A Unifying View of Geometry, Semantics, and Data Association in SLAM*

Nikolay Atanasov[1], **Sean L. Bowman**[2], **Kostas Daniilidis**[2] and **George J. Pappas**[2]
[1] Electrical and Computer Engineering, University of California, San Diego, CA 92093, USA
[2] GRASP Laboratory, University of Pennsylvania, PA 19104, USA
natanasov@ucsd.edu, seanbow@seas.upenn.edu, kostas@cis.upenn.edu, pappasg@seas.upenn.edu

Abstract

Traditional approaches for simultaneous localization and mapping (SLAM) rely on geometric features such as points, lines, and planes to infer the environment structure. They make hard decisions about the (data) association between observed features and mapped landmarks to update the environment model. This paper makes two contributions to the state of the art in SLAM. First, it generalizes the purely geometric model by introducing semantically meaningful objects, represented as structured models of mid-level part features. Second, instead of making hard, potentially wrong associations between semantic features and objects, it shows that SLAM inference can be performed efficiently with probabilistic data association. The approach not only allows building meaningful maps (containing doors, chairs, cars, etc.) but also offers significant advantages in ambiguous environments.

1 Introduction

This paper bridges the gap between the advances in SLAM, relying on purely geometric information, and those in visual recognition, relying on purely semantic information. We present an abridged version of [Bowman *et al.*, 2017] but also state key results leading to a unified view of geometry, semantics, and data association in SLAM. SLAM is the problem of estimating the motion of a sensor system, while simultaneously building a map of the environment, using a dense occupancy representation or a sparse set of *landmarks* (*e.g.*, corner, edge, or plane features). A comprehensive survey can be found in [Cadena *et al.*, 2016]. The problem is closely related to structure from motion (SfM) [Ma *et al.*, 2012] with a moot distinction that SfM is typically an offline process relying on camera measurements, while SLAM is an online operation using heterogeneous sensors found on mobile robots, including inertial measurement unit (IMU), camera, and LIDAR. The problem is also related to visual-inertial odometry (VIO) [Mourikis and Roumeliotis, 2006], which uses geometric features to infer the sensor's motion but does not construct a global map. SLAM approaches went through several transformations, starting with particle and Kalman filtering [Thrun *et al.*, 2005] and converging to iterative nonlinear optimization over the whole sensor trajectory [Strasdat *et al.*, 2010]. Current approaches model the problem as a (factor) graph of sensor and landmark states connected via edges capturing measurement constraints [Kümmerle *et al.*, 2011; Kaess *et al.*, 2012]. They can track visual-inertial systems over long trajectories in real time [Bloesch *et al.*, 2015; Forster *et al.*, 2016; Qin *et al.*, 2017], while recovering the sparse [Engel *et al.*, 2014; Mur-Artal and Tardós, 2017] or dense [Whelan *et al.*, 2016; Hornung *et al.*, 2013] environment structure. Surprisingly, most methods rely only on geometry and occurred in isolation from the impressive results in object recognition based on structured models [Felzenszwalb *et al.*, 2010] and deep neural networks [Krizhevsky *et al.*, 2012; Girshick *et al.*, 2014; Simonyan and Zisserman, 2014; He *et al.*, 2016]. Our goal is to provide a meaningful and efficient environment representation in SLAM by taking advantage of object recognition methods. A recent survey on semantic mapping can be found in [Kostavelis and Gasteratos, 2015]. The works closest to ours [Bao and Savarese, 2011; Salas-Moreno *et al.*, 2013; Gálvez-López *et al.*, 2016; Murthy *et al.*, 2017] overlook the *data association* problem (matching observations to correct landmarks in the map) which is critical for reconstructing a global object map with accurate loop closures (recognizing already visited locations). Data association is challenging

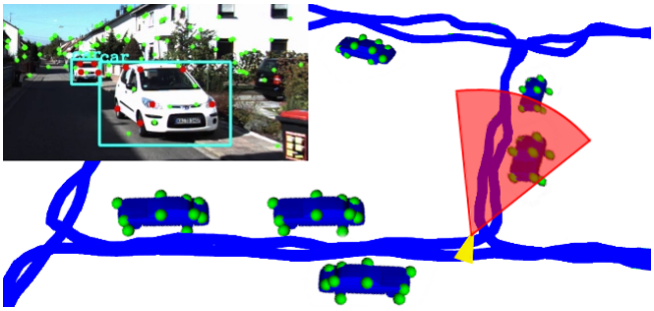

Figure 1: This paper proposes a SLAM approach that uses inertial, geometric, and semantic measurements (top left) to reconstruct the sensor trajectory (blue), detect objects of interest (cars), and estimate the positions of their parts (green points, *e.g.*, doors, wheels).

