Data Association Aware Semantic Mapping and Localization via a Viewpoint-Dependent Classifier Model

Vladimir Tchuiev, Yuri Feldman and Vadim Indelman

Abstract— **We present an approach for localization and semantic mapping in ambiguous scenarios by incrementally maintaining a hybrid belief over continuous states and discrete classification and data association variables. Unlike existing incremental approaches we explicitly maintain data association components over time, allowing us to deal with perceptual aliasing. Crucially, we utilize a viewpoint-dependent classifier model over rich classifier outputs and leverage the coupling between poses and semantic measurements both for disambiguating data association and in pose estimation. We demonstrate in simulation that incorporating semantic measurements with a viewpoint-dependent classifier model enhances disambiguation of both data association and localization over usage of only geometric measurements or viewpoint independent models, further contributing to the tractability of the approach in practice, and providing better estimates.**

I. Introduction

Localization and mapping in unknown and uncertain environments is a fundamental capability in robotics, with numerous applications, including search and rescue, autonomous car navigation, indoor navigation, and surveillance. The corresponding problem is known as simultaneous localization and mapping (SLAM) and has been extensively investigated in the last two decades, e.g. see a recent review [1]. Semantic perception and object-based SLAM have been actively investigated by the research community. In particular, object-based SLAM reasons about much fewer landmarks with richer information than geometric SLAM, allowing for faster computation and assistance in data association of features and objects between images.

One of the key challenges in SLAM is a reliable and robust operation in perceptually aliased environments. Data association (DA) is particularly difficult in these scenarios, as the measurement information can be interpreted in multiple ways, and DA errors may lead to critically incorrect estimations. Existing approaches maintain data association hypotheses, which is a computationally difficult problem on its own. Semantic information can assist in disambiguation of DA hypotheses. However, existing approaches that utilize semantic observations for DA disambiguation typically consider only most likely class measurements (e.g. [2]). Yet, different objects can appear visually similar when viewed from certain viewpoints, and thus lead to erroneous most likely class, which will cause these approaches to break. In contrast, we utilize a richer classifier output in the form of class probability vectors. Crucially, existing methods do not exploit the viewpoint dependency embedded in the semantic observations - indeed, the visual appearance of an object (scene) changes when observed from different viewpoints. We propose to exploit this viewpoint-dependency to assist in the data association task within a semantic perception framework, considering highly ambiguous scenarios.

In our approach the robot aims to localize itself and map geometrically and semantically the observed environment while reasoning about ambiguous data association. This kind of inference requires maintaining a hybrid belief and efficiently updating it with incoming information captured online by the robot's sensors. As our main contribution, we utilize a viewpoint dependent classifier model for DA disambiguation by leveraging the coupling between relative viewpoint and classifier outputs. We rigorously incorporate this viewpoint-dependent model within a recursive probabilistic formulation, building upon the DA-BSP framework by Pathak et al. [3], which however, considered only geometric observations. In addition, the proposed approach aids in SLAM, leading to a more accurate inference. Further, while DA-BSP assumes a single scene is observed per time step, we deal with multiple object detections. We demonstrate the strength of utilizing a viewpoint dependent classifier model for DA disambiguation in simulation considering a highly ambiguous environment.

II. Related Work

An important early work on DA is joint probability data association (JPDA) by Fortman et al. [4] which considers all possible DA options, therefore being computationally slow. Our approach utilizes semantic information and weight pruning to reduce the number of DA options considered. Wong et al. [5] presented a Dirichlet Process Mixture Model (DPMM) for data association for a partially observed environment. Sunderhauf and Protzel [6] proposed an approach to detect faulty loop closures that lead to erroneous data association in back-end optimization. Olson and Agarwal [7] proposed a robust approach that uses max-mixture models. Carlone et al. [8] classified measurements as coherent or not, thus predicting if they will result in erroneous data association. Indelman et al. [9] proposed a multi-robot framework for

V. Tchuiev and V. Indelman are with the Department of Aerospace Engineering, Technion - Israel Institute of Technology, Haifa 32000, Israel. Y. Feldman is with the Department of Computer Science, Technion - Israel Institute of Technology, Haifa 32000, Israel. {vovatch, vadim.indelman}@technion.ac.il, yurif@cs.technion.ac.il. This work was partially supported by the Israel Ministry of Science & Technology (MOST).

SLAM with ambiguous data association. A recent work by Pathak et al. [3] for data association aware belief space planning (DA-BSP) targets perceptual aliasing by explicitly reasoning about and probabilistically maintaining ambiguous DA hypotheses, in both inference and belief space planning. Yet, all of the above works are confined to geometric measurements.

Milan et al. [10] presented a method based on LSTM neural network forg data association, training it on the MOTChallenge dataset. Farazi and Behnke [11] expended on the above work to visually track and associate between identical robots using an LSTM based approach. [10], [11] are both deep learning based approaches, where a key question is how far the deployment scenario is from the training set, i.e. model uncertainty.

Several notable works utilized classifier models to improve class inference. Omidshafiei et al. [12] proposed a sequential classification algorithm that utilizes a classifier model that models the classifier output as a Dirichlet distribution. This model makes the algorithm robust to classification ambiguity, but is independent of the relative viewpoint between camera and object. Tchuiev and Indelman [13] maintained a distribution over posterior class probability, and in particular, provided access to posterior classification uncertainty, by incorporating model uncertainty within a sequential classification setting. A Dirichlet distributed classifier model was used as well, also independent on relative viewpoint. Teacy et al. [14], and Feldman and Indelman [15] utilized a Gaussian process viewpoint dependent classifier model to assist in classification tasks. Kopitkov and Indelman [16] utilized a viewpoint dependent classifier model, learned offline via deep learning, for probabilistic inference over robot trajectory. In contrast to the above works, which either consider a single object/scene or assume data association to be given and perfect, we utilize the viewpoint dependent classifier model to assist in data association disambiguation while addressing a localization and semantic mapping problem. Addressing this problem involves inference over a hybrid belief over continuous and discrete variables.

The following are the most relevant works that present approaches for hybrid belief inference. Segal and Reid [17] proposed a message passing algorithm to optimize hybrid factor graphs for inference. The discrete-continuous graphical model (DC-GM) approach by Lajoie et al. [18] performed inference on a hybrid factor graph that produces near-optimal estimates. Mu et al. [2] proposed a sampling based approach that uses most likely class semantic measurements; this approach performs batch inference using expectation maximization (EM). Bowman et al. [19] utilizes most likely class and bounding box measurements, in addition to geometric measurements, to perform SLAM and DA disambiguation using EM as well. The above approaches consider only the most likely class and do not reason about viewpoint dependency of classification results. In contrast, we utilize richer classifier output in conjunction with a viewpoint dependent model to perform object level SLAM, while maintaining classification and DA hypotheses.

