlusion and objects in close proximity. *Bottom-left:* Associations represented as a non-Gaussian sensor model. *Bottom-right*: Ambiguous measurements are incorporated into a factor graph as *multimodal semantic factors* (green).

The key insight of our approach is that measurement models, even those well-represented by Gaussian approximations, under data association and landmark class ambiguity can be represented by a non-Gaussian measurement model, as shown in Figure 1. Our primary contributions are as follows:

1) We provide a nonparametric belief propagation solution to full posterior inference for the semantic SLAM problem with ambiguous data associations.
2) We describe *multimodal semantic factors*, which allow us to incorporate uncertainty in data association and semantics as non-Gaussian factors in a factor graph and solve the continuous optimization over poses and landmarks using mm-iSAM.
3) We experimentally validate our approach in a simulated navigation task, as well as with the KITTI dataset, demonstrating robustness to data association and landmark class ambiguity.

The remainder of this paper proceeds as follows. In Section II, we review related efforts towards the problems of data association, non-Gaussian SLAM, and semantic SLAM. We describe the semantic SLAM and data association problems and introduce our approach in Section III. We give background on mm-iSAM and the proposed multimodal semantic factors in Section IV. Finally, experimental results validating our approach in a simulated environment and on real data

K. Doherty, D. Fourie, and J. Leonard are with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA {kdoherty, dehann, jleonard}@mit.edu

978-1-5386-6026-3/19/$31.00 ©2019 IEEE

from the KITTI dataset are provided in Section V.

II. Related Work

A. Data Association and Non-Gaussian SLAM

Early work on probabilistic data association (PDA) as a representation for ambiguous hypotheses stems from target-tracking literature, where it was incorporated into the "probabilistic data association filter" [5]. Similar filtering-based approaches to multi-hypothesis tracking (MHT) originated around the same time [6], later adapted to the SLAM problem [7], [8]. These approaches seek to explicitly represent several plausible hypotheses and over time "prune" those which become unlikely. In general these works focus on solutions to problems with linear (or approximately linear) dynamics and Gaussian noise models.

In the SLAM literature, FastSLAM [9] represents multiple hypotheses using a particle filter-based algorithm in which data association probabilities are computed separately for each particle representing a candidate robot state. Conceptually, FastSLAM is similar to our approach, but maintains separate parametric solutions each using extended Kalman filters (EKFs). In contrast, our approach directly approximates the non-Gaussian solution to the SLAM problem under ambiguous association. A similar approach focusing on filtering-based SLAM is the *sum of Gaussians* method described by Durrant-Whyte et al. [10]. The multimodal iSAM framework we leverage in this work can be viewed as a smoothing analogue to the sum of Gaussians approach.

A number of works in *robust SLAM* address the problem of SLAM with non-Gaussian noise. Sunderhauf et al. [11] introduce discrete switching variables which are estimated on the back-end to determine whether loop closure proposals from the front-end are accepted. We instead marginalize over all discrete variables and focus on functional approximation of the resulting non-Gaussian distribution. Pfingsthorn and Birk [12] proposed a maximum-likelihood optimization for multimodal distributions with a similar measurement representation to ours and the sum of Gaussians filter. The max-mixtures approach of Olson and Agarwal [13] side-steps the complexity typically associated with multi-hypothesis SLAM by selecting the most likely component of a mixture of Gaussians at all points in the measurement domain, rather than maintaining a potentially large number of individual parametric solutions. These works focus on the behavior of their respective approaches in the context of maximum likelihood, while our primary concern is to infer the full posterior distribution over poses and landmarks.

B. Semantic SLAM

The ability of semantic measurements obtained from an object detector to aid in data association when there is ambiguity in purely geometric features links the problems of semantic SLAM and data association. The majority of works in semantic SLAM thus far have considered questions of geometric representation and make use of variants on maximum-likelihood data association [14], [15], [16], [17], [18]; in this work, we instead opt for a simple geometric representation and focus on representing the multimodalities induced in the posterior by ambiguous data associations and unknown landmark classes. Bowman *et al.* [19] recently showed that the discrete problems of landmark class inference and data association could be combined and provided an expectation-maximization (EM) solution which replaces the marginalization over data associations in the PDA method with a geometric mean, preserving the Gaussian assumption. The EM formulation provably converges at least to a local optimum when iterated, but for computational reasons, it is undesirable to recompute the combinatorial number of plausible data associations for previous poses. We also proactively compute data association probabilities, but marginalize out data associations and perform nonparametric inference in the factor graph, retaining the plausible modes.

