2017 IEEE International Conference on Robotics and Automation (ICRA)
Singapore, May 29 - June 3, 2017

Probabilistic Data Association for Semantic SLAM

Sean L. Bowman Nikolay Atanasov Kostas Daniilidis George J. Pappas

Abstract— **Traditional approaches to simultaneous localization and mapping (SLAM) rely on low-level geometric features such as points, lines, and planes. They are unable to assign semantic labels to landmarks observed in the environment. Furthermore, loop closure recognition based on low-level features is often viewpoint-dependent and subject to failure in ambiguous or repetitive environments. On the other hand, object recognition methods can infer landmark classes and scales, resulting in a small set of easily recognizable landmarks, ideal for view-independent unambiguous loop closure. In a map with several objects of the same class, however, a crucial data association problem exists. While data association and recognition are discrete problems usually solved using discrete inference, classical SLAM is a continuous optimization over metric information. In this paper, we formulate an optimization problem over sensor states and semantic landmark positions that integrates metric information, semantic information, and data associations, and decompose it into two interconnected problems: an estimation of discrete data association and landmark class probabilities, and a continuous optimization over the metric states. The estimated landmark and robot poses affect the association and class distributions, which in turn affect the robot-landmark pose optimization. The performance of our algorithm is demonstrated on indoor and outdoor datasets.**

I. Introduction

In robotics, simultaneous localization and mapping (SLAM) is the problem of mapping an unknown environment while estimating a robot's pose within it. Reliable navigation, object manipulation, autonomous surveillance, and many other tasks require accurate knowledge of the robot's pose and the surrounding environment. Traditional approaches to SLAM rely on low-level geometric features such as corners [1], lines [2], and surface patches [3] to reconstruct the metric 3-D structure of a scene but are mostly unable to infer semantic content. On the other hand, recent methods for object recognition [4]–[6] can be combined with approximate 3D reconstruction of the environmental layout from single frames using priors [7], [8]. These are rather qualitative single 3D snapshots rather than the more precise mapping we need for a robot to navigate. The goal of this paper is to address the metric and semantic SLAM problems jointly, taking advantage of object recognition to tightly integrate both metric and semantic information into the sensor state and map estimation. In addition to providing a meaningful interpretation of the scene, semantically-labeled landmarks address two critical issues of geometric SLAM: data association (matching sensor observations to map landmarks) and loop closure (recognizing previously-visited locations).

The authors are with GRASP Lab, University of Pennsylvania, Philadelphia, PA 19104, USA, {seanbow, atanasov, kostas, pappasg}@seas.upenn.edu.

We gratefully acknowledge support by TerraSwarm, one of six centers of STARnet, a Semiconductor Research Corporation program sponsored by MARCO and DARPA and the following grants: ARL MAST-CTA W911NF-08-2-0004, ARL RCTA W911NF-10-2-0016.

Approaches to SLAM were initially most often based on filtering methods in which only the most recent robot pose is estimated [9]. This approach is in general very computationally efficient, however because of the inability to estimate past poses and relinearize previous measurement functions, errors can compound [1]. More recently, batch methods that optimize over entire trajectories have gained popularity. Successful batch methods typically represent optimization variables as a set of nodes in a graph (a "pose graph"). Two robot-pose nodes share an edge if an odometry measurement is available between them, while a landmark and a robot-pose node share an edge if the landmark was observed from the corresponding robot pose. This pose graph optimization formulation of SLAM traces back to Lu and Milios [10]. In recent years, the state of the art [11], [12] consists of iterative optimization methods (e.g., nonlinear least squares via the Gauss-Newton algorithm) that achieve excellent performance but depend heavily on linearization of the sensing and motion models. This becomes a problem when we consider including discrete observations, such as detected object classes, in the sensing model.

One of the first systems that used both spatial and semantic representations was proposed by Galindo et al. [13]. A spatial hierarchy contained camera images, local metric maps, and the environment topology, while a semantic hierarchy represented concepts and relations, which allowed room categories to be inferred based on object detections. Many other approaches [14]–[19] extract both metric and semantic information but typically the two processes are carried out separately and the results are merged afterwards. The lack of integration between the metric and the semantic mapping does not allow the object detection confidence to influence the performance of the metric optimization. Focusing on the localization problem only, Atanasov et al. [20] incorporated semantic observations in the metric optimization via a set-based Bayes filter. The works that are closest to ours [21]–[24] consider both localization and mapping and carry out metric and semantic mapping jointly. SLAM++ [22] focuses on a real-time implementation of joint 3-D object recognition and RGB-D SLAM via pose graph optimization. A global optimization for 3D reconstruction and semantic parsing has been proposed by [25], which is the closest work in semantic/geometric joint optmization. The main difference is that 3D space is voxelized and landmarks and/or semantic labels are assigned to voxels which are connected in a conditional random field while our approach allows the estimation of continuous pose of objects. Bao et al. [21] incorporate camera parameters, object geometry, and object classes into a structure from motion problem, resulting in a detailed and accurate but large and expensive optimization. A recent comprehensive survey of semantic mapping can be found in [26].

Most related work uses a somewhat arbitrary decomposition between data association, pose graph optimization, and object recognition. Our work makes the following ***contributions*** to the state of the art:

- our approach is the first to tightly couple inertial, geometric, and semantic observations into a single optimization framework,
- we provide a formal decomposition of the joint metric-semantic SLAM problem into continuous (pose) and discrete (data association and semantic label) optimization sub-problems,
- we carry out experiments on several long-trajectory real indoor and outdoor datasets, which include odometry and visual measurements in cluttered scenes and varying lighting conditions.

II. Probabilistic Data Association in SLAM

Consider the classical localization and mapping problem, in which a mobile sensor moves through an unknown environment, modeled as a collection $\mathcal{L} \triangleq \{\ell_m\}_{m=1}^{M}$ of M static landmarks. Given a set of sensor measurements $\mathcal{Z} \triangleq \{\mathbf{z}_k\}_{k=1}^{K}$, the task is to estimate the landmark positions $\mathcal{L}$ and a sequence of poses $\mathcal{X} \triangleq \{\mathbf{x}_t\}_{t=1}^{T}$ representing the sensor trajectory. Most existing work focuses on estimating $\mathcal{X}$ and $\mathcal{L}$ and rarely emphasizes that the data association $\mathcal{D} \triangleq \{(\alpha_k, \beta_k)\}_{k=1}^{K}$ stipulating that measurement z_k of landmark ℓ_{β_k} was obtained from sensor state x_{α_k} is in fact unknown. A complete statement of the SLAM problem involves maximum likelihood estimation of $\mathcal{X}$, $\mathcal{L}$, and $\mathcal{D}$ given the measurements $\mathcal{Z}$:

$$\hat{\mathcal{X}}, \hat{\mathcal{L}}, \hat{\mathcal{D}} = \arg\max_{\mathcal{X},\mathcal{L},\mathcal{D}} \log p(\mathcal{Z}|\mathcal{X}, \mathcal{L}, \mathcal{D}) \tag{1}$$

The most common approach to this maximization has been to decompose it into two separate estimation problems. First, given prior estimates $\mathcal{X}^0$ and $\mathcal{L}^0$, the maximum likelihood estimate $\hat{\mathcal{D}}$ of the data association $\mathcal{D}$ is computed (e.g., via joint compatability branch and bound [27] or the Hungarian algorithm [28]). Then, given $\hat{\mathcal{D}}$, the most likely landmark and sensor states are estimated[1]:

$$\hat{\mathcal{D}} = \arg\max_{\mathcal{D}} p(\mathcal{D}|\mathcal{X}^0, \mathcal{L}^0, \mathcal{Z}) \tag{2a}$$

$$\hat{\mathcal{X}}, \hat{\mathcal{L}} = \arg\max_{\mathcal{X},\mathcal{L}} \log p(\mathcal{Z}|\mathcal{X}, \mathcal{L}, \hat{\mathcal{D}}) \tag{2b}$$

The second optimization above is typically carried out via filtering [30]–[32] or pose-graph optimization [11], [12].