III. Notations and Problem Formulation

Consider a robot operating in a partially known environment containing different, possibly perceptually similar or identical, objects. The robot aims to localize itself, and map the environment geometrically and semantically while reasoning about ambiguous data association (DA). We consider a closed-set setting where each object is assumed to be one of M classes. Moreover, in this work we consider the number of objects in the environment is known. The objects are assumed to be stationary.

Let x_k denote the robot's camera pose at time k, and x_n^o and c_n represent the n-th object pose and class, respectively. We denote the set of all object poses and classes by $\mathcal{X}^o \doteq \{x_1^o, ..., x_N^o\}$ and $C \doteq \{c_1, ..., c_n\}$. To shorten notations, denote $\mathcal{X}_k \doteq \{x_{0:k}, \mathcal{X}^o\}$.

Further, we denote the data association realization at time k as β_k: given n_k object observations at time k, $\beta_k \in \mathbb{R}^{n_k}$; each element in β_k corresponds to an object observation, and is equal to an object's identity label. For example, if at time k the camera observes 2 objects with hypothesized identity labels 4 in observation 1 and 6 in observation 2, then $\beta_k = [\beta_{k,1}, \beta_{k,2}]^T \in \mathbb{R}^2$, and $\beta_{k,1} = 4$ and $\beta_{k,2} = 6$. Denote $\mathcal{Z}_k \doteq \{z_{k,1}, ..., z_{k,n_k}\}$ as the set of n_k measurements at time k, and a_k as the robot's action at time k.

Each measurement $z_{k,i} \in \mathcal{Z}_k$ consists of two parts: a geometric part $z_{k,i}^{geo}$, e.g. range or bearing measurements to an object, and a semantic part $z_{k,i}^{sem}$. The set of all geometric measurements for time k is denoted $\mathcal{Z}_k^{geo}$, and similarly for semantic measurements $\mathcal{Z}_k^{sem}$, such that $\mathcal{Z}_k = \mathcal{Z}_k^{geo} \cup \mathcal{Z}_k^{sem}$. We assume the geometric and semantic measurements are independent from each other. In addition, we assume independence between measurements at different time steps.

We consider standard motion and geometric observation Gaussian models, such that $\mathbb{P}(x_{k+1}|x_k, a_k) = \mathcal{N}(f(x_k, a_k), \Sigma_w)$ and $\mathbb{P}(z_k^{geo}|x_k, x^o) = \mathcal{N}(h^{geo}(x_k, x^o), \Sigma_v^{geo})$. The process and geometric measurement covariance matrices, Σ_w and Σ_v^{geo}, as well as the functions $f(.), h^{geo}(.)$ are assumed to be known.

For the semantic measurements, we utilize a (deep learning) classifier that provides a vector of class probabilities where $z_{k,i}^{sem} \doteq \mathbb{P}(c_i|I_{k,i})$ given sensor raw observation $I_{k,i}$, e.g. an image cropped from a bounding box of a larger image taken by the camera of object i at time k. To simplify notations we drop index i, as the measurements, both semantic and geometric, apply to each bounding box. Thus, $z_k^{sem} \in \mathbb{R}^M$ with

$$z_k^{sem} \doteq [\mathbb{P}(c = 1|I_k) \quad \cdots \quad \mathbb{P}(c = M|I_k)]^T . \qquad (1)$$

A crucial observation, following [15], is that z_k^{sem} is dependent on the camera's pose relative to the object (see Fig. 1). In this work we contribute an approach that leverages this coupling to assist in inference and data association disambiguation.

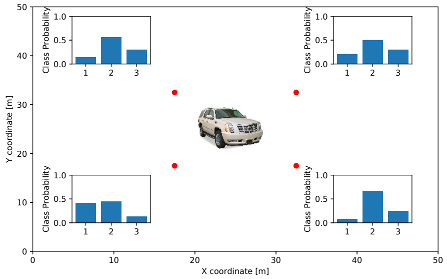


Fig. 1: A classifier observing an object from multiple viewpoints will produce different classification scores for each viewpoint.

Specifically, we model this dependency via a classifier model $\mathbb{P}(z_k^{sem}|c = m, x_k, x^o)$. The classifier model represents the distribution over classifier output, i.e. class probability vector z_k^{sem}, when an object with a class hypothesis m is observed from relative pose $x^o \ominus x_k$. Note that for M classes we require M classifier models, one for each class. The model can be represented with a Gaussian Process (see [14], [15]) or a deep neural network (see [16]). In this work, we use a Gaussian classifier model, given by

$$\mathbb{P}(z_k^{sem} \mid c, x_k, x^o) = \mathcal{N}(h_c(x_k, x^o), \Sigma_c(x_k, x^o)), \quad (2)$$

where the viewpoint-dependent functions $h_c(x_k, x^o)$ and $\Sigma_c(x_k, x^o)$ are learned offline. Note that unlike [14], [15] we do not model correlations in classifier scores among viewpoints. Conversely, we do not assume data association is known.

We assume a prior on initial camera and object poses, x_0 and X^o respectively, and class realization probability $\mathbb{P}(C)$. For simplicity, we assume independent variable priors (although this assumption is not required by our approach, and is not true in general, as e.g. some objects are more likely to appear together than others), thus we can write the prior as follows:

$$\mathbb{P}(x_0, X^o, C) = \mathbb{P}(x_0) \prod_{i=1}^{N} \mathbb{P}(x_i^o)\mathbb{P}(c_i). \quad (3)$$

In this paper we use a Gaussian prior for the continuous variables, and uninformative (uniform) prior for the object classes.

Problem formulation: We aim to efficiently maintain the following hybrid belief

$$\mathbb{P}(\mathcal{X}_k, C, \beta_{1:k} \mid \mathcal{H}_k), \quad (4)$$

with history $\mathcal{H}_k \doteq \{\mathcal{Z}_{1:k}, a_{0:k-1}\}$. The belief (4) is both over continuous variables, i.e. robot and object poses $\mathcal{X}_k$ (continuous variables), and over discrete variables, i.e. object classes C and data association hypotheses thus far, $\beta_{1:k}$. In the following, we incorporate a viewpoint-dependent classifier model and develop a recursive formulation to update that hybrid belief with incoming information captured by the robot as it moves in the environment.

IV. Approach

In this section we develop a recursive scheme to compute and maintain the hybrid belief from Eq. (4). We start by factorizing using the chain rule as

$$\mathbb{P}(\mathcal{X}_k, C, \beta_{1:k}|\mathcal{H}_k) = \underbrace{\mathbb{P}(\mathcal{X}_k|C, \beta_{1:k}, \mathcal{H}_k)}_{b[\mathcal{X}_k]^C_{\beta_{1:k}}} \underbrace{\mathbb{P}(C, \beta_{1:k}|\mathcal{H}_k)}_{w^C_{\beta_{1:k}}}, \quad (5)$$

where $b[\mathcal{X}_k]^C_{\beta_{1:k}} \doteq \mathbb{P}(\mathcal{X}_k \mid C, \beta_{1:k}, \mathcal{H}_k)$ is the conditional belief over the continuous variables, and $w^C_{\beta_{1:k}} \doteq \mathbb{P}(C, \beta_{1:k} \mid \mathcal{H}_k)$ is the marginal belief over the discrete variables, and can be considered as the conditional belief weight. Thus, each realization of the discrete variables, i.e. data association and class hypotheses, has its own probability (weight) and gives rise to a different belief over the continuous variables.