III. Semantic SLAM with Ambiguous Data Associations

In the following sections, we give an overview of the semantic SLAM and data association problems and introduce our solution for posterior inference in the joint data association and semantic SLAM problem.

A. Semantic SLAM with Known Data Associations

In its general form, the problem of semantic SLAM is the estimation of the most probable set of robot poses $\mathcal{X} \triangleq \{x_t\}_{t=1}^T$ and landmark positions and semantic classes $\mathcal{L} \triangleq \{(l^p, l^c)_j\}_{j=1}^M$ given a set of sensor measurements $\mathcal{Z} \triangleq \{\mathcal{Z}_t\}_{t=1}^T$ made at each robot pose. We have $x_t \in SE(2)$ in the two-dimensional case, and $x_t \in SE(3)$ in 3D. Similarly, we take $l^p \in \mathbb{R}^2$ in 2D and $\mathbb{R}^3$ in 3D. The landmark class l^c is assumed to come from a finite set of discrete, known, class labels: $\mathcal{C} = \{1, 2, \dots, C\}$.

Landmarks and poses in the SLAM problem are generally inferred using a maximum-likelihood approach, i.e.

$$\hat{\mathcal{X}}_{ML}, \hat{\mathcal{L}}_{ML} = \underset{\mathcal{X},\mathcal{L}}{\operatorname{argmax}}\, p(\mathcal{Z} \mid \mathcal{X}, \mathcal{L}). \tag{1}$$

In our case, we aim to infer the posterior over latent variables $\mathcal{X}$ and $\mathcal{L}$: $p(\mathcal{X}, \mathcal{L} \mid \mathcal{Z})$. Once we have obtained the posterior, we can obtain maximum *a posteriori* (MAP) estimates for any of the variables as

$$\hat{\mathcal{X}}_{MAP}, \hat{\mathcal{L}}_{MAP} = \underset{\mathcal{X},\mathcal{L}}{\operatorname{argmax}}\, p(\mathcal{X}, \mathcal{L} \mid \mathcal{Z}). \tag{2}$$

In this work, we will assume $\mathcal{Z}_t$ consists of a set of inertial measurements and object measurements, denoted $\mathcal{U}_t$ and $\mathcal{Y}_t$, respectively. We assume object measurements consist of a geometric component, e.g. range and bearing, as well as a semantic component; such information can be easily obtained using, for example, a stereo camera in conjunction with an object detector, which we discuss further in Sections IV-B and V.

B. Probabilistic Data Association

When correspondences between measurements and landmarks are not known *a priori*, they must also be inferred. Let d_t^k denote a data association for measurement k taken at pose x_t, such that $d_t^k = j$ signifies that measurement z_t^k corresponds to landmark l_j. Let $\mathcal{D} \triangleq \{\mathcal{D}_t\}_{t=1}^T$ denote the set of all associations of measurements to landmarks. One of the most common solutions to the data association problem is maximum-likelihood estimation. That is, given an initial estimate of poses and landmarks $\mathcal{X}^{(0)}$ and $\mathcal{L}^{(0)}$, respectively, maximum-likelihood data association performs the following optimization:

$$\hat{\mathcal{D}} = \operatorname*{argmax}_{\mathcal{D}} p(\mathcal{D} \mid \mathcal{X}^{(0)}, \mathcal{L}^{(0)}, \mathcal{Z}) \tag{3}$$

$$\hat{\mathcal{X}}_{ML}, \hat{\mathcal{L}}_{ML} = \operatorname*{argmax}_{\mathcal{X},\mathcal{L}} p(\mathcal{Z} \mid \mathcal{X}, \mathcal{L}, \hat{\mathcal{D}}). \tag{4}$$

Maximum-likelihood data associations are computed and fixed, then the SLAM solution is optimized assuming the fixed set of data associations. This method is typically performed in a proactive fashion, after each state update, for example using the Hungarian algorithm [20] or joint compatibility branch and bound [21] to simultaneously compute an optimal assignment of all measurements $z_t^k,\ k = 1, \ldots, K_t$ observed at a pose x_t to landmarks. While very efficient and easy to implement, this method can be brittle.