The above process has the disadvantage that an incorrectly chosen data association may have a highly detrimental effect on the estimation performance. Moreover, if ambiguous measurements are discarded to avoid incorrect association choices, they will never be reconsidered later when refined estimates of the sensor pose (and hence their data association) are available. Instead of a simple one step process, then, it is possible to perform **coordinate descent**, which iterates the two maximization steps as follows:

$$\mathcal{D}^{i+1} = \arg\max_{\mathcal{D}} p(\mathcal{D}|\mathcal{X}^i, \mathcal{L}^i, \mathcal{Z}) \tag{3a}$$

$$\mathcal{X}^{i+1}, \mathcal{L}^{i+1} = \arg\max_{\mathcal{X},\mathcal{L}} \log p(\mathcal{Z}|\mathcal{X}, \mathcal{L}, \mathcal{D}^{i+1}) \tag{3b}$$

This resolves the problem of being able to revisit association decisions once state estimates improve but does little to resolve the problem with ambiguous measurements since a hard decision on data associations is still required. To address this, rather than simply selecting $\hat{\mathcal{D}}$ as the mode of $p(\mathcal{D}|\mathcal{X}, \mathcal{L}, \mathcal{Z})$, we should consider the entire density of $\mathcal{D}$ when estimating $\mathcal{X}$ and $\mathcal{L}$. Given initial estimates $\mathcal{X}^i$, $\mathcal{L}^i$, an improved estimate that utilizes the whole density of $\mathcal{D}$ can be computed by maximizing the expected measurement likelihood via **expectation maximization** (EM):

$$\begin{aligned}\mathcal{X}^{i+1}, \mathcal{L}^{i+1} &= \arg\max_{\mathcal{X},\mathcal{L}} \mathbb{E}_{\mathcal{D}}\big[\log p(\mathcal{Z}|\mathcal{X}, \mathcal{L}, \mathcal{D}) \mid \mathcal{X}^i, \mathcal{L}^i, \mathcal{Z}\big] \\ &= \arg\max_{\mathcal{X},\mathcal{L}} \sum_{\mathcal{D}\in\mathbb{D}} p(\mathcal{D}|\mathcal{X}^i, \mathcal{L}^i, \mathcal{Z}) \log p(\mathcal{Z}|\mathcal{X}, \mathcal{L}, \mathcal{D})\end{aligned} \tag{4}$$

where $\mathbb{D}$ is the space of all possible values of $\mathcal{D}$. This EM formulation has the advantage that no hard decisions on data association are required since it "averages" over all possible associations. To compare this with the coordinate descent formulation in (3), we can rewrite (4) as follows:

$$\begin{aligned}&\arg\max_{\mathcal{X},\mathcal{L}} \sum_{\mathcal{D}\in\mathbb{D}} \sum_{k=1}^{K} p(\mathcal{D}|\mathcal{X}^i, \mathcal{L}^i, \mathcal{Z}) \log p(\mathbf{z}_k|\mathbf{x}_{\alpha_k}, \ell_{\beta_k}) \\ &\quad = \arg\max_{\mathcal{X},\mathcal{L}} \sum_{k=1}^{K} \sum_{j=1}^{M} w_{kj}^i \log p(\mathbf{z}_k|\mathbf{x}_{\alpha_k}, \ell_j)\end{aligned} \tag{5}$$

where $w_{kj}^i \triangleq \sum_{\mathcal{D}\in\mathbb{D}(k,j)} p(\mathcal{D}|\mathcal{X}^i, \mathcal{L}^i, \mathcal{Z})$ is a weight, independent of the optimization variables $\mathcal{X}$ and $\mathcal{L}$, that quantifies the influence of the "soft" data association, and $\mathbb{D}(k, j) \triangleq \{\mathcal{D} \in \mathbb{D} \mid \beta_k = j\} \subseteq \mathbb{D}$ is the set of all data associations such that measurement k is assigned to landmark j. Note that the coordinate descent optimization (3b) has a similar form to (5), except that for each k there is exactly one j such that $w_{kj}^i = 1$ and $w_{kl}^i = 0$ for all $l \neq j$.

We can also show that the EM formulation, besides being a generalization of coordinate descent, is equivalent to the following matrix permanent maximization problem.

Proposition 1. *If $p(\mathcal{D} \mid \mathcal{X}^i, \mathcal{L}^i)$ is uniform, the maximizers of the EM formulation in (4) and the optimization below are equal:*

$$\mathcal{X}^{i+1}, \mathcal{L}^{i+1} = \arg\max_{\mathcal{X},\mathcal{L}} \mathbf{per}(\mathbf{Q}^i(\mathcal{X}, \mathcal{L})),$$

where $\mathbf{per}$ *denotes the matrix permanent*[2]*, $\mathbf{Q}^i(\mathcal{X}, \mathcal{L})$ is a matrix with elements $\big[\mathbf{Q}^i\big]_{kj} := p(\mathbf{z}_k|\mathbf{x}_j^i, \ell_j^i)p(\mathbf{z}_k|\mathbf{x}_j, \ell_j)$ and $\{(\mathbf{x}_j^i, \ell_j^i)\}$ and $\{(\mathbf{x}_j, \ell_j)\}$ are enumerations of the sets $\mathcal{X}^i \times \mathcal{L}^i$ and $\mathcal{X} \times \mathcal{L}$, respectively.*

Proof. See Appendix I. □

Similar to the coordinate descent formulation, the EM for-

[1]Note that the first maximization in (2a) assumes that $p(\mathcal{D}|\mathcal{X}^0, \mathcal{L}^0)$ is uniform. This is true when there are no false positive measurements or missed detections. A more sophisticated model can be obtained using ideas from [29].

[2]The permanent of an $n \times m$ matrix $A = [A(i, j)]$ with $n \leq m$ is defined as $\mathbf{per}(A) := \sum_{\pi} \prod_{i=1}^{n} A(i, \pi(i))$, where the sum is over all one-to-one functions $\pi : \{1, \ldots, n\} \to \{1, \ldots, m\}$.