Moreover, the factorization (5) facilitates computation of marginal distributions that are of interest in practice. In particular, the posterior over robot and object poses can be calculated via

$$\mathbb{P}(\mathcal{X}_k \mid \mathcal{H}_k) = \sum_{\beta_{1:k}} \sum_{C} w^C_{\beta_{1:k}} b[\mathcal{X}_k]^C_{\beta_{1:k}}, \quad (6)$$

while the marginal distributions over object classes and data association hypotheses are given by

$$\mathbb{P}(C \mid \mathcal{H}_k) = \sum_{\beta_{1:k}} w^C_{\beta_{1:k}}, \quad (7)$$

$$\mathbb{P}(\beta_{1:k} \mid \mathcal{H}_k) = \sum_{C} w^C_{\beta_{1:k}}. \quad (8)$$

The posterior $\mathbb{P}(\mathcal{X}_k \mid \mathcal{H}_k)$ in Eq. (6) is a mixture belief that accounts for all hypotheses regarding data association and classification. Without semantic observations, our approach degenerates to passive DA-BSP. The term $\mathbb{P}(C \mid \mathcal{H}_k)$ is the distribution over classes of all objects while accounting for both localization uncertainty and ambiguous data association. As such, it is important for robust semantic perception. Finally, the posterior over data association hypotheses, $\mathbb{P}(\beta_{1:k} \mid \mathcal{H}_k)$ accounts for all class realizations for all objects.

Next, we derive a recursive formulation for calculating the continuous and marginal distributions in the factorization (5). As will be seen, semantic observations along with the viewpoint-dependent classifier model (2) impact both of the terms in the factorization (5), and as a result assist in inference of robot and objects poses (via Eq. (6)) and helps in disambiguation between data association realizations (via Eq. (8)). Furthermore, as discussed in Sec. 4.D, while the number of objects' classes and data association hypotheses (number of weights $w^C_{\beta_{1:k}}$) is intractable, in practice many of these are negligible and can be pruned.

A. *Conditional Belief Over Continuous Variables:* $b[\mathcal{X}_k]^C_{\beta_{1:k}}$

Using Bayes law we get the following expression:

$$\begin{aligned} b[\mathcal{X}_k]^C_{\beta_{1:k}} &\equiv \mathbb{P}(\mathcal{X}_k \mid C, \beta_{1:k}, \mathcal{H}_k) \propto \\ \mathbb{P}(\mathcal{Z}_k \mid \mathcal{X}_k, C, \beta_k) &\cdot \mathbb{P}(\mathcal{X}_k \mid C, \beta_{1:k-1}, \mathcal{H}_k^-), \end{aligned} \quad (9)$$

where $\mathcal{H}_k^- \doteq \{\mathcal{Z}_{1:k-1}, a_{0:k-1}\}$, the normalization constant is omitted as it does not depend on $\mathcal{X}_k$, and β_k is dropped in the second term because it refers to association of $\mathcal{Z}_k$ which is not present.

The expression $\mathbb{P}(\mathcal{Z}_k \mid \mathcal{X}_k, C, \beta_k)$ in Eq. (9) is the joint measurement likelihood for all geometric and semantic observations obtained at time k. Given classifications, associations and robot pose at time k, history $\mathcal{H}_k^-$ and past associations $\beta_{1:k-1}$ can be omitted. The joint measurement likelihood can be explicitly written as

$$\begin{gathered}\mathbb{P}(\mathcal{Z}_k \mid \mathcal{X}_k, C, \beta_k) = \\ \prod_{i=1}^{n_k} \mathbb{P}(z_{k,i}^{geo} \mid x_k, x^o_{\beta_{k,i}}) \cdot \mathbb{P}(z_{k,i}^{sem} \mid x_k, x^o_{\beta_{k,i}}, c_{\beta_{k,i}}),\end{gathered} \tag{10}$$

where $x^o_{\beta_{k,i}}$ and $c_{\beta_{k,i}}$ are the object pose and class corresponding to the measurement respectively, given DA realization β_k and n_k the number of measurements obtained at time k as before. Note that the viewpoint-dependent semantic measurement term above $\mathbb{P}(z_{k,i}^{sem} \mid x_k, x^o_{\beta_{k,i}}, c_{\beta_{k,i}})$ couples between semantic measurement and robot pose relative to object, making it useful for inference of both.

The term $b^-[\mathcal{X}_k]^C_{\beta_{1:k-1}} \doteq \mathbb{P}(\mathcal{X}_k \mid C, \beta_{1:k-1}, \mathcal{H}_k^-)$ in Eq. (9) is the propagated belief over continuous variables, which, using chain rule, can be written as

$$b^-[\mathcal{X}_k]^C_{\beta_{1:k-1}} = \mathbb{P}(x_k|x_{k-1}, a_{k-1}) b[\mathcal{X}_{k-1}]^C_{\beta_{1:k-1}}. \tag{11}$$

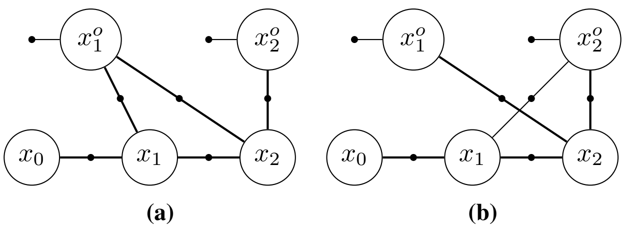
Overall, the conditional belief (9) can be represented as a factor graph (Kschischang et al. [20]). Note that each realization of $\beta_{1:k}$ has a different factor graph topology (observation factors are affected, motion model factors are not). For a fixed $\beta_{1:k}$ with different class assignments C the corresponding conditional belief factor graph topology remains the same (geometric and semantic observation factors connect the same nodes), but semantic factors change, according to class models.

Fig. 2 presents an example for 2 factor graphs, in which $k = 2$, $N = 2$, and the DA hypothesis is that at time $k = 1$ the camera observes object 1 for the first graph and 2 for the second, at time $k = 2$ the camera observes objects 1 and 2 for both graphs. To efficiently infer $\mathcal{X}_k$ for every realization of C and $\beta_{1:k}$, state of the art incremental inference approaches, such as iSAM2 [21] can be used. The joint posterior from Eq. (4) can thus be maintained following Eq. (5) as a set of continuous beliefs $b[\mathcal{X}_k]^C_{\beta_{1:k}}$ conditioned on the discrete variables $\beta_{1:k}$ and C each represented with a factor graph, along with their corresponding component weights $w^C_{\beta_{1:k}}$, describing the marginal belief over discrete variables. In the next section, we describe how the latter can be calculated.