*Supported by ARL DCIST CRA W911NF-17-2-0181.

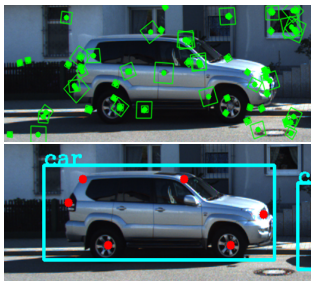

Figure 2: Geometric (ORB) features (green points) and semantic (car) features (red points).

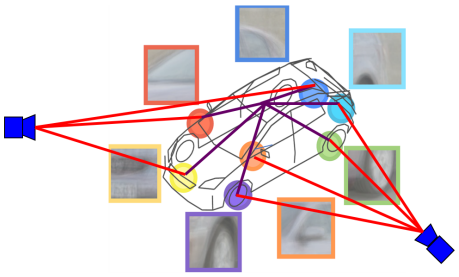

Figure 3: Structured object model with prior shape constraints (purple lines) and online camera constraints (red lines).

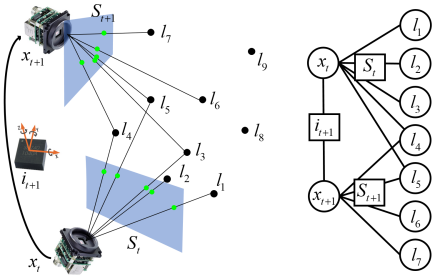

Figure 4: Factor graph (right) of two sensor states x_t, x_{t+1} constrained by IMU measurement i_{t+1} and semantic observations S_t, S_{t+1} (green) of landmarks l_i.

when many areas look alike (perceptual aliasing) and a single wrong choice can drastically affect the estimation accuracy. We make the following **contributions**.

- We generalize the geometric environment model of SLAM by introducing structured object models of mid-level semantic landmarks (Fig. 1).
- Instead of making hard, potentially wrong associations among observations and landmarks, we show that SLAM inference can be done efficiently using a probability distribution over all possible associations, leading to significant advantages in ambiguous environments.

2 Metric-Semantic SLAM

2.1 Representation

Consider a sensor system consisting of an IMU and monocular camera that are time synchronized. Choose a subset of the camera images as *keyframes* (*e.g.*, by selecting every nth frame) and let x_t be the system state (6-D pose, velocity, and IMU biases) corresponding to the tth keyframe. Let $\mathcal{X} := \{x_t\}_{t=0}^{T}$ be the sensor state trajectory and let the environment be represented as a collection of landmarks $\mathcal{L} := \{l_m\}_{m=1}^{M}$ with positions $l_m \in \mathbb{R}^3$. We distinguish between *geometric landmarks* λ_j estimated from *geometric features* Y_t (*e.g.*, Harris corners, SIFT, ORB, etc.) and *semantic landmarks* l_m estimated from *semantic features* S_t (parts of detected objects) extracted from each keyframe image as shown in Fig. 2.

Geometric Structure. The IMU provides a set I_{t+1} of linear acceleration and rotational velocity measurements between keyframes t and $t+1$. These measurements are integrated using the approach of [Forster *et al.*, 2015] to obtain a single constraint that relates x_t and x_{t+1} via an IMU measurement model f with noise covariance U:

$$i_{t+1} = f(x_{t+1}, x_t) + u_t, \quad u_t \sim \mathcal{N}(0, U). \tag{1}$$

The model in (1) relates two system states and specifies a Gaussian density function $p(i_{t+1}|x_t, x_{t+1})$ for the IMU measurement i_{t+1}. Similarly, each geometric feature $y_{tk} \in Y_t$ for $k = 1, \ldots, K_t$ extracted at time t relates the position $\lambda_j \in \mathbb{R}^3$ of a corresponding geometric landmark to the system state x_t via a camera projection model g [Ma *et al.*, 2012, Ch.3]:

$$y_{tk} = g(x_t, \lambda_j) + v_t, \quad v_t \sim \mathcal{N}(0, V) \tag{2}$$

To recover the landmark positions $\{\lambda_j\}_{j=1}^{J}$, it is necessary to know the data association $\pi_t(k) = j$ of each feature y_{tk} to the correct landmark λ_j. Estimating $\pi_t : \{1, \ldots, K_t\} \to \{1, \ldots, J\}$ can done by tracking the features y_{tk} in consecutive keyframes. For example, we extract ORB features and track them by minimizing the ORB descriptor distance between consecutive frames [Rublee *et al.*, 2011]. As discussed in Sec. 2.2, this visual odometry allows us to infer the sensor trajectory $\mathcal{X}$ locally but is not sufficient for loop closure and estimation of the global landmark positions $\{\lambda_j\}$.

Semantic Structure. Instead of using geometric landmarks to model the environment, **our first innovation** is to extract semantic features (object parts) and represent the environment as a set of objects, with inferred positions, orientations, and classes (Fig. 1). We assume a known list of object classes (with prior training data) and detect bounding boxes in each keyframe via a real-time detector such as DPM [Dubout and Fleuret, 2013], RCNN [Ren *et al.*, 2015], or YOLO [Redmon and Farhadi, 2016]. Within each bounding box, we extract semantic features specific to the object class via the stacked hourglass convolutional network of [Pavlakos *et al.*, 2017]. Let S_t be the set of all $|S_t| = N_t$ semantic features extracted from the tth keyframe. Similar to the geometric case, each semantic feature $s_{tn} \in S_t$ is related to the position $l_m \in \mathbb{R}^3$ of a semantic landmark via the camera perspective projection:

$$s_{tn} = g(x_t, l_m) + w_t, \quad w_t \sim \mathcal{N}(0, W) \tag{3}$$

and an *unknown* function $d_t : \{1, \ldots, N_t\} \to \{1, \ldots, M\}$ stipulating the data associations among S_t and $\mathcal{L}$. Unlike in the geometric case, we cannot simply track the semantic features because we are interested in recovering their global positions $\mathcal{L}$, which requires loop closure (recognizing previously observed landmarks instead of instantiating new ones). We focus on mapping semantic landmarks because they are fewer (leading to efficient performance) and yet more robust to viewpoint changes than geometric ones. This allows us to perform accurate visual-inertial odometry relying on both geometric and semantic features locally and efficient global loop closure over the map of semantic landmarks (Fig. 1).

To obtain the poses of the objects associated with the semantic landmarks, we use a representation similar to the deformable parts model [Felzenszwalb *et al.*, 2010]. We introduce structure constraints among an object's landmarks, given by a root location and displacements of the object parts with respect to it (Fig. 3). For example, the front-left wheel of a car should not deviate significantly from an expected posi-

tion relative to its front-left door. The structure constraint c_{ij} between the positions l_i, l_j of two semantic parts belonging to the same object class are captured by a function h, while the part displacement costs are captured by a Gaussian model:

$$c_{ij} = h(l_i, l_j) + \eta, \quad \eta \sim \mathcal{N}(0, E). \tag{4}$$

Given its semantic part positions, an object's orientation can be recovered via the Kabsch algorithm [Kabsch, 1978].

Problem. Determine the sensor trajectory $\hat{\mathcal{X}}$ and semantic landmark positions $\hat{\mathcal{L}}$ that maximize the likelihood of the inertial, geometric, and semantic measurements $\mathcal{Z} := \{(I_t, Y_t, S_t)\}_{t=0}^{T}$ under the nonlinear Gaussian models (1)-(4)

$$\hat{\mathcal{X}}, \hat{\mathcal{L}} = \underset{\mathcal{X},\mathcal{L}}{\arg\max}\ \log p(\mathcal{Z} \mid \mathcal{X}, \mathcal{L}) \tag{5}$$