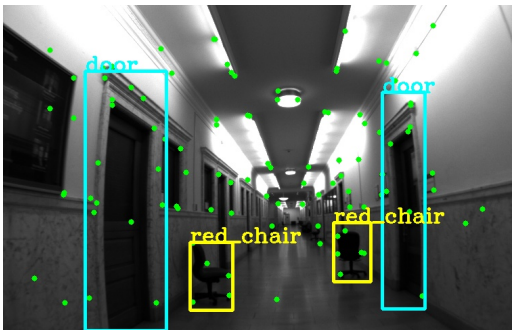

Fig. 1: Example keyframe image overlaid with ORB features (green points) and object detections

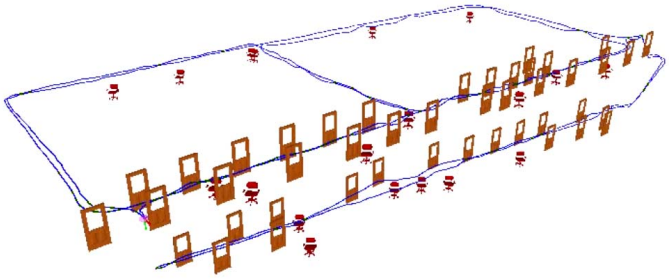

Fig. 2: Estimated sensor trajectory (blue) and landmark positions and classes using inertial, geometric, and semantic measurements such as those in Fig. 1. The accompanying video shows the estimation process in real time.

mulation (5) allows us to solve the permanent maximization problem iteratively. First, instead of estimating a maximum likelihood data association, we estimate the data association distribution $p(\mathcal{D}|\mathcal{X}^i, \mathcal{L}^i, \mathcal{Z})$ in the form of the weights w^i_{kj} (the "E" step). Then, we maximize the expected measurement log likelihood over the previously computed distribution (the "M" step).

III. Semantic SLAM

In the rest of the paper, we focus on a particular formulation of the SLAM problem that in addition to sensor and landmark poses involves *landmark classes* (*e.g.*, door, chair, table) and *semantic measurements* in the form of object detections. We will demonstrate that the expectation maximization formulation (5) is an effective way to solve the semantic SLAM problem.

Let the state ℓ of each landmark consist of its position $\ell^p \in \mathbb{R}^3$ as well as a class label ℓ^c from a discrete set $\mathcal{C} = \{1, \ldots, C\}$. To estimate the landmark states $\mathcal{L}$ and sensor trajectory $\mathcal{X}$, we utilize three sources of information: inertial, geometric point features, and semantic object observations. Examples of geometric features and semantic observations can be seen in Figure 1.

A. Inertial information

We assume that the sensor package consists of an inertial measurement unit (IMU) and one monocular camera. A subset of the images captured by the camera are chosen as *keyframes* (*e.g.*, by selecting every nth frame as a keyframe). The sensor state corresponding to the tth keyframe is denoted x_t and consists of the sensor 6-D pose, velocity, and IMU bias values. We assume that the IMU and camera are time synchronized, so between keyframes t and $t+1$, the sensor also collects a set $\mathcal{I}_t$ of IMU measurements (linear acceleration and rotational velocity).

B. Geometric information

In addition to the inertial measurements $\mathcal{I}_t$, we utilize geometric point measurements (*e.g.*, Harris corners, SIFT, SURF, FAST, BRISK, ORB, etc.) $\mathcal{Y}_t$. From each keyframe image, these geometric point features are extracted and tracked forward to the subsequent keyframe. In our experiments we extract ORB features [33] from each keyframe and match them to the subsequent keyframe by minimizing the ORB descriptor distance. Since these features are matched by an external method, we assume that their data association is known.

C. Semantic information

The last type of measurement used are object detections $\mathcal{S}_t$ extracted from every keyframe image. An object detection $\mathbf{s}_k = (s^c_k, s^s_k, s^b_k) \in \mathcal{S}_t$ extracted from keyframe t consists of a detected class $s^c_k \in \mathcal{C}$, a score s^s_k quantifying the detection confidence, and a bounding box s^b_k. Such information can be obtained from any modern approach for object recognition such as [5], [34]–[36]. In our implementation, we use a deformable parts model (DPM) detector [4], [37], [38], which runs on a CPU in real time. If the data association $\mathcal{D}_k = (\alpha_k, \beta_k)$ of measurement $\mathbf{s}_k$ is known, the measurement likelihood can be decomposed as follows: $p(\mathbf{s}_k|\mathbf{x}_{\alpha_k}, \ell_{\beta_k}) = p(s^c_k|\ell^c_{\beta_k}) p(s^s_k|\ell^c_{\beta_k}, s^c_k) p(s^b_k|\mathbf{x}_{\alpha_k}, \ell^p_{\beta_k})$. The density $p(s^c_k|\ell^c_{\beta_k})$ corresponds to the confusion matrix of the object detector and is learned offline along with the score distribution $p(s^s_k|\ell^c_{\beta_k}, s^c_k)$. The bounding-box likelihood $p(s^b_k|\mathbf{x}_{\alpha_k}, \ell^p_{\beta_k})$ is assumed normally distributed with mean equal to the perspective projection of the centroid of the object onto the image plane and covariance proportional to the dimensions of the detected bounding box.

Problem (Semantic SLAM)**.** *Given inertial* $\mathcal{I} \triangleq \{\mathcal{I}_t\}_{t=1}^T$, *geometric* $\mathcal{Y} \triangleq \{\mathcal{Y}_t\}_{t=1}^T$, *and semantic* $\mathcal{S} \triangleq \{\mathcal{S}_t\}_{t=1}^T$ *measurements, estimate the sensor state trajectory* $\mathcal{X}$ *and the positions and classes* $\mathcal{L}$ *of the objects in the environment.*

The inertial and geometric measurements are used to track the sensor trajectory locally and, similar to a visual odometry approach, the geometric structure is not recovered. The semantic measurements, in contrast, are used to construct a map of objects that can be used to perform loop closure that is robust to ambiguities and viewpoint and is more efficient than a SLAM approach that maintains full geometric structure.

IV. Semantic SLAM using EM

Following the observations from Sec. II, we apply expectation maximization to robustly handle the semantic data association. In addition to treating data association as a latent variable, we also treat the discrete landmark class labels as latent variables in the optimization, resulting in a clean and efficient separation between discrete and continuous variables. As mentioned in Sec. III, the data association of the geometric measurements is provided by the feature tracking algorithm,

so the latent variables we use are the data association $\mathcal{D}$ of the semantic measurements measurements and the object classes $\ell^c_{1:M}$. The following proposition specifies the EM steps necessary to solve the semantic SLAM problem. The initial guess $\mathcal{X}^{(0)}$ is provided by odometry integration; the initial guess $\mathcal{L}^{(0)}$ can be obtained from $\mathcal{X}^{(0)}$ by initializing a landmark along the detected camera ray.