An alternative solution is to consider *probabilistic* data associations. If we had access to the probability of each data association, we could marginalize out data associations when computing the solution to the SLAM problem,

$$\hat{\mathcal{X}}_{ML}, \hat{\mathcal{L}}_{ML} = \operatorname*{argmax}_{\mathcal{X},\mathcal{L}} \mathbb{E}_{\mathcal{D}}\left[p(\mathcal{Z} \mid \mathcal{X}, \mathcal{L}, \mathcal{D})\right]. \tag{5}$$

The approximate marginal distribution represented by this expectation over data association hypotheses—even when the individual measurement likelihoods are well-represented by Gaussian distributions—is almost always multimodal in practice. Consequently, maximum likelihood estimation finds a suboptimal solution somewhere between plausible modes. Other solutions maintain a potentially exponential set of Gaussian solutions that branch with each set of new hypotheses. Work in the latter area has primarily focused on methods to prune the space of plausible hypotheses (e.g. [8], [22], and more recently [23]).

A recent solution making use of expectation-maximization iterates between computing the data association probabilities and the conditional log-likelihood [19]:

$$\hat{\mathcal{X}}_{ML}^{(i+1)}, \hat{\mathcal{L}}_{ML}^{(i+1)} = \operatorname*{argmax}_{\mathcal{X},\mathcal{L}} \mathbb{E}_{\mathcal{D}}\left[\log p(\mathcal{Z} \mid \mathcal{X}, \mathcal{L}, \mathcal{D}) \mid \mathcal{X}_{ML}^{(i)}, \mathcal{L}_{ML}^{(i)}, \mathcal{Z}\right]. \tag{6}$$

This effectively replaces the sum of Gaussians in the marginal with a geometric mean, preserving the Gaussian assumption, and iterating in this fashion provides guaranteed convergence; we refer to this method as *Gaussian PDA*. This approach solves for point estimates of poses and landmarks at each iteration effectively using weighted maximum likelihood estimation [24], where weights are determined by estimated data association probabilities. In practice, recomputing data association probabilities for all previous measurements is a computational burden, so typically data association probabilities are computed once proactively (i.e. after each keyframe), resulting in solutions somewhere *between* the modes induced by the plausible association hypotheses. For posterior inference, we would like to consider the *distribution* over possible poses and landmarks at every time step and account for possible multimodalities.

To solve the problem of semantic SLAM with ambiguous data associations, we consider an alternative representation in which we first marginalize out poses and landmarks to compute data association probabilities, then marginalize out data associations to obtain a distribution over poses and landmarks. That is, letting $\Theta \triangleq \{\mathcal{X}, \mathcal{L}\}$, given a previous estimate of the distribution over landmarks and poses $\hat{p}^{(i)}(\Theta)$, we compute the marginal probability for each set of data associations:

$$\hat{p}^{(i+1)}(\mathcal{D}) = \eta_{\mathcal{D}} \int p(\mathcal{Z} \mid \mathcal{D}, \Theta)\hat{p}^{(i)}(\Theta) d\Theta, \tag{7}$$

where $\eta_{\mathcal{D}}$ is a normalizing constant, then update the estimate of the posterior over poses and landmarks:

$$\hat{p}^{(i+1)}(\Theta) = \mathbb{E}_{\mathcal{D}\sim\hat{p}^{(i+1)}(\mathcal{D})}\left[p(\Theta \mid \mathcal{Z}, \mathcal{D})\right] \tag{8}$$

$$\propto \hat{p}^{(i)}(\Theta) \sum_{\mathcal{D}} p(\mathcal{Z} \mid \Theta, \mathcal{D})\hat{p}^{(i+1)}(\mathcal{D}). \tag{9}$$

We proactively compute data association probabilities when measurements are received and consider a single iteration of this approach, which relieves the computational challenge of recomputing the set of data association probabilities for all previous measurements after every observation.

IV. Multimodal Semantic SLAM

We have thus far introduced a solution to posterior inference for semantic SLAM relying on alternating computation of marginals over data associations and robot poses and landmarks. In this section, we describe nonparametric belief propagation, which we use to obtain the approximate (non-Gaussian) marginals in our solution. We then show how data association and landmark class ambiguity can be represented as *multimodal semantic factors* that we incorporate into a factor graph and solve using nonparametric belief propagation.

A. Multimodal iSAM

We use multimodal iSAM [3] to compute the posterior over poses and landmarks given a factor graph. In the factor graph representation of the SLAM problem, we can compute the posterior over poses and landmarks as follows:

$$p(\mathcal{X}, \mathcal{L} \mid \mathcal{Z}) \propto \prod_{\varphi} \varphi(\mathcal{X}, \mathcal{L}, \mathcal{Z}) \prod_{\psi} \psi(\mathcal{X}, \mathcal{L}), \tag{10}$$

where φ denotes a measurement factor and ψ denotes a prior factor. Here, the factor graph is an undirected graphical model where poses and landmarks are latent variables linked by measurement factors and priors. The marginal distribution over each variable can be obtained using belief propagation, a solution which yields a convenient analytical form when all of the variables are Gaussian.