B. Marginal Belief Over Discrete Variables: $w^C_{\beta_{1:k}}$

To compute the DA and class realization weight $w^C_{\beta_{1:k}}$ we marginalize over all continuous variables:

$$w^C_{\beta_{1:k}} \equiv \mathbb{P}(C, \beta_{1:k} \mid \mathcal{H}_k) = \int_{\mathcal{X}_k} \mathbb{P}(\mathcal{X}_k, C, \beta_{1:k}|\mathcal{H}_k) d\mathcal{X}_k. \tag{12}$$



(a) **(b)**

Fig. 2: A toy example for two factor graphs in our approach, each for a different data association realization. The edges that connect between camera poses correspond to motion model $\mathbb{P}(x_k|x_{k-1}, a_{k-1})$. The edges that connect directly between camera and object poses correspond to the measurement model (10) for both semantic and geometric measurements. Thus a viewpoint-dependent semantic measurement model results in geometric constraints on robot-to-object relative pose.

Using Bayes law, we can expand the above as follows:

$$\begin{gathered}\mathbb{P}(\mathcal{X}_k, C, \beta_{1:k} \mid \mathcal{H}_k) = \\ \eta \cdot \mathbb{P}(\mathcal{Z}_k \mid \mathcal{X}_k, C, \beta_{1:k}) \cdot \mathbb{P}(\mathcal{X}_k, C, \beta_{1:k}|\mathcal{H}_k^-)\end{gathered} \tag{13}$$

where $\eta = \mathbb{P}(\mathcal{Z}_k \mid \mathcal{H}_k^-)^{-1}$ is a normalization constant and the joint measurement likelihood $\mathbb{P}(\mathcal{Z}_k \mid \mathcal{X}_k, C, \beta_{1:k})$ can be explicitly written as in Eq. (10).

We further expand $\mathbb{P}(\mathcal{X}_k, C, \beta_{1:k}|\mathcal{H}_k^-)$ using chain rule:

$$\begin{gathered}\mathbb{P}(\mathcal{X}_k, C, \beta_{1:k}|\mathcal{H}_k^-) = \mathbb{P}(\beta_k|\beta_{1:k-1}, \mathcal{X}_k, C, \mathcal{H}_k^-) \cdot \\ \cdot \mathbb{P}(x_k|x_{k-1}, a_{k-1}) \cdot \mathbb{P}(\mathcal{X}_{k-1}, C, \beta_{1:k-1} \mid \mathcal{H}_{k-1}),\end{gathered} \tag{14}$$

where $\mathbb{P}(\mathcal{X}_{k-1}, C, \beta_{1:k-1}|\mathcal{H}_{k-1}) = \sum w^C_{\beta_{1:k-1}} b[\mathcal{X}_{k-1}]^C_{\beta_{1:k-1}}$ is the prior belief calculated at time $k-1$ and represented as a set of continuous belief components along with corresponding weights as described above. The term $\mathbb{P}(\beta_k|\beta_{1:k-1}, \mathcal{X}_k, C, H_k^-)$ from Eq. (14) is the object observation model that represents the probability of observing a scene given a hypothesis of camera and object poses. In this paper we use a simple model that depends only on camera and object poses at current time step, thus it can be written as $\mathbb{P}(\beta_k \mid x_k, \mathcal{X}^o_{\beta_k})$, where $\mathcal{X}^o_{\beta_k} \doteq \{x^o_{\beta_{k,i}}\}_{i=1}^{n_k}$. If the model predicts observation of all objects corresponding to β_k then $\mathbb{P}(\beta_k \mid x_k, \mathcal{X}^o_{\beta_k})$ is equal to a constant, otherwise it is zero.

Plugging the above into Eq. (12) yields a recursive rule for calculating component weights at time k

$$\begin{gathered}w^C_{\beta_{1:k}} = \eta \cdot \int_{\mathcal{X}_k} \mathbb{P}(\mathcal{Z}_k|\mathcal{X}_k, C, \beta_k) \cdot \mathbb{P}(\beta_k|x_k, \mathcal{X}^o_{\beta_k}) \cdot \\ \cdot b^-[\mathcal{X}_k]^C_{\beta_{1:k-1}} w^C_{1:k-1} \, d\mathcal{X}_k.\end{gathered} \tag{15}$$

The normalization constant η (from Eq. (13)) does not depend on variables and cancels out when weights are normalized to sum to 1. It is therefore dropped out in subsequent calculations. Note that the realization weight from the previous time step $w^C_{\beta_{1:k-1}}$ is independent from $\mathcal{X}_k$, and thus can be taken out of the integral. Recalling Eq. (10), the continuous variables participating in $\mathbb{P}(\mathcal{Z}_k \mid \mathcal{X}_k, C, \beta_{1:k})$ are x_k and $\mathcal{X}^o_{\beta_k}$. Those variables are participating also in $\mathbb{P}(\beta_k \mid x_k, x^o_k)$. As $b^-[\mathcal{X}_k]^C_{\beta_{1:k-1}}$ is Gaussian, all other continuous variables can be marginalized easily. On the other hand, x_k and $\mathcal{X}^o_{\beta_k}$ must be sampled because of the object observation model $\mathbb{P}(\beta_k \mid x_k, \mathcal{X}^o_k)$, which is commonly not Gaussian. If the observation model predicts that the objects will not be observed for most of the samples, then $w^C_{\beta_{1:k}}$

will be small and likely to be pruned. We can express the realization weight as follows:

$$w^C_{\beta_{1:k}} \propto w^C_{\beta_{1:k-1}} \iint_{x_k, \mathcal{X}^o_{\beta_k}} \mathbb{P}(\mathcal{Z}_k \mid x_k, \mathcal{X}^o_{\beta_k}, C, \beta_{1:k}) \cdot \\ \cdot \, \mathbb{P}(\beta_k \mid x_k, \mathcal{X}^o_{\beta_k}) \; b^-[x_k, \mathcal{X}^o_{\beta_k}]^C_{\beta_{1:k-1}} \; dx_k \; d\mathcal{X}^o_{\beta_k}, \tag{16}$$

where:

$$b^-[x_k, \mathcal{X}^o_{\beta_k}]^C_{\beta_{1:k-1}} \doteq \mathbb{P}(x_k, \mathcal{X}^o_{\beta_k} | C, \beta_{1:k-1}, \mathcal{H}^-_k) = \\ \int_{\mathcal{X}^-_k \setminus \{x_k, \mathcal{X}^o_{\beta_k}\}} b^-[\mathcal{X}_k]^C_{\beta_{1:k-1}} \; d\left\{\mathcal{X}^-_k \setminus \{x_k, \mathcal{X}^o_{\beta_k}\}\right\}. \tag{17}$$

The viewpoint dependent classifier model contributes to data association disambiguation by acting as reinforcement or contradiction to the geometric model. If both 'agree' on the poses' hypothesis, $w^C_{\beta_{1:k}}$ will be large relative to cases where both 'disagree'.