Proposition 2. *If $p(\mathcal{D}|\mathcal{X},\mathcal{L})$ is uniform and the semantic measurement data associations are independent across keyframes, i.e., $p(\mathcal{D}|\mathcal{S},\mathcal{X},\mathcal{L}) = \prod_{t=1}^{T} p(\mathcal{D}_t|\mathcal{S}_t,\mathcal{X},\mathcal{L})$,[3] the semantic SLAM problem can be solved via the expectation maximization algorithm by iteratively solving for (1) data association weights w^t_{ij} (the "E" step) and (2) continuous sensor states $\mathcal{X}$ and landmark positions $\ell^p_{1:M}$ (the "M" step) via the following equations:*

$$w^{t,(i)}_{kj} = \sum_{\ell^c\in\mathcal{C}} \sum_{\mathcal{D}_t\in\mathbb{D}_t(k,j)} \kappa^{(i)}(\mathcal{D}_t,\ell^c) \quad \forall t,k,j \tag{6}$$

$$\mathcal{X}^{(i+1)},\ell^{p,(i+1)}_{1:M} = \arg\min_{\mathcal{X},\ell^p_{1:M}} \sum_{t=1}^{T}\sum_{\mathbf{s}_k\in\mathcal{S}_t}\sum_{j=1}^{M} -w^{t,(i)}_{kj}\log p(\mathbf{s}_k|\mathbf{x}_t,\ell_j) - \log p(\mathcal{Y}|\mathcal{X}) - \log p(\mathcal{I}|\mathcal{X}) \tag{7}$$

where

$$\kappa^{(i)}(\mathcal{D}_t,\ell^c) = \frac{p(\mathcal{S}_t|\mathcal{X}^{(i)},\mathcal{L}^{(i)},\mathcal{D}_t)}{\sum_{\ell^c}\sum_{\mathcal{D}_t\in\mathbb{D}_t} p(\mathcal{S}_t|\mathcal{X}^{(i)},\mathcal{L}^{(i)},\mathcal{D}_t)},$$

$\mathbb{D}_t$ is the set of all possible data associations for measurements received at timestep t, and $\mathbb{D}_t(i,j)\subseteq\mathbb{D}_t$ is the set of all possible data associations for measurements received at time t such that measurement i is assigned to landmark j.

Proof. See Appendix II. □

A. Object classes and data association (E step)

The computation of the weights for a single keyframe require several combinatorial sums over all possible data associations. However, due to the assumption of independent associations among keyframes and the fact that only few objects are present within the sensor field-of-view, it is feasible to compute the summations and hence w^t_{kj} for all keyframes t, measurements k, and landmarks j extremely efficiently in practice. Once the weights $w^{t,(i)}_{kj}$ are computed for each measurement-landmark pair, they are used within the continuous optimization over sensor states and landmark positions. Additionally, maximum likelihood landmark class estimates ℓ^c can be recovered from the computed κ values:

$$\hat{\ell}^c_{1:M} = \arg\max_{\ell^c} p(\ell^c_{1:M}|\theta,\mathcal{Z}) = \arg\max_{\ell^c}\prod_{t=1}^{T}\sum_{\mathcal{D}_t\in\mathbb{D}_t}\kappa(\mathcal{D}_t,\ell^c)$$

B. Pose graph optimization (M step)

Equation (7) forms the basis of our pose graph optimization over sensor states and landmark positions. A pose graph is a convenient way of representing an optimization problem for which there exists a clear physical structure or a sparse constraint set. The graph consists of a set of vertices $\mathcal{V}$, each of which corresponds to an optimization variable, and a set of factors $\mathcal{F}$ among the vertices that correspond to individual components of the cost function. Graphically, a factor is a generalization of an edge that allows connectivity between more than two vertices. A factor f in the graph is associated with a cost function that depends on a subset of the variables $\mathcal{V}$ such that the entire optimization is of the form

$$\hat{\mathcal{V}} = \arg\min_{\mathcal{V}} \sum_{f\in\mathcal{F}} f(\mathcal{V}) \tag{8}$$

In addition to providing a useful representation, factor graphs are advantageous in that there exist computational tools that allow efficient optimization [11], [39].

Our graph has a vertex for each sensor state $\mathbf{x}_t$ and for each landmark position ℓ^p_i. Contrary to most prior work in which a hard data association decision results in a measurement defining a single factor between a sensor pose and a landmark, we consider soft semantic data association multiple factors.

1) Semantic Factors: A measurement $\mathbf{s}_k$ from sensor state $\mathbf{x}_i$ defines factors $f^s_{kj}(\mathbf{x}_i,\ell_j)$ for each visible landmark j. Assuming the number of visible landmarks and the number of received measurements are approximately equal, with this method the number of semantic factors in the graph is roughly squared. Note that since ℓ^c is fixed in (7), $p(s^s|\ell^c,s^c)$ and $p(s^c|\ell^c)$ are constant. Thus, $\log p(\mathbf{s}|\mathbf{x},\ell) = \log p(s^b|\mathbf{x},\ell^p) + \log p(s^s|\ell^c,s^c)p(s^c|\ell^c)$ and so the latter term can be dropped from the optimization.

Let $h_\pi(\mathbf{x},\ell^p)$ be the standard perspective projection of a landmark ℓ^p onto a camera at pose $\mathbf{x}$. We assume that the camera measurement of a landmark ℓ^p from camera pose $\mathbf{x}$ is Gaussian distributed with mean $h_\pi(\mathbf{x},\ell^p)$ and covariance $\mathbf{R}_s$. Thus, a camera factor corresponding to sensor state t, measurement k, and landmark j, f^s_{kj}, becomes

$$f^s_{kj}(\mathcal{X},\mathcal{L}) = -w^{t,(i)}_{kj}\log p(s^b_k|\mathbf{x}_t,\ell^p_j) \tag{9}$$

$$= \|s^b_k - h_\pi(\mathbf{x}_t,\ell_j)\|^2_{\mathbf{R}_s/w^{t,(i)}_{kj}} \tag{10}$$

Those semantic factors due to the re-observation of a previously seen landmark are our method's source of loop closure constraints.

2) Geometric Factors: Following [30], [40], we incorporate geometric measurements into the pose graph as structureless constraints between the camera poses that observed them. We can rewrite the term corresponding to geometric factors in (7) as

$$-\log p(\mathcal{Y}|\mathcal{X}) = -\sum_{i=1}^{N_y}\sum_{k:\beta^y_k=i}\log p(\mathbf{y}_k|\mathbf{x}_{\alpha^y_k}) \tag{11}$$

where N_y is the total number of distinct feature tracks, i.e. the total number of observed physical geometric landmarks.

Letting $\rho_{\beta^y_k}$ be the 3D position in the global frame of the landmark that generated measurement $\mathbf{y}_k$, and assuming as before that the projection has Gaussian pixel noise with covariance $\mathbf{R}_y$, we have

$$-\log p(\mathcal{Y}|\mathcal{X}) = \sum_{i=1}^{N_y}\sum_{k:\beta^y_k=i}\|\mathbf{y}_k - h_\pi(\mathbf{x}_{\alpha^y_k},\rho_i)\|^2_{\mathbf{R}_y} \tag{12}$$

[3]This "naïve Bayes" assumption might not always hold perfectly in practice but it significantly simplifies the optimization and allows for efficient implementation.

For a single observed landmark ρ_i, the factor constraining the camera poses which observed it takes the form

$$f_i^y(\mathcal{X}) = \sum_{k:\beta_k^y=i} \|\mathbf{y}_k - h_\pi(\mathbf{x}_{\alpha_k^y}, \rho_i)\|^2_{\mathbf{R}_y} \tag{13}$$

Because we use iterative methods to optimize the full pose graph, it is necessary to linearize the above cost term. The linearization of the above results in a cost term of the form

$$\sum_{k:\beta_k^y=i} \|\mathbf{H}_{ik}^\rho \delta\rho_i + \mathbf{H}_{ik}^{\mathbf{x}} \delta\mathbf{x}_{\alpha_k^y} + \mathbf{b}_{ik}\|^2 \tag{14}$$

where $\mathbf{H}_{ik}^\rho$ is the Jacobian of the cost function with respect to $\rho_{\beta_k^y}$, $\mathbf{H}_{ik}^{\mathbf{x}}$ is the Jacobian with respect to $\mathbf{x}_{\alpha_k^y}$, $\mathbf{b}_{ik}$ is a function of the measurement and its error, and the linearized cost term is in terms of deltas $\delta\mathbf{x}$, $\delta\rho$ rather than the true values $\mathbf{x}$, ρ.