To accommodate non-Gaussian variables in the factor graph, multimodal iSAM makes use of nonparametric belief propagation [4]. Nonparametric belief propagation approximates the belief over all continuous state variables absent the assumption of Gaussianity using a combination of Gibbs sampling and kernel density estimation. That is, for a random variable X, we approximate the marginal over X as

$$\hat{p}(X) = \sum_{n=1}^{N} w^{[n]} \mathcal{N}\left(x^{[n]}, \Sigma^{[n]}\right), \tag{11}$$

where $\mathcal{N}$ is a multivariate Gaussian kernel, each kernel is centered at a sample $x^{[n]}$, $w^{[n]}$ is the weight associated with the n-th kernel, and $\Sigma^{[n]}$ is the associated Gaussian kernel bandwidth, determined using leave-one-out cross-validation. The weights $w^{[n]}$ are chosen uniformly such that the resulting sum is a valid probability density function.

A beneficial aspect of the functional approximation of marginals is that we no longer need to explicitly represent the potentially many modes in the posterior. This *implicit* representation decouples the complexity of inference from the number of hypotheses, as the computation involved in the approximation of the marginals depends only on a fixed number of samples. The result is that modes with very low probability are unlikely to be represented in the approximate marginal density. However, we do not explicitly prune these modes, and since they still exist in the factor graph, modes which later become more probable can be recovered.

B. Multimodal Semantic Factors

To incorporate semantic measurements into the factor graph, we use multimodal semantic factors, which introduce constraints between a pose and potentially many landmarks. We assume a factorized semantic measurement model $p(y_t^k \mid x_t, l_j) = p(y_t^{k,c} \mid l_j^c) p(y_t^{k,r} \mid x_t, l_j) p(y_t^{k,b} \mid x_t, l_j)$ consisting of the class estimate $y_t^{k,c}$ from an object detector, the estimated range to the object $y_t^{k,r}$, and the estimated bearing to the object $y_t^{k,r}$. The distribution $p(y_t^{k,c} \mid l_j^c)$ correspond to the confusion matrix for the classifier, learned offline, while $p(y_t^{k,r} \mid x_t, l_j)$ and $p(y_t^{k,b} \mid x_t, l_j)$ are each assumed Gaussian with means $y_t^{k,r}$ and $y_t^{k,b}$ and variances $\sigma_t^{2,k,r}$ and $\sigma_t^{2,k,b}$, respectively. The latter terms can be obtained by considering the range and bearing to the set of 3D points estimated by a stereo vision system which project into the bounding box for the object detection corresponding to measurement y_t^k.

At each time step t, we update the factor graph solution according to (10), which provides marginals for all poses $x_{1:t}$ and known landmarks. Given the semantic measurement model, we compute the probability of an association as the total posterior probability of all associations at time t of measurement k to landmark j given the measurements. Specifically, let $\mathbb{D}_t$ denote the set of all possible associations of measurements at time t to known landmarks. Similarly, define $\mathbb{D}\left\{d_t^k = j\right\} \triangleq \{\mathcal{D}_t \in \mathbb{D}_t \mid d_t^k = j\}$, the set of all possible data associations at time t in which measurement k is associated to landmark j. Assuming a uniform prior on data associations, we have:

$$\hat{p}(d_t^k = j) = \eta_{\mathcal{D}} \sum_{\mathcal{D}_t \in \mathbb{D}\left\{d_t^k = j\right\}} \prod_k p(y_t^k \mid \mathcal{D}_t), \tag{12}$$

where $\eta_{\mathcal{D}}$ is the total probability of the set $\mathbb{D}$. We compute the likelihood of each measurement y_t^k given its association d_t^k by marginalizing out the pose estimate at x_t and the landmark position and class[1]:

$$\begin{aligned} p(y_t^k \mid \mathcal{D}_t) &= \iint p(y_t^k \mid x_t, l_{d_t^k}, d_t^k) \hat{p}(x_t) \hat{p}(l_{d_t^k}) dx_t dl_{d_t^k} \\ &\approx \sum_{n=1}^{N} \int p(y_t^k \mid x_t^{[n]}, l_{d_t^k}, d_t^k) \hat{p}(x_t^{[n]}) \hat{p}(l_{d_t^k}) dl_{d_t^k}, \end{aligned} \tag{13}$$

where we have replaced the integral over the pose distribution by a sampled approximation. For data association computation, we adopt a maximum likelihood sensor model to simplify the integral over the landmark position. We find this works well, empirically, when the sensor model is Gaussian, but non-Gaussian sensor models can be accommodated by making a sample-based approximation, for example.