Next, we provide an overview of the inference scheme, then address computational aspects.

C. Overall Algorithm

The proposed scheme is outlined in Alg. 1. For every time step we are input the prior belief $\mathbb{P}(\mathcal{X}_{k-1}, C, \beta_{1:k-1} \mid \mathcal{H}_{k-1})$ represented following Eq. (5) as a set of weights $w^C_{\beta_{1:k-1}} \doteq \mathbb{P}(C, \beta_{1:k-1} \mid \mathcal{H}_{k-1})$ and corresponding continuous (Gaussian) belief components $b[\mathcal{X}_{k-1}]^C_{\beta_{1:k-1}} \doteq \mathbb{P}(\mathcal{X}_{k-1} \mid C, \beta_{1:k-1}, \mathcal{H}_{k-1})$. In our implementation we maintain a separate factor graph for each such component. We also obtain an action a_{k-1} and observations $\mathcal{Z}_k$, separated into geometric $\mathcal{Z}^{geo}_k$, and semantic $\mathcal{Z}^{sem}_k$. We propagate each prior belief component using the motion model $\mathbb{P}(x_k \mid x_{k-1}, a_{k-1})$ (step 3). Each component then splits to a number of subcomponents, one for each possible assignments of data associations β_k at current time (generally a vector of length n_k). Procedure `PropWeights` at step 5 computes the normalized weight of each subcomponent via Eq. (16) as a product of the component (prior) weight $w^C_{\beta_{1:k-1}}$ with an update term comprising the measurement likelihood $\mathbb{P}(\mathcal{Z}_k \mid x_k, \mathcal{X}^o_{\beta_k}, C, \beta_{1:k})$ (both geometric and semantic, see Eq. (10)) and object observation model $\mathbb{P}(\beta_k \mid x_k, \mathcal{X}^o_{\beta_k})$, averaged over the propagated belief $b^-[x_k, \mathcal{X}^o_{\beta_k}]^C_{\beta_{1:k-1}}$ from Eq. (17). In step 7 we prune low-weight subcomponents by setting their weights to 0 and re-normalizing remaining weights to 1, in an approximation to true posterior (other pruning strategies are equally possible). In step 11 we update the posterior for non-zero weight subcomponents using current measurements. Finally, we return posterior as a set of Gaussian components and corresponding weights.

We next address aspects of computational tractability of the scheme.

D. Computational Complexity and Tractability

With M candidate classes, and N objects, the number of possible class realizations, and consequently initial number of belief components, is M^N. At time step k each prior component splits into up to N^{n_k} subcomponents as each measurement can in general be associated to any scene object. The maximum number of components at time k is thus $M^N \cdot \prod_{j=1}^{k} N^{n_j} = O\left(M^N \cdot N^{\psi \cdot k}\right)$ if ψ is an upper bound on n_k, making the approach computationally intractable in theory without pruning. In practice, as observed by [3], the number of components that need to be accounted for is limited by the belief, and is much smaller than the theoretical maximum, with the rest getting negligible weights that can be safely pruned under any scheme. Further, our empirical results suggest that semantic information added through the viewpoint-dependent factors leads to even stronger disambiguation than observed in DA-BSP (which uses only geometric information), both in data association and localization, resulting in smaller number of non-negligible weights.

Algorithm 1 Data Association-Aware Mapping and Localization. Inference at time k

```
Input: Prior belief $\mathbb{P}(\mathcal{X}_{k-1}, C, \beta_{1:k-1} \mid \mathcal{H}_{k-1})$,
    observations $\mathcal{Z}_k = (\mathcal{Z}^{geo}, \mathcal{Z}^{sem})$, action $a_{k-1}$
 1: for every component $\beta_{1:k-1}, C$ s.t. $w^C_{\beta_{1:k-1}} > 0$ do
 2:     ▷ Propagate component according to motion model
 3:     $b[\mathcal{X}^-_k]^C_{\beta_{1:k-1}} \leftarrow \mathbb{P}(x_k|x_{k-1}, a_{k-1}) \cdot b[\mathcal{X}_{k-1}]^C_{\beta_{1:k-1}}$
 4:     ▷ Propagate weights Eq. (15), Eq. (16)
 5:     $w^C_{\beta_{1:k}} \leftarrow$ PropW. $\left(b[\mathcal{X}^-_k]^C_{\beta_{1:k-1}}, w^C_{\beta_{1:k-1}}, \mathcal{Z}_k\right)$
 6:     ▷ Prune low-probability components
 7:     $w^C_{\beta_{1:k}} \leftarrow$ PruneAndNormalize$(w^C_{\beta_{1:k}})$
 8:     ▷ Propagate non-zero weight components
 9:     for $\beta_{1:k}, C$ s.t. $w^C_{\beta_{1:k}} > 0$ do
10:         ▷ Add observation factors, Eqs. (9), and (10)
11:         $b[\mathcal{X}_k]^C_{\beta_{1:k}} \leftarrow b[\mathcal{X}^-_k]^C_{\beta_{1:k-1}} \cdot \mathbb{P}(\mathcal{Z}_k \mid \mathcal{X}_k, C, \beta_k)$
12:     end for
13: end for
14: return $\mathbb{P}(\mathcal{X}_k, C, \beta_{1:k} \mid \mathcal{H}_k) \equiv \{(b[\mathcal{X}_k]^C_{\beta_{1:k}}, w^C_{\beta_{1:k}})\}$

 1: procedure PropWeights$(b[\mathcal{X}^-_k]^C_{\beta_{1:k-1}}, w^C_{\beta_{1:k-1}}, \mathcal{Z}_k)$
 2:     for every possible assignment of $\beta_k$ do
 3:         ▷ Sample current poses by Eq. (17)
 4:         Sample $\{x^{(i)}_k, \mathcal{X}^{o\ (i)}_{\beta_k}\}^{n_s}_{i=1} \sim b^-[x_k, \mathcal{X}^o_{\beta_k}]^C_{\beta_{1:k-1}}$
 5:         ▷ Calculate update factor and propagate Eq. (16)
 6:         $\psi \leftarrow (1/n_s) \cdot \sum_{i=1}^{n_s} \mathbb{P}(\mathcal{Z}_k, \beta_k \,|\, x^{(i)}_k, \mathcal{X}^{o\,(i)}_{\beta_k}, C, \beta_{1:k-1})$
 7:         $\widetilde{w}^C_{\beta_{1:k}} \leftarrow w^C_{\beta_{1:k-1}} \cdot \psi$
 8:     end for
 9:     ▷ Normalize weights and return
         return $w^C_{\beta_{1:k}} \leftarrow \widetilde{w}^C_{\beta_{1:k}} / \sum_{\beta_k} \widetilde{w}^C_{\beta_{1:k}}$
10: end procedure

 1: procedure PruneAndNormalize$(w^C_{\beta_{1:k}})$
 2:     for $\beta_{1:k}, C$ s.t. $w^C_{\beta_{1:k}} <$ THRESHOLD do
 3:         $\widetilde{w}^C_{\beta_{1:k}} \leftarrow 0$
 4:     end for
         return $w^C_{\beta_{1:k}} \leftarrow \widetilde{w}^C_{\beta_{1:k}} / \sum_{\beta_k} \widetilde{w}^C_{\beta_{1:k}}$
 5: end procedure
```