Writing the inner summation in one matrix form by stacking the individual components, we can write this simply as $\|\mathbf{H}_i^\rho \delta\rho_i + \mathbf{H}_i^{\mathbf{x}} \delta\mathbf{x}_{\alpha^y(i)} + \mathbf{b}_i\|^2$. To avoid optimizing over ρ values, and hence to remove the dependence of the cost function upon them, we project the cost into the null space of its Jacobian. We premultiply each cost term by $\mathbf{A}_i$, a matrix whose columns span the left nullspace of $\mathbf{H}_i^\rho$. The cost term for the structureless geometric features thus becomes a function of only the states which observe it:

$$\|\mathbf{A}_i \mathbf{H}_i^{\mathbf{x}} \delta\mathbf{x}_{\alpha^y(i)} + \mathbf{A}_i \mathbf{b}_i\|^2 \tag{15}$$

3) Inertial Factors: To incorporate the accelerometer and gyroscope measurements into the pose graph, we use the method of preintegration factors detailed in [40]. The authors provide an efficient method of computing inertial residuals between two keyframes $\mathbf{x}_i$ and $\mathbf{x}_j$ in which several inertial measurements were received. By "preintegrating" all IMU measurements received between the two keyframes, the relative pose difference (*i.e.* difference in position, velocity, and orientation) between the two successive keyframes is estimated. Using this estimated relative pose, the authors provide expressions for inertial residuals on the rotation ($\mathbf{r}_{\Delta R_{ij}}$), velocity ($\mathbf{r}_{\Delta \mathbf{v}_{ij}}$), and position ($\mathbf{r}_{\Delta \mathbf{p}_{ij}}$) differences between two keyframes as a function of the poses $\mathbf{x}_i$ and $\mathbf{x}_j$. Specifically, they provide said expressions along with their noise covariances $\mathbf{\Sigma}$ such that

$$f_i^{\mathcal{I}}(\mathcal{X}) = -\log p(\mathcal{I}_{ij}|\mathcal{X}) \tag{16}$$

$$= \|\mathbf{r}_{\mathcal{I}_{ij}}\|^2_{\mathbf{\Sigma}_{ij}} \tag{17}$$

The full pose graph optimization corresponding to equation (7) is then a nonlinear least squares problem involving semantic observation terms (see (10)), geometric observation terms (see (15)), and inertial terms (see (17)).

$$\hat{\mathbf{x}}_{1:T}, \hat{\ell}_{1:M} = \arg\min_{\mathcal{X}, \ell_{1:M}} \sum_{k=1}^{K} \sum_{j=1}^{M} f_{kj}^s(\mathcal{X}, \ell_{1:M}^p) + \sum_{i=1}^{N_y} f_i^y(\mathcal{X}) + \sum_{t=1}^{T-1} f_t^{\mathcal{I}}(\mathcal{X}) \tag{18}$$

We solve this within the iSAM2 framework [12], which is able to provide a near-optimal solution with real-time performance.

V. Experiments

We implemented our algorithm in C++ using GTSAM [39] and its iSAM2 implementation as the optimization back-end. All experiments were able to be computed in real-time.

The front-end in our implementation simply selects every 15th camera frame as a keyframe. As mentioned in section III-B, the tracking front-end extracts ORB features [33] from every selected keyframe and tracks them forward through the images by matching the ORB descriptors. Outlier tracks are eliminated by estimating the essential matrix between the two views using RANSAC and removing those features which do not fit the estimated model. We assume that the timeframe between two subsequent images is short enough that the orientation difference between the two frames can be estimated accurately by integrating the gyroscope measurements. Thus, only the unit translation vector between the two images needs to be estimated. We can then estimate the essential matrix using only two point correspondences [41].

The front-end's object detector is an implementation of the deformable parts model detection algorithm [38]. On the acquisition of the semantic measurements from a new keyframe, the Mahalanobis distance from the measurement to all known landmarks is computed. If all such distances are above a certain threshold, a new landmark is initialized in the map, with initial position estimate along the camera ray, with depth given by the median depth of all geometric feature measurements within its detected bounding box (or some fixed value if no such features were tracked successfully).

While ideally we would iterate between solving for constraint weights w_{ij} and poses as proposition 2 suggests, in practice for computational reasons we solve for the weights just once per keyframe.

Our experimental platform was a VI-Sensor [42] from which we used the IMU and left camera. We performed three seperate experiments. The first consists of a medium length (approx. 175 meters) trajectory around one floor of an office building, in which the object classes detected and kept in the map were two types of chairs (red office chairs and brown four-legged chairs). The second experiment is a long (approx. 625 meters) trajectory around two different floors of an office building. The classes in the second experiment are red office chairs and doors. The third and final trajectory is several loops around a room equipped with a vicon motion tracking system, in which the only class of objects detected is red office chairs. In addition to our own experiments, we applied our algorithm to the KITTI dataset [43] odometry sequences 05 and 06.

The final trajectory estimate along with the estimated semantic map for the first office experiment is shown in Fig. 3. The trajectories estimated by our algorithm, by the ROVIO visual-inertial odometry algorithm [31], and by the ORB-SLAM2 visual SLAM algorithm [44], [45], projected into the x-y plane, are shown in Fig. 4. Due to a lack of inertial information and a relative lack of visual features in the environment, ORB-SLAM2 frequently got lost and much of the trajectory estimate is missing, but was always able to recover when entering a previously mapped region.

The second office experiment trajectory along with the

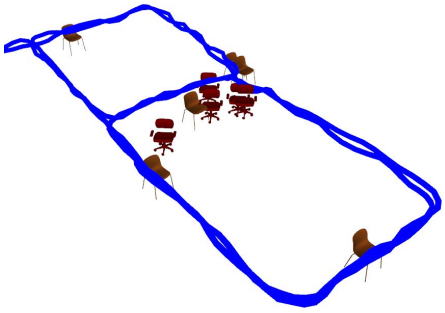

Fig. 3: Sensor trajectory and estimated landmarks for the first office experiment

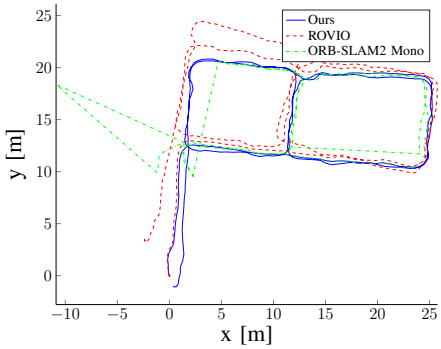

Fig. 4: Estimated trajectories in first office experiment.