Given $\hat{p}(d_t^k = j)$ for all landmarks l_j in a set $\mathcal{J} \subseteq \mathcal{L}$ of candidate landmarks, a multimodal semantic factor links a pose x_t and each candidate in $\mathcal{J}$:

$$\varphi_t^k(\mathcal{X}, \mathcal{L}, \mathcal{Z}) = \sum_{l_j \in \mathcal{J}} p(y_t^k \mid x_t, l_j, d_t^k = j) \hat{p}(d_t^k = j), \tag{14}$$

which for a Gaussian measurement model is a weighted sum of Gaussians.

Finally, MAP estimates for each landmark class, assuming a uniform prior, can be computed as in [19]:

$$l_j^c = \operatorname*{argmax}_c \prod_t \sum_{\mathcal{D}_t} p(\mathcal{D}_t, l_j^c = c \mid \mathcal{Z}), \tag{15}$$

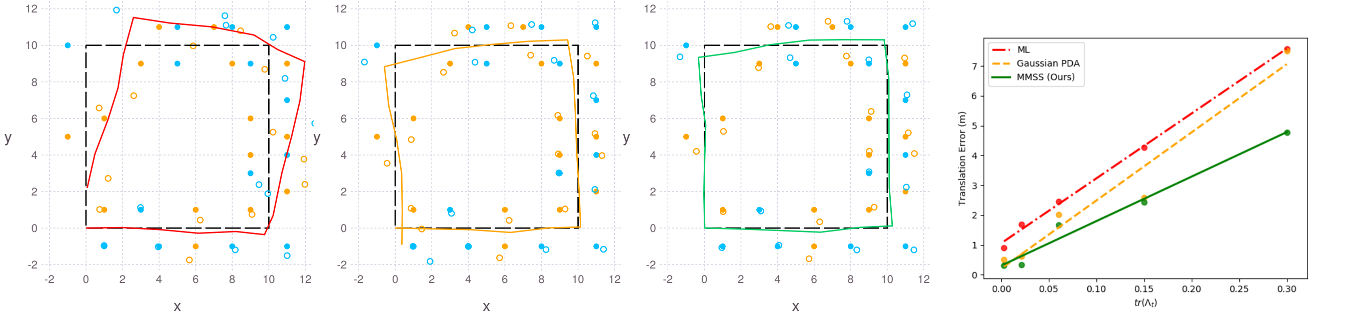
which are obtained by maximizing over the measurement likelihoods found using Equation 13 with respect to the landmark classes, rather than marginalizing them out, and instead marginalizing out data associations.

V. Experimental Results

Experiments with mm-iSAM were implemented in the Julia programming language using the Caesar.jl library[2]. We demonstrate the proposed approach both in simulation, with a hallway environment, and using real data from the KITTI dataset [25], [26]. All experiments were run offline using 10 cores of a 2.2 GHz i7 CPU and factor graph computation time was roughly identical across the three methods (approximately 1 minute for simulated data and

[1] The integral with respect to $l_{d_t^k}$ is a combined discrete summation over the possible landmark classes in $\mathcal{C}$ and integral over the domain of landmark positions, e.g. $\mathbb{R}^2$ in the 2-dimensional case.

[2] `https://github.com/JuliaRobotics/Caesar.jl`



(a) Maximum-Likelihood (b) Gaussian PDA (c) MMSS (Ours) (d) Avg. translation error vs. $tr(\Lambda_t)$

Fig. 2: (a-c): Comparison of trajectories estimated using each approach in a simulated hallway environment. Ground-truth trajectories are shown as dashed black lines. Ground-truth landmarks are shown as circles and colored by semantic class. Landmark position estimates from each method are shown as rings and colored similarly by class. (d) Comparison of translation error on simulated navigation tasks under five odometry noise models, Λ_t, with best fit line for each method.

3 minutes for the KITTI dataset). In both tests, we compared our method, multimodal semantic SLAM (MMSS) with maximum-likelihood (ML) data association and Gaussian probabilistic data (Gaussian PDA). The ML method selects the maximum-likelihood association considering all measurements in a keyframe. We implement the Gaussian PDA method using Gaussian factors with variance inversely weighted by data association probabilities[3].