Additionally, we hypothesize that classification uncertainty is in practice usually limited to only a few classes, and thus would not cause a computational bottleneck even with numerous candidate classes. One can further avoid explicitly maintaining the initially exponential number of components (M^N) by noting that the classes of objects that were not observed yet under an association hypothesis $\beta_{1:k}$ do not participate in the inference process for that belief component, and thus do not need to be maintained separately. That is, for two class realizations C and C', if $C_{\beta_{1:k}} = C'_{\beta_{1:k}}$ with $C_{\beta_{1:k}} \doteq \{\forall 1 \leq j \leq k, i \ c_{\beta_{j,i}}\}$ (i.e. classifications for all associated objects are the same) and $C_{\neg\beta_{1:k}} \neq C'_{\neg\beta_{1:k}}$ (i.e. realizations differ on classifications for objects that do not participate in $\beta_{1:k}$), then $w^{C'}_{\beta_{1:k}} = w^{C}_{\beta_{1:k}}$ (assuming uninformative prior on classes) and $b[\mathcal{X}_k]^{C}_{\beta_{1:k}} = b[\mathcal{X}_k]^{C'}_{\beta_{1:k}}$ (always), without need to compute or maintain those separately.

Finally we note that parts of Alg. 1 can be readily parallelized (”embarrasingly parallel”), thanks to computations being independent across components and wide availability of massively parallel processors (e.g. GPUs), contributing to its practical applicability.

V. Experiments

In this section we evaluate the performance of our approach in a 2D simulation and demonstrate the advantage of using a viewpoint dependent classifier model for disambiguating between DA realizations and improving inference accuracy. Our implementation uses the GTSAM library [22] with a Python wrapper; all experiments were run on an Intel i7-7700 CPU running at 2800 GHz and with 16GB RAM.

We consider a scenario where the robot navigates in an uncertain perceptually aliased environment represented by a set of scattered objects of the same class, i.e. objects differ in their position and orientation. In this scenario $M = 2$ and $N = 6$, thus the number of possible class realizations is $M^N = 64$. Fig. 3a shows the ground truth object poses and robot trajectory.

The prior (3) comprises a highly uncertain initial robot pose, and an uninformative prior on object classes. Object poses are assumed to be known up to a certain accuracy (i.e. uncertain map). The prior covariance off the objects is $\Sigma_o = diag(0.05, 0.05, 0.5 \cdot 10^{-3})$, and initial robot pose is $\Sigma_p = diag(100, 100, 0.04)$. The process and geometric measurement covariance matrices are $\Sigma_w = diag(0.75 \cdot 10^{-3}, 0.75 \cdot 10^{-3}, 0.25 \cdot 10^{-3})$ (corresponds to spatial coordinates and orientation), and $\Sigma_v^{geo} = diag(0.1, 0.05)$ (corresponds to range and bearing).

The semantic measurement model (2) is defined as:

$$h_c(c = 1, \theta) = \begin{bmatrix} \alpha \sin^2(\theta/2) + (1 - \alpha) \\ \alpha - \alpha \sin^2(\theta/2) \end{bmatrix} \tag{18}$$

where θ is the relative angle from the object to camera, calculated from the relative pose $x_k^{rel} \doteq x_k \ominus x^o$. This chosen model represents a mirror symmetrical object (e.g. a car) with a parameter α that corresponds to the viewpoint dependency 'strength', i.e. $\frac{\partial h_c}{\partial \theta}$ values are larger when α increases (for computation details, see [16]). We assume the classifier scores are independent from range from camera to object as the observations are cropped from bounding boxes, and unless the camera is very close to the object the perspective distortion is negligible. In practice, the classifier model can be learned from images of an object from different viewpoints with corresponding classifier outputs via a neural network or GP for example. The measurement covariance matrix $\Sigma_c \doteq (R^T R)^{-1}$ is defined as $R = K \begin{bmatrix} 1 & -0.5 \\ 0 & 1 \end{bmatrix}$. We note that in general, also Σ_c can be viewpoint-dependent [15], [16]. The parameters α and K are constants and take the values $\alpha = 0.25$ and $K = 15$ by default. We sample measurements from our motion, geometric, and semantic models.

Further, we sample 1000 sets of x_k and $x^o_{\beta_k}$ for each computation of $w^C_{\beta_{1:k}}$, see procedure PropWeights in Alg. 1, and compute them as shown in Eq. (16). At each time k we prune components with weight w below threshold $\frac{\{w_k\}_{max}}{150} \leq w$, where $\{w_k\}_{max}$ is the highest weight component at time k.

We compare performance of our approach that utilizes semantic observations along with a viewpoint-dependent classifier model against an alternative that does not use this information, with the latter roughly corresponding to the passive instance of DA-BSP [3]. To quantify performance as a function of α we compare between the following metrics:

1) Entropy over data association weights: for N_k non-pruned weights $\{w_i\}_{i=1}^{N_k}$ we compute the entropy $H(w)$ with $H(w) \doteq -\sum_{i=1}^{N_k} w_i \log(w_i)$.
2) Determinant of position covariance $det(\Sigma)$ of x_k for the highest weight realization at each time k.
3) Estimation error $\tilde{x}^{w_{max}}$, which is the Euclidean distance from ground truth to highest weight estimation for the last pose.
4) Estimation error $\tilde{x}^{w-avg}$, which is the weighted average of all estimation errors for the last pose.

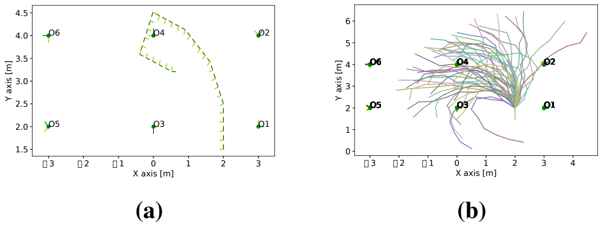


Fig. 3: (a) An example scenario with ground truth camera trajectory, represented in terms of camera poses (green line is the camera heading) and objects $O1$ to $O6$ (green dots indicate position, orientation is not shown). **(b)** Multiple sampled paths for the statistical study, each path realization is presented with different color.