2.2 Inference

A *factor graph* is a probabilistic graphical model [Koller and Friedman, 2009] that allows representing an optimization problem such as (5) as a graph of random variables (nodes) with constraints among them (factors). In our case, the nodes are Gaussian variables corresponding to $\mathcal{X}$ and $\mathcal{L}$. The factors are constraints imposed by the inertial I_t, geometric Y_t, and semantic S_t measurements via the models (1)-(4). Fig. 4 shows an example graph of two sensor states related by inertial and semantic measurements. Inference over the graph can be formulated as a maximization (5) of the likelihood of all measurements with respect to the latent variables $\mathcal{X}$, $\mathcal{L}$:

$$\begin{aligned}\hat{\mathcal{X}}, \hat{\mathcal{L}} &= \underset{\mathcal{X},\mathcal{L}}{\arg\max}\ \log \prod_{t=0}^{T} p(I_t, Y_t, S_t \mid \mathcal{X}, \mathcal{L}) \\ &= \underset{\mathcal{X},\mathcal{L}}{\arg\min}\ \sum_{t=1}^{T} \|i_t - f(x_t, x_{t-1})\|_U^2 \\ &\quad + \sum_{t=0}^{T}\sum_{k=1}^{K_t}\sum_{j=1}^{J} \mathbb{1}_{\{\pi_t(k)=j\}} \|y_{tk} - g(x_t, \lambda_j)\|_V^2 \\ &\quad + \sum_{t=0}^{T}\sum_{n=1}^{N_t}\sum_{m=1}^{M} \mathbb{1}_{\{d_t(n)=m\}} \|s_{tn} - g(x_t, l_m)\|_W^2\end{aligned} \tag{6}$$

where we have skipped prior and structure constraints (4) to save and have used $\|e\|_\Sigma^2 := e^T \Sigma^{-1} e$. The second equality utilizes that the models (1)-(4) are Gaussian; *e.g.*, $p(i_{t+1} \mid x_{t+1}, x_t) \propto \exp\left\{-\frac{1}{2}\|i_{t+1} - f(x_{t+1}, x_t)\|_U^2\right\}$. We also emphasized the dependence on the *unknown* data associations π_t, d_t via the indicator functions. Assuming for now that π_t, d_t are known, (6) is a nonlinear least-squares problem, which can be solved efficiently via incremental linearization (e.g., Gauss-Newton algorithm). Given initial estimates $\mathcal{X}^0$, $\mathcal{L}^0$, *e.g.*, obtained from IMU integration and feature triangulation, the functions f and g are linearized to obtain the Jacobians F^x, F^y and G^x, G^y with respect to each input, leading to:

$$\begin{aligned}&\min_{\delta\mathcal{X},\delta\mathcal{L}} \sum_{t=1}^{T} \|i_t - f(x_t^0, x_{t-1}^0) - F_{t,t-1}^x \delta x_t - F_{t,t-1}^y \delta x_{t-1}\|_U^2 + \\ &\sum_{t=0}^{T}\sum_{k=1}^{K_t}\sum_{j=1}^{J} \mathbb{1}_{\{\pi_t(k)=j\}} \|y_{tk} - g(x_t^0, \lambda_j^0) - G_{t,j}^x \delta x_t - G_{t,j}^y \delta\lambda_j\|_V^2 \\ &\sum_{t=0}^{T}\sum_{n=1}^{N_t}\sum_{m=1}^{M} \mathbb{1}_{\{d_t(n)=m\}} \|s_{tn} - g(x_t^0, l_m^0) - G_{t,m}^x \delta x_t - G_{t,m}^y \delta l_m\|_W^2\end{aligned}$$

To avoid estimating the geometric landmark positions $\{\lambda_j\}$, a key idea used in *structureless* VIO approaches [Mourikis and Roumeliotis, 2006] is to multiply the second term above by a unitary matrix whose columns form the the basis of the left nullspace of $G_{t,j}^y$, making the term $G_{t,j}^y \delta\lambda_j$ disappear. Collecting the Jacobians into one large but *sparse* matrix A and the error vectors $i_t - f(x_t^0, x_{t-1}^0)$, $y_{tk} - g(x_t^0, \lambda_j^0)$, and $s_{tn} - g(x_t^0, l_m^0)$ into one vector b, leads to a standard least-squares problem $\min_\theta \|A\theta - b\|^2$ where $\theta := (\delta\mathcal{X}, \delta\mathcal{L})$. Even though we are estimating the whole history $\mathcal{X}$ of sensor states, since the models (1)-(4) depend only on pairs of sensor-landmark states, A remains sparse and the problem can be solved efficiently via QR factorization. The optimization can be performed incrementally as new measurements arrive by updating (rather than recomputing) the QR factorization [Kaess *et al.*, 2008; 2012].