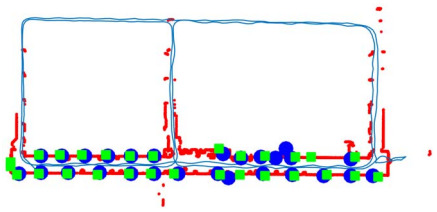

Fig. 5: Estimated trajectory in second office experiment from our algorithm (blue line) along with our estimated door landmark positions (blue circles), overlaid onto partial ground truth map (red) along with ground truth door locations (green squares)

estimated map is shown in Fig. 2. An example image overlaid with object detections from near the beginning of this trajectory is displayed in Fig. 1. We constructed a partial map of the top floor in the experiment using a ground robot equipped with a lidar scanner. On this ground truth map, we manually picked out door locations. The portion of the estimated trajectory on the top floor is overlayed onto this partial truth map (the two were manually aligned) in Fig. 5, Due to the extremely repetitive nature of the hallways in this experiment, bag-of-words based loop closure detections are subject to false positives and incorrect matches. ORB-SLAM2 was unable to successfully estimate the trajectory due to such false loop closures. A partial trajectory estimate after an incorrect loop closure detection is shown in Fig. 6.

The vicon trajectory and the estimated map of chairs is shown in Fig. 7. We evaluated the position error with respect to the vicon's estimate for our algorithm, ROVIO, and ORB-SLAM2 and the results are shown in Fig. 8. Note that the spikes in the estimate errors are due to momentary occlusion from the vicon cameras.

We also evaluated our algorithm on the KITTI outdoor dataset, using odometry sequences 05 and 06. The semantic objects detected and used in our algorithm were cars. Rather than use inertial odometry in this experiment, we used the VISO2 [46] visual odometry algorithm as the initial guess $\mathcal{X}^{(0)}$ for a new keyframe state. Similarly, we replaced the preintegrated inertial relative pose (cf. Sec. IV-B.3) with the relative pose obtained from VISO in the odometry factors. The absolute position errors over time for KITTI sequence 05 with respect to ground truth for our algorithm, VISO2, and ORB-SLAM2 with monocular and stereo cameras are shown in Fig. 9. The same for sequence 06 are shown in Fig. 10. Finally, the mean translational and rotational errors over all possible subpaths of length (100, 200, ..., 800) meters are shown in Fig. 11.

VI. Conclusion

The experiments demonstrated that in complex and cluttered real-world datasets our method can be used to reconstruct the full 6-D pose history of the sensor and the positions and classes of the objects contained in the environment. The advantage of our work is that by having semantic features directly into the optimization, we include a relatively sparse and easily distinguishable set of features that allows for improved localization performance and loop closure, while only slightly impacting the computational cost of the algorithm. Furthermore, semantic information about the environment is valuable in and of itself in aiding autonomous operation of robots within a human-centric environment.

In future work, we plan to expand our algorithm to estimate the full pose of the semantic objects (*i.e.*, orientation in addition to position). We also plan to fully exploit our EM decomposition by reconsidering data associations for past keyframes, and to consider systems with multiple sensors and non-stationary objects.

Appendix I: Proof of Proposition 1

First, we rewrite the optimization in (4) without a logarithm and similarly expand the expectation:

$$\mathcal{X}^{i+1},\mathcal{L}^{i+1} = \arg\max_{\mathcal{X},\mathcal{L}} \sum_{\mathcal{D}\in\mathbb{D}} p(\mathcal{D}|\mathcal{X}^i,\mathcal{L}^i,\mathcal{Z})p(\mathcal{Z}|\mathcal{X},\mathcal{L},\mathcal{D})$$

The data association likelihood can then be rewrite as

$$p(\mathcal{D}|\mathcal{X}^i,\mathcal{L}^i,\mathcal{Z}) = \frac{p(\mathcal{Z}|\mathcal{X}^i,\mathcal{L}^i,\mathcal{D})p(\mathcal{D}|\mathcal{X}^i,\mathcal{L}^i)}{\sum_{\mathcal{D}} p(\mathcal{Z}|\mathcal{X}^i,\mathcal{L}^i,\mathcal{D})p(\mathcal{D}|\mathcal{X}^i,\mathcal{L}^i)} \quad (19)$$

$$= \frac{p(\mathcal{Z}|\mathcal{X}^i,\mathcal{L}^i,\mathcal{D})}{\sum_{\mathcal{D}} p(\mathcal{Z}|\mathcal{X}^i,\mathcal{L}^i,\mathcal{D})} \quad (20)$$

with the last equality due to the assumption that $p(\mathcal{D}|\mathcal{X},\mathcal{L})$ is uniform. We can next decompose the measurement likelihood $p(\mathcal{Z}|\mathcal{X},\mathcal{L},\mathcal{D}) = \prod_k p(\mathbf{z}_k|\mathbf{x}_{\alpha_k},\ell_{\beta_k})$, and so

$$\mathcal{X}^{i+1},\mathcal{L}^{i+1} = \arg\max_{\mathcal{X},\mathcal{L}} \sum_{\mathcal{D}\in\mathbb{D}} p(\mathcal{D}|\mathcal{X}^i,\mathcal{L}^i,\mathcal{Z})p(\mathcal{Z}|\mathcal{X},\mathcal{L},\mathcal{D}) \quad (21)$$

$$= \arg\max_{\mathcal{X},\mathcal{L}} \sum_{\mathcal{D}\in\mathbb{D}} \prod_k \frac{p(\mathbf{z}_k|\mathbf{x}^i_{\alpha_k},\ell^i_{\beta_k})p(\mathbf{z}_k|\mathbf{x}_{\alpha_k},\ell_{\beta_k})}{\sum_{\mathcal{D}} p(\mathcal{Z}|\mathcal{X}^i,\mathcal{L}^i,\mathcal{D})}$$

The result then follows by noting that the normalizing denominator is independent of the optimization variables and from the definition of the matrix permanent.

Appendix II: Proof of Proposition 2

Suppose we have some initial guess given by $\theta^{(i)} = \{\mathcal{X}^{(i)},\ell^{p,(i)}\}$. We can then compute an improved estimate of $\theta = \{\mathcal{X},\ell^p\}$ by maximizing the expected log likelihood:

$$\theta^{(i+1)} = \arg\max_{\theta} \mathbb{E}_{\mathcal{D},\ell^c|\theta^{(i)}}\left[\log p(\mathcal{D},\ell^c,\mathcal{S},\mathcal{Y},\mathcal{I}|\theta)\right] \quad (22)$$

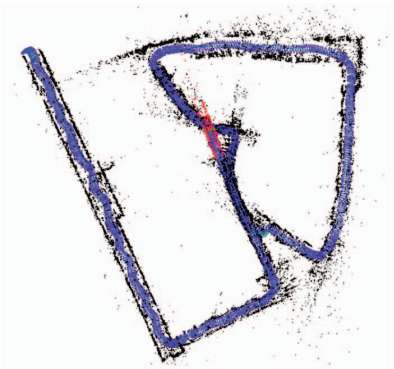

Fig. 6: Partial ORB-SLAM2 trajectory after incorrect loop closure in second office experiment.