Fig. 4 shows results of an example scenario for different time steps, comparing between using the viewpoint-dependent classifier model (middle row) and without semantic information (upper row), essentially utilizing passive DA-BSP [3]. At each time k, the plots show the mixture posterior $\mathbb{P}(x_k|\mathcal{H}_k)$ over camera pose x_k, calculated from (6), where each component is a Gaussian, thus represented by mean and covariance. Estimated camera poses are shown in red and blue lines, where the blue line represents the camera orientation. Components with higher weight are shown with thicker covariance ellipse lines. To reduce clutter, the posterior over the rest of the continuous variables, i.e. object poses and past robot poses, is not shown. Additionally, the plots show the ground truth trajectory (from Fig. 3a) of the robot. The bottom row reports the probabilities of DA hypotheses from (8) for different time instances for both compared cases. The correct association is marked with a green circle.

As seen from the upper row of Fig. 4, inference without incorporating viewpoint-dependent semantic information results in the first time steps in multiple DA realizations with similar weights. The reason is that given only geometric range and bearing measurements without observing all the objects, inference results can be interpreted in multiple ways, i.e. perceptually aliased (see Fig. 4i). Only at time $k = 25$ the DA was disambiguated once the camera observed objects $O1$ and $O2$.

In contrast, utilizing a viewpoint-dependent classifier model admits faster DA disambiguation, as shown in the bottom row of Fig. 4. In particular, already at time $k = 1$ the posterior $\mathbb{P}(x_k|\mathcal{H}_k)$ has only two non-negligible components, while at time $k = 5$ there is a single DA realization with significant weight. This shows an improvement over Fig. 4b where there are multiple DA realizations with significant weight when not using the classifier model.

The bottom row in Fig. 4 presents the realization weights for the times $k = 1, 5, 15, 25$, and compare between weights without and with classifier model. For each realization $\beta_{1:k}$, we present $\mathbb{P}(\beta_{1:k} \mid \mathcal{H}_k)$ after pruning without classifier model as a blue bar, and with as a red bar. If the bar is missing, then $\mathbb{P}(\beta_{1:k} \mid \mathcal{H}_k) = 0$. In all sub-figures the classifier model reduces the number of non pruned DA realizations, and for time $k = 15$ and $k = 25$ the DA is disambiguated with the classifier model. We observe more DA realizations when the classifier model is not used, and at time $k = 15$ the DA with a classifier model fully disambiguated.

Further, we quantify the performance improvement due to the viewpoint-dependent classifier model in a statistical study by sampling multiple ground truth tracks in the scenario, while keeping the same landmarks. The sampled tracks are shown in Fig. 3b. For this study, we sampled 50 different tracks with 10 time steps of path length, and performed a statistical analysis on the performance parameters. In all paths, the starting position is identical.

The results of this study are shown in Fig. 5, which shows average over each of the mentioned metrics ($H(w)$, $\det(\Sigma)$, $\tilde{x}^{w_{max}}$, $\tilde{x}^{w\text{-avrg}}$). In that figure we also study sensitivity to α, which controls the level of viewpoint-dependency in the considered classifier model (18). The plots show a significant improvement of utilizing a classifier model, both for DA disambiguation and inference where the estimation error (Fig. 5c, 5d) and uncertainty (Fig. 5b) are lower when the model is utilized. From all the plots, the most notable performance increase occurs for DA disambiguation (Fig. 5a), where stronger viewpoint dependence assists more significantly; Overall, Fig. 5 presents a strong advantage for utilizing a viewpoint classifier model in the presented scenario.

VI. Conclusions

We presented a recursive Bayesian approach for localization and semantic mapping in ambiguous environments, which maintains and updates incrementally a hybrid belief over camera and object poses, and classification and data association hypotheses. As a key contribution, we incorporated semantic observations and a viewpoint dependent classifier model within the probabilistic formulation and showed these contribute both to data association disambiguation and inference over continuous variables (camera and objects poses). Our simulation results demonstrate the improved performance due to using the viewpoint-dependent classifier model in highly aliased scenarios, yielding faster data association disambiguation, improved localization accuracy and lower estimation uncertainty. Future work will examine the proposed method in larger scenarios and real world experiments.

References

[1] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, Jose Neira, Ian D Reid, and John J Leonard. Simultaneous localization and mapping: Present, future, and the robust-perception age. *IEEE Trans. Robotics*, 32(6):1309 – 1332, 2016.

[2] Beipeng Mu, Shih-Yuan Liu, Liam Paull, John Leonard, and Jonathan How. Slam with objects using a nonparametric pose graph. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2016.

[3] S. Pathak, A. Thomas, and V. Indelman. A unified framework for data association aware robust belief space planning and perception. *Intl. J. of Robotics Research*, 32(2-3):287–315, 2018.

[4] T.E. Fortmann, Y. Bar-Shalom, and M. Scheffe. Multi-target tracking using joint probabilistic data association. In *Proc. 19th IEEE Conf. on Decision & Control*, 1980.

[5] L. Wong, L. P. Kaelbling, and T. Lozano-Pérez. Data association for semantic world modeling from partial views. In *International Symposium for Robotics Research*. Intl. Foundation of Robotics Research, 2013.

[6] N. Sünderhauf and P. Protzel. Switchable constraints for robust pose graph SLAM. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2012.

[7] E. Olson and P. Agarwal. Inference on networks of mixtures for robust robot mapping. *Intl. J. of Robotics Research*, 32(7):826–840, 2013.