In practice, new landmarks are determined using a threshold on their likelihood given each known landmark (similar to a Mahalanobis distance threshold in the Gaussian case) and we compute data association probabilities for each candidate landmark within a conservative range of the estimated pose at time t (this determines the set $\mathcal{J}$ in Equation 14).

A. Simulated Data

Our simulated navigation experiments consist of a two-dimensional hallway environment with landmarks of two classes. The robot in this simulation makes noisy measurements to each landmark within its limited field of view (120° up to 3.5 m), and each range measurement has an associated distribution over class probabilities. We model semantic measurements as samples from a categorical model having a confusion matrix with 90% accuracy for all landmark classes. Range and bearing measurements were corrupted with zero-mean Gaussian noise with variance 0.01. We also simulate an odometry model corrupted by Gaussian noise with diagonal covariance Λ_t, which we vary in our experiments.

In Figures 2a-c, trajectories and landmark estimates from each method are compared qualitatively for a simulated run with $\Lambda_t = \mathrm{diag}(0.01; 0.01; 0.001)$. In this example, we find that ML data association fails in the presence of substantial perceptual aliasing. Both Gaussian PDA and our method are more robust to errors in data association, but we find that ours is the only method that accurately closes the loop after executing the full trajectory. In Figure 2d, we show the average trajectory error for the three methods, plotted against $tr(\Lambda_t)$. Error for our approach increases the least as the odometry becomes more noisy, suggesting improved robustness to odometry uncertainty.

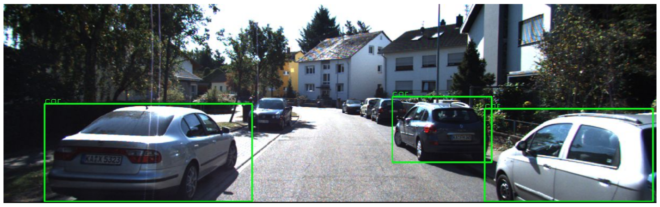


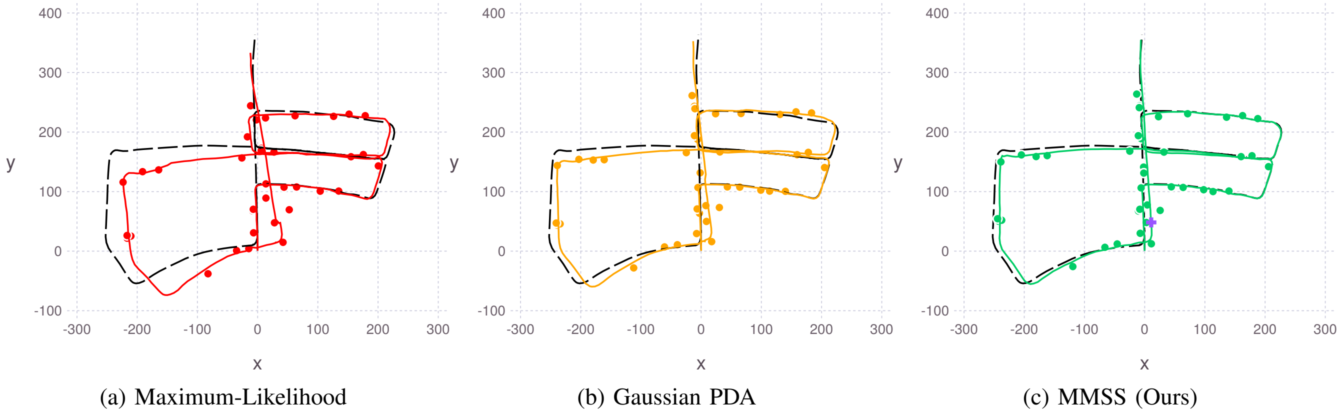
Fig. 3: Object detections from MobileNet-SSD on KITTI sequence 5.

B. Real Data

We evaluated the three approaches for a navigation using a stereo camera with data from KITTI odometry sequence 5 [25]. Odometry is provided by VISO2 stereo odometry [27], and probabilistic data associations with objects provide loop closures. We sample keyframes at 1 Hz and objects are detected in the left camera image using the MobileNet-SSD neural network [28] (with the single-shot detector (SSD) and MobileNets proposed respectively in [29] and [30]) trained on the PASCAL Visual Object Categories (VOC) dataset [31]. We accept measurements for which the neural network reports a confidence greater than 0.8. Examples of bounding box measurements from the object detector are shown in Figure 3. Semantic measurements are produced in the KITTI dataset by detections of cars and are represented by the average range and bearing to all 3D points that project into the detection bounding box. We assume that the stereo pair has fixed height and is constrained in pitch and roll, so the resulting estimation procedure is carried out with respect to the vehicle translation along the ground plane and yaw.