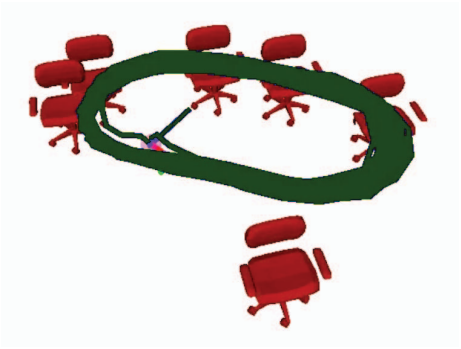

Fig. 7: Sensor trajectory and estimated landmarks for the vicon experiment

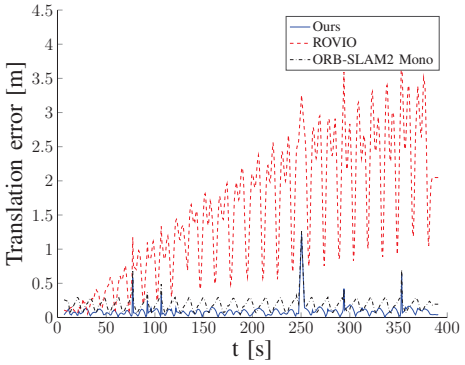

Fig. 8: Position errors with respect to vicon ground truth.

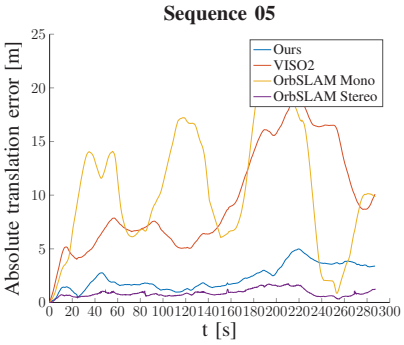

Fig. 9: Norm of position error between estimate and ground truth, KITTI seq. 05

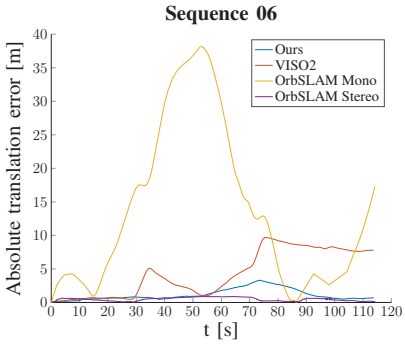

Fig. 10: Norm of position error between estimate and ground truth, KITTI seq. 06

KITTI Sequence 05		
Method	Trans. err [%]	Rot. err [deg/m]
Ours	1.31	0.0038
VISO2	4.08	0.0050
ORBSLAM2 Mono	5.39	0.0019
ORBSLAM2 Stereo	0.63	0.0017

KITTI Sequence 06		
Method	Trans. err [%]	Rot. err [deg/m]
Ours	0.77	0.0037
VISO2	1.81	0.0036
ORBSLAM2 Mono	6.71	0.0015
ORBSLAM2 Stereo	0.29	0.0013

Fig. 11: KITTI mean translational and rotational error over path lengths (100, 200, . . . , 800) meters.

Expanding the expectation,

$$\mathbb{E}_{\mathcal{D},\ell^c|\theta^{(i)}}\left[\log p(\mathcal{D},\ell^c,\mathcal{S},\mathcal{Y},\mathcal{I}|\theta)\right] \tag{23}$$
$$= \sum_{\mathcal{D},\ell^c} p(\mathcal{D},\ell^c|\mathcal{S},\theta^{(i)}) \log[p(\mathcal{S},\mathcal{D},\ell^c|\theta)p(\mathcal{Y}|\theta)p(\mathcal{I}|\theta)]$$

Letting $\kappa(\mathcal{D},\ell^c) \triangleq p(\mathcal{D},\ell^c|\mathcal{S},\theta^{(i)})$, a constant with respect to the optimization variables, we continue:

$$\mathbb{E}[\cdot] = \sum_{\mathcal{D},\ell^c} \kappa(\mathcal{D},\ell^c)\log p(\mathcal{S},\mathcal{D},\ell^c|\theta) + \sum_{\mathcal{D},\ell^c} \kappa(\mathcal{D},\ell^c)\log[p(\mathcal{Y}|\theta)p(\mathcal{I}|\theta)] \tag{24}$$
$$= \sum_{\mathcal{D},\ell^c} \kappa(\mathcal{D},\ell^c)\log p(\mathcal{S},\mathcal{D},\ell^c|\theta) + \log p(\mathcal{Y}|\theta) + \log p(\mathcal{I}|\theta)$$

Focusing on the leftmost summation over data associations and landmark classes,

$$\sum_{\mathcal{D},\ell^c} \kappa(\mathcal{D},\ell^c)\log p(\mathcal{S},\mathcal{D},\ell^c|\theta) \tag{25}$$
$$= \sum_{\mathcal{D},\ell^c} \kappa(\mathcal{D},\ell^c)\log p(\mathcal{S}|\mathcal{D},\ell^c,\theta) + \sum_{\mathcal{D},\ell^c} \kappa(\mathcal{D},\ell^c)\log p(\mathcal{D},\ell^c|\theta)$$

Using the assumption that $p(\mathcal{D},\ell^c|\theta)$ is a uniform distribution over the space of data associations and landmark classes, this term doesn't affect which θ maximizes the objective, so for optimization purposes we have

$$\sum_{\mathcal{D},\ell^c} \kappa(\mathcal{D},\ell^c)\log p(\mathcal{S},\mathcal{D},\ell^c|\theta)$$
$$= \sum_{\mathcal{D},\ell^c} \kappa(\mathcal{D},\ell^c)\log p(\mathcal{S}|\mathcal{D},\ell^c,\theta) \tag{26}$$
$$= \sum_t \sum_i \sum_{\mathcal{D}_t,\ell^c} \kappa(\mathcal{D}_t,\ell^c)\log p(\mathbf{s}_i|\mathbf{x}_t,\ell_{\beta_i}) \tag{27}$$

Note that if we let $\mathbb{D}(i,j)$ be the subset of all possible data associations that assign measurment i to landmark j, we can further decompose this summation as

$$\sum_{\mathcal{D},\ell^c} \kappa(\mathcal{D},\ell^c)\log p(\mathcal{S},\mathcal{D},\ell^c|\theta)$$
$$= \sum_t \sum_i \sum_j \sum_{\ell^c} \sum_{\mathcal{D}_t \in \mathbb{D}(i,j)} \kappa(\mathcal{D}_t,\ell^c)\log p(\mathbf{s}_i|\mathbf{x}_t,\ell_j) \tag{28}$$

Finally, letting $w_{ij}^t \triangleq \sum_{\ell^c} \sum_{\mathcal{D}_t \in \mathbb{D}(i,j)} \kappa(\mathcal{D}_t,\ell^c)$, we can write the final expectation maximization as

$$\theta^{(i+1)} = \arg\max_\theta \sum_t \sum_i \sum_j w_{ij}^t \log p(\mathbf{s}_i|\mathbf{x}_t,\ell_j) + \log p(\mathcal{Y}|\theta) + \log p(\mathcal{I}|\theta) \tag{29}$$

References

[1] J. A. Hesch, D. G. Kottas, S. L. Bowman, and S. I. Roumeliotis, "Consistency Analysis and Improvement of Vision-aided Inertial Navigation," *IEEE Trans. on Robotics (TRO)*, vol. 30, no. 1, pp. 158–176, 2014.