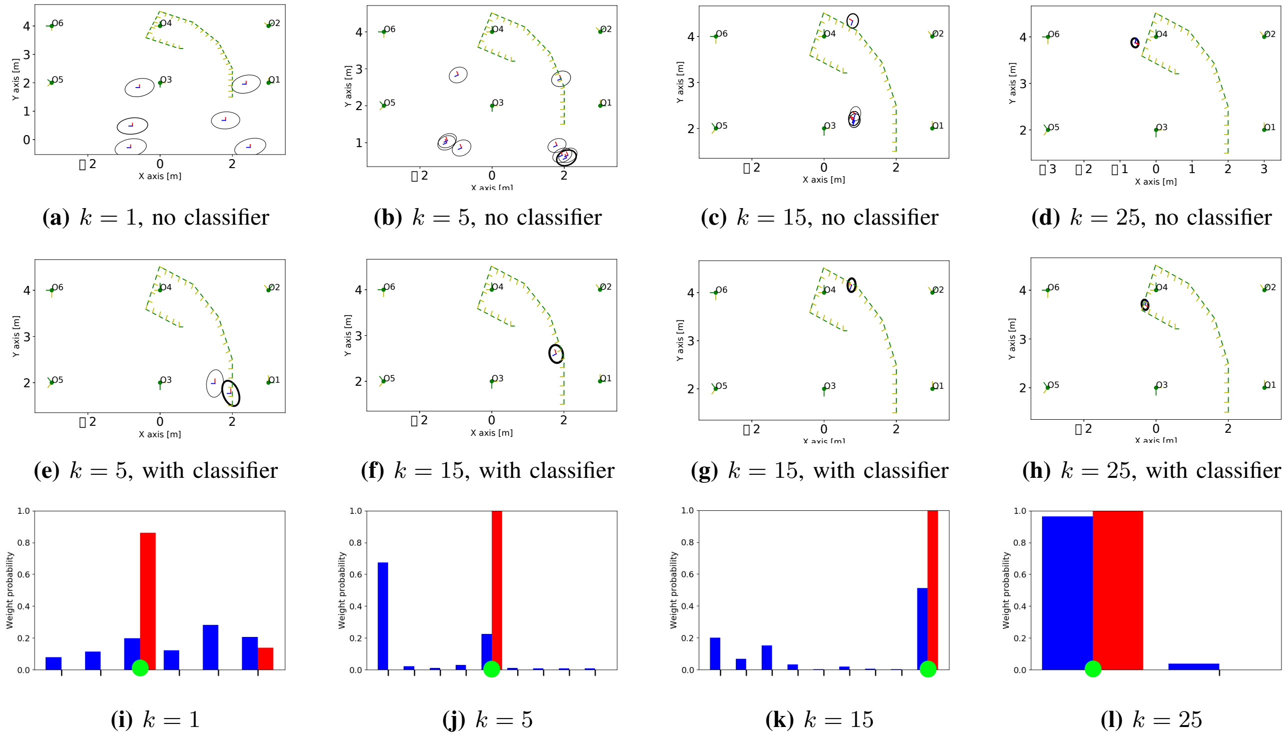
[8] L. Carlone, A. Censi, and F. Dellaert. Selecting good measurements via l1 relaxation: A convex approach for robust estimation over graphs. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 2667–2674. IEEE, 2014.

[9] V. Indelman, E. Nelson, J. Dong, N. Michael, and F. Dellaert. Incremental distributed inference from arbitrary poses and unknown data association: Using collaborating robots to establish a common reference. *IEEE Control Systems Magazine (CSM), Special Issue on Distributed Control and Estimation for Robotic Vehicle Networks*, 36(2):41–74, 2016.

[10] A. Milan, S.H. Rezatofighi, A.R. Dick, I.D. Reid, and K. Schindler. Online multi-target tracking using recurrent neural networks. In *Nat. Conf. on Artificial Intelligence (AAAI)*, pages 4225–4232, 2017.

[11] Hafez Farazi and Sven Behnke. Online visual robot tracking and identification using deep lstm networks. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 6118–6125. IEEE, 2017.

[12] Shayegan Omidshafiei, Brett T Lopez, Jonathan P How, and John Vian. Hierarchical bayesian noise inference for robust real-time probabilistic object classification. *arXiv preprint arXiv:1605.01042*, 2016.

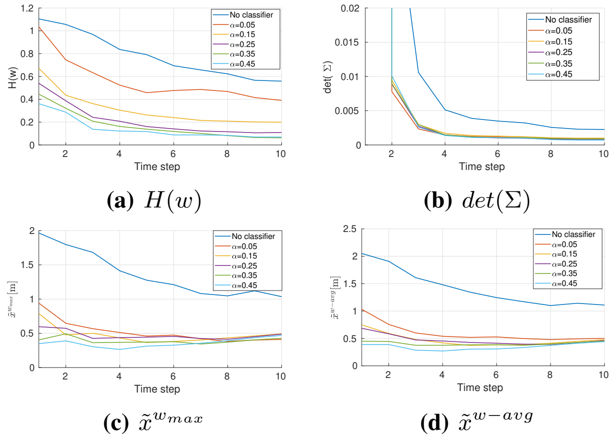


(a) $k = 1$, no classifier **(b)** $k = 5$, no classifier **(c)** $k = 15$, no classifier **(d)** $k = 25$, no classifier

(e) $k = 5$, with classifier **(f)** $k = 15$, with classifier **(g)** $k = 15$, with classifier **(h)** $k = 25$, with classifier

(i) $k = 1$ **(j)** $k = 5$ **(k)** $k = 15$ **(l)** $k = 25$

Fig. 4: **(a)** - **(h)**: Posterior over robot poses of all non-pruned realizations for times $k = 1, 5, 15, 25$, without (first row) and with a classifier model (second row). **Bolder** lines correspond to higher weights. Ground truth trajectory is shown in each of the plots (in terms of camera poses). **(i)**-**(l)**: Corresponding posterior over data association hypotheses, $\mathbb{P}(\beta_{1:k} \mid \mathcal{H}_k)$, at each time. **Blue** bars are without classifier model, **red** bars are with. **Green** circles represent ground truth data associations.



(a) $H(w)$ **(b)** $det(\Sigma)$

(c) $\tilde{x}^{w_{max}}$ **(d)** $\tilde{x}^{w-avg}$

Fig. 5: Effects of different α values on DA disambiguation ability, estimation uncertainty and accuracy in terms of the metrics ($H(w)$, $det(\Sigma)$, $\tilde{x}^{wmax}$, and $\tilde{x}^{w-avg}$), averaged over 50 sampled tracks (see Fig. 3b).

[13] Vladimir Tchuiev and Vadim Indelman. Inference over distribution of posterior class probabilities for reliable bayesian classification and object-level perception. *IEEE Robotics and Automation Letters (RA-L)*, 3(4):4329–4336, 2018.

[14] WT Teacy, Simon J Julier, Renzo De Nardi, Alex Rogers, and Nicholas R Jennings. Observation modelling for vision-based target search by unmanned aerial vehicles. In *Intl. Conf. on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1607–1614, 2015.

[15] Y. Feldman and V. Indelman. Bayesian viewpoint-dependent robust classification under model and localization uncertainty. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2018.

[16] D. Kopitkov and V. Indelman. Robot localization through information recovered from cnn classificators. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*. IEEE, October 2018.

[17] Aleksandr V Segal and Ian D Reid. Hybrid inference optimization for robust pose graph estimation. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2675–2682. IEEE, 2014.

[18] Pierre-Yves Lajoie, Siyi Hu, Giovanni Beltrame, and Luca Carlone. Modeling perceptual aliasing in slam via discrete-continuous graphical models. *IEEE Robotics and Automation Letters*, 2019.

[19] S. Bowman, N. Atanasov, K. Daniilidis, and G. Pappas. Probabilistic data association for semantic slam. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 1722–1729. IEEE, 2017.

[20] F.R. Kschischang, B.J. Frey, and H-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Trans. Inform. Theory*, 47(2):498–519, February 2001.

[21] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. Leonard, and F. Dellaert. iSAM2: Incremental smoothing and mapping using the Bayes tree. *Intl. J. of Robotics Research*, 31:217–236, Feb 2012.

[22] F. Dellaert. Factor graphs and GTSAM: A hands-on introduction. Technical Report GT-RIM-CP&R-2012-002, Georgia Institute of Technology, September 2012.