2.3 Data Association

We return to the data association problem. Since we do not aim to recover $\{\lambda_j\}$, it is sufficient to obtain the geometric associations π_t via feature tracking as mentioned in Sec. 2.1. It is, however, necessary to estimate the semantic associations $\mathcal{D} := \{d_t\}_{t=0}^{T}$, if we are to perform loop closure and reconstruct the semantic landmark positions $\mathcal{L}$. We revisit the max likelihood estimation in (5), emphasizing that the semantic associations $\mathcal{D}$ are unknown and should also be inferred:

$$\hat{\mathcal{X}}, \hat{\mathcal{L}}, \hat{\mathcal{D}} = \underset{\mathcal{X},\mathcal{L},\mathcal{D}}{\arg\max} \log p(\mathcal{Z} \mid \mathcal{X}, \mathcal{L}, \mathcal{D}) \tag{7}$$

Existing work decomposes this joint SLAM and data association problem into two subproblems. First, given prior estimates $\mathcal{X}^0, \mathcal{L}^0$, the max likelihood estimate $\hat{\mathcal{D}}$ of the associations is computed, *e.g.*, via JCBB [Neira and Tardós, 2001] or the Hungarian algorithm [Munkres, 1957]. Then, given $\hat{\mathcal{D}}$, the most likely landmark and sensor states are estimated. This is a special case of *coordinate descent*:

$$\begin{aligned}\hat{\mathcal{D}}^{i+1} &= \underset{\mathcal{D}}{\arg\max}\ p(\mathcal{D} \mid \mathcal{X}^i, \mathcal{L}^i, \mathcal{Z}) \\ \mathcal{X}^{i+1}, \mathcal{L}^{i+1} &= \underset{\mathcal{X},\mathcal{L}}{\arg\max}\ \log p(\mathcal{Z} \mid \mathcal{X}, \mathcal{L}, \mathcal{D}^{i+1})\end{aligned} \tag{8}$$

which allows revisiting association decisions once state estimates improve but does little to resolve the problem with ambiguous measurements since a hard decision on data associations is required and may have a highly detrimental effect on the estimation performance. Rather than simply selecting $\mathcal{D}$ as the mode of $p(\mathcal{D} \mid \mathcal{X}, \mathcal{L}, \mathcal{Z})$, **our second innovation** is to consider the entire distribution of $\mathcal{D}$ when estimating $\mathcal{X}$ and $\mathcal{L}$. Since $\mathcal{D}$ is a latent variable whose realization we are not interested in inferring but is needed to estimate $\mathcal{X}$ and $\mathcal{L}$, probability theory suggests that we should marginalize $\mathcal{D}$. Our main result is that *expectation maximization* (EM) allows us to maximize the expected (w.r.t. $\mathcal{D}$) measurement likelihood, utilizing the *whole distribution* of $\mathcal{D}$:

$$\mathcal{X}^{i+1}, \mathcal{L}^{i+1} = \underset{\mathcal{X},\mathcal{L}}{\arg\max} \sum_{\mathcal{D}\in\mathbb{D}} p(\mathcal{D} \mid \mathcal{X}^i, \mathcal{L}^i, \mathcal{Z}) \log p(\mathcal{Z} \mid \mathcal{X}, \mathcal{L}, \mathcal{D}) \tag{9}$$

To compare this with coordinate descent (8), we drop the inertial and geometric terms in (6) since their data associations

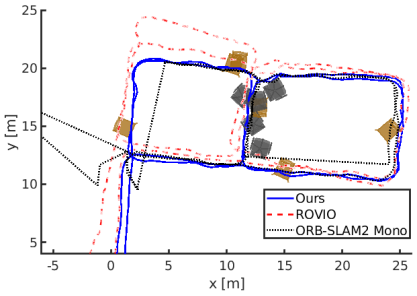

Figure 5: Estimated sensor trajectory and poses of swivel and four-legged chairs in first office experiment.