[2] D. G. Kottas and S. I. Roumeliotis, "Efficient and Consistent Vision-aided Inertial Navigation using Line Observations," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2013, pp. 1540–1547.

[3] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox, "RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments," *The International Journal of Robotics Research (IJRR)*, vol. 31, no. 5, pp. 647–663, 2012.

[4] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 32, no. 9, pp. 1627–1645, 2010.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[6] P. Agrawal, R. Girshick, and J. Malik, "Analyzing the performance of multilayer neural networks for object recognition," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 329–344.

[7] X. Liu, Y. Zhao, and S.-C. Zhu, "Single-view 3d scene parsing by attributed grammar," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 684–691.

[8] X. Chen, K. Kundu, Y. Zhu, A. Berneshawi, H. Ma, S. Fidler, and R. Urtasun, "3d object proposals for accurate object class detection," in *NIPS*, 2015.

[9] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: part i," *Robotics Automation Magazine, IEEE*, vol. 13, no. 2, pp. 99–110, June 2006.

[10] F. Lu and E. Milios, "Globally Consistent Range Scan Alignment for Environment Mapping," *Auton. Robots*, vol. 4, no. 4, pp. 333–349, 1997.

[11] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "g2o: A General Framework for Graph Optimization," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2011, pp. 3607–3613.

[12] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. Leonard, and F. Dellaert, "iSAM2: Incremental Smoothing and Mapping Using the Bayes Tree," *The International Journal of Robotics Research (IJRR)*, vol. 31, no. 2, pp. 216–235, 2012.

[13] C. Galindo, A. Saffiotti, S. Coradeschi, P. Buschka, J. Fernandez-Madrigal, and J. Gonzalez, "Multi-hierarchical Semantic Maps for Mobile Robotics," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2005, pp. 2278–2283.

[14] J. Civera, D. Galvez-Lopez, L. Riazuelo, J. Tardos, and J. Montiel, "Towards Semantic SLAM Using a Monocular Camera," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2011, pp. 1277–1284.

[15] A. Pronobis, "Semantic Mapping with Mobile Robots," dissertation, KTH Royal Institute of Technology, 2011.

[16] J. Stückler, B. Waldvogel, H. Schulz, and S. Behnke, "Dense real-time mapping of object-class semantics from RGB-D video," *Journal of Real-Time Image Processing*, pp. 1–11, 2013.

[17] V. Vineet, O. Miksik, M. Lidegaard, M. Nießner, S. Golodetz, V. A. Prisacariu, O. Kähler, D. W. Murray, S. Izadi, P. Perez, and P. H. S. Torr, "Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2015.

[18] B. Leibe, N. Cornelis, K. Cornelis, and L. Van Gool, "Dynamic 3d scene analysis from a moving vehicle," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2007, pp. 1–8.

[19] S. Pillai and J. Leonard, "Monocular slam supported object recognition," in *Proceedings of Robotics: Science and Systems (RSS)*, Rome, Italy, July 2015.

[20] N. Atanasov, M. Zhu, K. Daniilidis, and G. Pappas, "Semantic Localization Via the Matrix Permanent," in *Robotics: Science and Systems (RSS)*, 2014.

[21] S. Bao and S. Savarese, "Semantic Structure from Motion," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 2025–2032.

[22] R. Salas-Moreno, R. Newcombe, H. Strasdat, P. Kelly, and A. Davison, "SLAM++: Simultaneous Localisation and Mapping at the Level of Objects," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 1352–1359.

[23] D. Gálvez-López, M. Salas, J. Tardós, and J. Montiel, "Real-time Monocular Object SLAM," *arXiv:1504.02398*, 2015.

[24] I. Reid, "Towards Semantic Visual SLAM," in *Int. Conf. on Control Automation Robotics Vision (ICARCV)*, 2014.

[25] A. Kundu, Y. Li, F. Dellaert, F. Li, and J. Rehg, "Joint semantic segmentation and 3d reconstruction from monocular video," in *Computer Vision ECCV 2014*, ser. Lecture Notes in Computer Science, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Springer International Publishing, 2014, vol. 8694, pp. 703–718. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-10599-4_45

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

[27] J. Neira and J. Tardós, "Data Association in Stochastic Mapping Using the Joint Compatibility Test," *IEEE Trans. on Robotics and Automation (TRO)*, vol. 17, no. 6, pp. 890–897, 2001.

[28] J. Munkres, "Algorithms for the Assignment and Transportation Problems," *Journal of the Society for Industrial & Applied Mathematics (SIAM)*, vol. 5, no. 1, pp. 32–38, 1957.

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

[30] A. Mourikis and S. I. Roumeliotis, "A multi-state constraint kalman filter for vision-aided inertial navigation," in *Robotics and Automation, 2007 IEEE International Conference on*. IEEE, 2007, pp. 3565–3572.

[31] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart, "Robust visual inertial odometry using a direct ekf-based approach," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2015, pp. 298–304.

[32] C. Forster, M. Pizzoli, and D. Scaramuzza, "Svo: Fast semi-direct monocular visual odometry," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2014, pp. 15–22.

[33] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *Int. Conf. on Computer Vision*, 2011, pp. 2564–2571.

[34] S. Gidaris and N. Komodakis, "Object detection via a multi-region and semantic segmentation-aware cnn model," in *IEEE Int. Conf. on Computer Vision*, 2015, pp. 1134–1142.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[36] Z. Cai, Q. Fan, R. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *European Conference on Computer Vision (ECCV)*, 2016.

[37] M. Zhu, N. Atanasov, G. Pappas, and K. Daniilidis, "Active Deformable Part Models Inference," in *European Conference on Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science. Springer, 2014, vol. 8695, pp. 281–296.

[38] C. Dubout and F. Fleuret, "Deformable part models with individual part scaling," in *British Machine Vision Conference*, no. EPFL-CONF-192393, 2013.

[39] F. Dellaert, "Factor graphs and gtsam: A hands-on introduction," GT RIM, Tech. Rep. GT-RIM-CP&R-2012-002, Sept 2012. [Online]. Available: https://research.cc.gatech.edu/borg/sites/edu.borg/files/downloads/gtsam.pdf

[40] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "Imu preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation," in *Proceedings of Robotics: Science and Systems*, Rome, Italy, July 2015.

[41] D. G. Kottas, K. Wu, and S. I. Roumeliotis, "Detecting and dealing with hovering maneuvers in vision-aided inertial navigation systems," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Nov 2013, pp. 3172–3179.

[42] J. Nikolic, J. Rehder, M. Burri, P. Gohl, S. Leutenegger, P. T. Furgale, and R. Siegwart, "A synchronized visual-inertial sensor system with fpga pre-processing for accurate real-time slam," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, 2014, pp. 431–437.

[43] 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.

[44] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: a versatile and accurate monocular SLAM system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[45] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras," *CoRR*, vol. abs/1610.06475, 2016. [Online]. Available: http://arxiv.org/abs/1610.06475

[46] A. Geiger, J. Ziegler, and C. Stiller, "StereoScan: Dense 3d Reconstruction in Real-time," in *Intelligent Vehicles Symposium (IV)*, 2011, pp. 963–968.