Figure 4 shows estimated trajectories and landmark positions for each method on KITTI sequence 5, and corresponding average translation and rotation errors can be

[3] Our implementation of the Gaussian PDA method uses the approximate marginal likelihood of each observation to compute data association probabilities, rather than a point estimate of poses and landmarks; thus, it can be viewed as an extension of the EM formulation in [19] from maximum-likelihood estimation to MAP estimation.



(a) Maximum-Likelihood (b) Gaussian PDA (c) MMSS (Ours)

Fig. 4: Comparison of trajectories (lines) and landmark position estimates (points) for each method applied to KITTI sequence 5. Ground truth trajectory is plotted as a black dashed line. The contour plot for the pose marked with a purple cross in (c) is shown in Figure 5.

Method	Avg. Trans. Error (m)	Avg. Rot. Error (rad)
ML	20.427	0.0810
GPDA	8.814	0.0446
MMSS (Ours)	**5.718**	**0.0255**

TABLE I: Comparison of translation and rotation error on KITTI sequence 5 for the different methods tested.

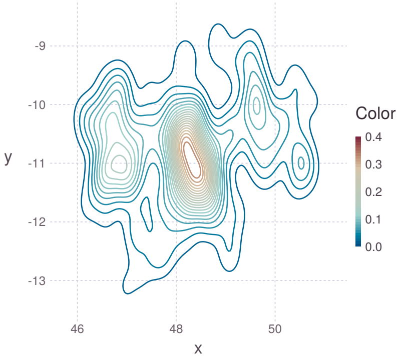


Fig. 5: Contour plot for the marginal distribution of the marked pose in Figure 4c. Multimodality is induced by odometry uncertainty and data association ambiguity.

found in Table I. As a result of perceptual aliasing due to long rows of parked cars, maximum-likelihood associations cause a number of incorrect loop closures that are hard to recover from. Gaussian PDA makes “soft” measurements in these cases, and produces a much better solution. By representing the full posterior, however, our method obtains a more accurate solution, recovering from the uncertainty in growth in the largest loop. We additionally mark a pose near this loop closure in Figure 4c and display the contour plot of its distribution in Figure 5, which shows that odometry uncertainty coupled with data association ambiguity results in a non-Gaussian posterior. A supplemental video provides visualization of the object detections and estimated vehicle trajectory using our approach on the KITTI dataset[4].

VI. Conclusion and Future Work

We have proposed a solution to semantic SLAM with unknown data associations that implicitly represents multiple association hypotheses as a multimodal sensor model. This formulation leads to a non-Gaussian SLAM problem, which we solve using mm-iSAM [3]. We validated our approach on a simulated navigation task under variety of odometry noise characteristics, as well as on data from the KITTI dataset. In addition to representing non-Gaussian belief over poses and landmarks, our multimodal semantic SLAM approach showed improved robustness to odometry noise and perceptual aliasing as compared with other methods.

Though our method represents uncertain associations, like many previous efforts, we rely on hard decisions about whether or not to add landmarks. Representing this uncertainty is an important step toward more tightly coupling the data association and SLAM problems, but computation of association probabilities may become expensive. Dirichlet process priors on associations, as in [32] provide one avenue for future work, while the approximate matrix permanent methods of [33] may help address computational complexity.

Our approach also enables semantic SLAM with non-traditional sensing modalities. By choosing a representation that does not make assumptions about the measurement distribution, we are able to deal with ambiguous data associations that arise from non-Gaussian sensor models, for example in the case of multiple returns by a sonar.

Finally, we assumed a simple geometric model and focused on comparison of data association methods. Another area for future work is the application of our approach using novel geometric representations, e.g. quadrics [14], [18].

Acknowledgments

This work was partially supported by the Office of Naval Research under grants N00014-18-1-2832 and N00014-16-1-2628. K. Doherty acknowledges support from an NSF Graduate Research Fellowship.

[4] https://youtu.be/9hEonD8KDrs

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] N. Sünderhauf, O. Brock, W. Scheirer, R. Hadsell, D. Fox, J. Leitner, B. Upcroft, P. Abbeel, W. Burgard, M. Milford *et al.*, "The limits and potentials of deep learning for robotics," *The International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 405–420, 2018.

[3] D. Fourie, J. Leonard, and M. Kaess, "A nonparametric belief solution to the Bayes tree," in *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*. IEEE, 2016, pp. 2189–2196.