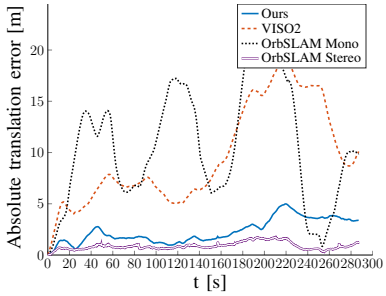

Figure 6: Norm of the position error between the estimated and ground truth vehicle trajectories on KITTI sequence 05.

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

Figure 7: Mean translational and rotational error over subpath lengths (100, 200, . . . , 800) meters on the KITTI odometry dataset.

are known or estimated via feature tracking and re-write (9):

$$\max_{\mathcal{X},\mathcal{L}} \sum_{t=0}^{T} \sum_{n=1}^{N_t} \sum_{m=1}^{M+N_t} \underbrace{\sum_{d_t \in \mathbb{D}_t(n,m)} p(d_t \mid \mathcal{X}^i, \mathcal{L}^i, S_t)}_{=: w_{n,m}^{t,i}} \log p(s_{tn} \mid x_t, l_m) \quad (10)$$

Above $\mathbb{D}_t(n,m) := \{d_t \mid d_t(n) = m\}$ is the set of possible data associations at time t that assign semantic feature s_{tn} to landmark l_m and $w_{n,m}^{t,i}$ quantifies the probability that s_{tn} was produced by landmark l_m via (3). Note that (6) and (8) have a similar form to (10), except that for each n there is exactly one m such that $w_{n,m}^{t,i} = 1$ and $w_{n,m'}^{t,i} = 0$ for all $m' \neq m$. The EM formulation has the advantage that no hard decisions are required since it "averages" over all possible data associations. The following proposition shows that, besides generalizing coordinate descent, EM can be performed efficiently via a connection between the data association distribution $w_{n,m}^{t,i}$ and the permanent[1] of a suitable matrix.

Proposition. *Optimization* (10) *can be performed by iterating:*

- ***E step**: given $\mathcal{X}^i, \mathcal{L}^i$ compute the associations distribution:*

$$\omega_{n,m}^{t,i} \propto Q_{nm}^t \mathbf{per}\, Q_{-nm}^t$$

where Q^t is a matrix of data association probabilities[2] with elements Q_{nm}^t and Q_{-nm}^t is the matrix with the nth row and mth column removed.

- ***M step**: given the data association distribution $\omega_{n,m}^{t,i}$, update the sensor and landmark state estimates:*

$$\mathcal{X}^{i+1}, \mathcal{L}^{i+1} = \arg\max_{\mathcal{X},\mathcal{L}} \sum_{t=0}^{T} \sum_{n=1}^{N_t} \sum_{m=1}^{M+N_t} w_{n,m}^{t,i} \log p(s_{tn} \mid x_t, l_m)$$

This result allows us to take advantage of matrix permanent approximation algorithms [Jerrum *et al.*, 2004] to efficiently summarize the combinatorial data association space in polynomial time. Instead of max likelihood data association as in (8), we estimate a data association distribution via the weights $w_{nm}^{t,i}$ ("E" step) and then maximize the expected measurement log likelihood over it ("M" step).

3 Evaluation

Our algorithm uses GTSAM and its iSAM2 implementation [Kaess *et al.*, 2012] to perform the M step. Our front-end selects every 15th camera frame as a keyframe and tracks geometric ORB features. Outlier tracks that do not fit the essential matrix constraint between consecutive views are eliminated via RANSAC. The essential matrix is estimated using relative orientation from gyro measurements and two-point correspondences. Objects are detected via the deformable parts model [Dubout and Fleuret, 2013] and semantic features are extracted via a stacked hourglass convolutional network [Pavlakos *et al.*, 2017]. Experiments were performed indoors, building a map of doors, swivel chairs and four-legged chairs, and outdoors, building a map of cars on the KITTI odometry dataset [Geiger *et al.*, 2012]. The indoor experiments included loops around a room equipped with motion tracking, a medium trajectory (175 m) spanning one floor, and a long trajectory (625 m) spanning two floors. The estimated sensor trajectory and chair poses in the medium sequence are shown in Fig. 5. The performance is compared qualitatively to ROVIO [Bloesch *et al.*, 2015] and ORB-SLAM2 [Mur-Artal and Tardós, 2017]. Since ORB-SLAM2 uses only geometric features and no inertial information, it loses tracking for long durations in feature-deprived areas but is able to recover when entering previously mapped regions. In the long indoor sequence due to the repetitive nature of the hallways, the bag-of-words loop-closure approach of ORB-SLAM2 also makes incorrect associations leading to unsuccessful tracking. The outdoor experiments used visual odometry from VISO2 [Geiger *et al.*, 2011] instead of inertial odometry. The translation and rotation errors of our algorithm, VISO2, and ORB-SLAM2 with respect to the ground truth vehicle trajectory are compared in Fig. 6 and Fig. 7.