[4] E. B. Sudderth, A. T. Ihler, M. Isard, W. T. Freeman, and A. S. Willsky, "Nonparametric belief propagation," *Communications of the ACM*, vol. 53, no. 10, pp. 95–103, 2010.

[5] Y. Bar-Shalom and E. Tse, "Tracking in a cluttered environment with probabilistic data association," *Automatica*, vol. 11, no. 5, pp. 451–460, 1975.

[6] D. Reid, "An algorithm for tracking multiple targets," *IEEE transactions on Automatic Control*, vol. 24, no. 6, pp. 843–854, 1979.

[7] I. J. Cox and J. J. Leonard, "Probabilistic data association for dynamic world modeling: A multiple hypothesis approach," in *Advanced Robotics, 1991.'Robots in Unstructured Environments', 91 ICAR., Fifth International Conference on*. IEEE, 1991, pp. 1287–1294.

[8] ——, "Modeling a dynamic environment using a bayesian multiple hypothesis approach," *Artificial Intelligence*, vol. 66, no. 2, pp. 311–344, 1994.

[9] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit, "FastSLAM: A factored solution to the simultaneous localization and mapping problem," in *Proc. of the AAAI National Conference on Artificial Intelligence, 2002*, 2002.

[10] H. Durrant-Whyte, S. Majumder, S. Thrun, M. De Battista, and S. Scheding, "A Bayesian algorithm for simultaneous localisation and map building," in *Robotics Research*. Springer, 2003, pp. 49–60.

[11] N. Sünderhauf and P. Protzel, "Switchable constraints for robust pose graph SLAM," in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*. IEEE, 2012, pp. 1879–1884.

[12] M. Pfingsthorn and A. Birk, "Simultaneous localization and mapping with multimodal probability distributions," *The International Journal of Robotics Research*, vol. 32, no. 2, pp. 143–171, 2013.

[13] E. Olson and P. Agarwal, "Inference on networks of mixtures for robust robot mapping," *The International Journal of Robotics Research*, vol. 32, no. 7, pp. 826–840, 2013.

[14] N. Sünderhauf and M. Milford, "Dual quadrics from object detection bounding boxes as landmark representations in SLAM," *arXiv preprint arXiv:1708.00965*, 2017.

[15] 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," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 1352–1359.

[16] S. Yang and S. Scherer, "CubeSLAM: Monocular 3D object detection and SLAM without prior models," *arXiv preprint arXiv:1806.00557*, 2018.

[17] J. McCormac, R. Clark, M. Bloesch, A. J. Davison, and S. Leutenegger, "Fusion++: Volumetric Object-Level SLAM," *arXiv preprint arXiv:1808.08378*, 2018.

[18] L. Nicholson, M. Milford, and N. Sünderhauf, "QuadricSLAM: Constrained Dual Quadrics from Object Detections as Landmarks in Semantic SLAM," *IEEE Robotics and Automation Letters (RA-L)*, 2018.

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

[20] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics (NRL)*, vol. 2, no. 1-2, pp. 83–97, 1955.

[21] J. Neira and J. D. Tardós, "Data association in stochastic mapping using the joint compatibility test," *IEEE Transactions on robotics and automation*, vol. 17, no. 6, pp. 890–897, 2001.

[22] C. M. Smith and J. J. Leonard, "A multiple-hypothesis approach to concurrent mapping and localization for autonomous underwater vehicles," in *Field and Service Robotics*. Springer, 1998, pp. 237–244.

[23] J. Wang and B. Englot, "Robust exploration with multiple hypothesis data association," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 3537–3544.

[24] S. X. Wang, "Maximum weighted likelihood estimation," Ph.D. dissertation, University of British Columbia, 2001.

[25] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

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

[27] A. Geiger, J. Ziegler, and C. Stiller, "StereoScan: Dense 3d Reconstruction in Real-time," in *IEEE Intelligent Vehicles Symposium*, Baden-Baden, Germany, June 2011.

[28] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama *et al.*, "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7310–7311.

[29] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[30] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[31] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[32] B. Mu, S.-Y. Liu, L. Paull, J. Leonard, and J. P. How, "SLAM with objects using a nonparametric pose graph," in *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*. IEEE, 2016, pp. 4602–4609.

[33] N. Atanasov, M. Zhu, K. Daniilidis, and G. J. Pappas, "Semantic localization via the matrix permanent," in *Robotics: Science and Systems*, vol. 2, 2014.