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

[2] $Q^t := [G^t \; H^t]$, where $G^t \in \mathbb{R}^{N_t \times M}$ is a matrix of scaled likelihoods $G_{nm}^t = \frac{p_{tm}}{\lambda(1-p_{tn})} p(s_{tn} \mid x_t, l_m)$, p_{tm} is the probability that l_m is detected from x_t according to a Poisson process with mean λ, and $H^t = \mathbf{diag}\left([p_{new}(s_{t1}) \; \cdots \; p_{new}(s_{tN_t})]\right) \in \mathbb{R}^{N_t \times N_t}$ is a matrix of probabilities that new landmarks should be created. See [Atanasov *et al.*, 2015] for details.

References

[Atanasov *et al.*, 2015] N. Atanasov, M. Zhu, K. Daniilidis, and G. Pappas. Localization from semantic observations via the matrix permanent. *IJRR*, 35:73–99, 2015.

[Bao and Savarese, 2011] S. Bao and S. Savarese. Semantic Structure from Motion. In *IEEE CVPR*, pages 2025–2032, 2011.

[Bloesch *et al.*, 2015] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart. Robust visual inertial odometry using a direct EKF-based approach. In *IEEE/RSJ IROS*, pages 298–304, 2015.

[Bowman *et al.*, 2017] S. Bowman, N. Atanasov, K. Daniilidis, and G. Pappas. Probabilistic Data Association for Semantic SLAM. In *IEEE ICRA*, 2017. (**Best Conference Paper Award**).

[Cadena *et al.*, 2016] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. Leonard. Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age. *IEEE T-RO*, 32(6):1309–1332, 2016.

[Dubout and Fleuret, 2013] C. Dubout and F. Fleuret. Deformable part models with individual part scaling. In *BMVC*, 2013.

[Engel *et al.*, 2014] J. Engel, T. Schöps, and D. Cremers. LSD-SLAM: Large-scale direct monocular SLAM. In *ECCV*, 2014.

[Felzenszwalb *et al.*, 2010] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Trans. on PAMI*, 32(9):1627–1645, 2010.

[Forster *et al.*, 2015] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza. IMU Preintegration on Manifold for Efficient Visual-Inertial Maximum-a-Posteriori Estimation. In *Robotics: Science and Systems*, 2015.

[Forster *et al.*, 2016] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza. SVO: Semi-Direct Visual Odometry for Monocular and Multi-Camera Systems. *IEEE T-RO*, 2016.

[Gálvez-López *et al.*, 2016] D. Gálvez-López, M. Salas, J. Tardós, and J. Montiel. Real-time monocular object SLAM. *Robotics and Autonomous Systems*, 75:435–449, 2016.

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

[Geiger *et al.*, 2012] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *IEEE CVPR*, 2012.

[Girshick *et al.*, 2014] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE CVPR*, 2014.

[He *et al.*, 2016] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE CVPR*, 2016.

[Hornung *et al.*, 2013] A. Hornung, K. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard. OctoMap: An efficient probabilistic 3D mapping framework based on octrees. *Autonomous Robots*, 34(3):189–206, 2013.

[Jerrum *et al.*, 2004] M. Jerrum, A. Sinclair, and E. Vigoda. A Polynomial-time Approximation Algorithm for the Permanent of a Matrix with Nonnegative Entries. *Journal of the ACM*, 51(4):671–697, 2004.

[Kabsch, 1978] W. Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 34(5):827–828, 1978.

[Kaess *et al.*, 2008] M. Kaess, A. Ranganathan, and F. Dellaert. iSAM: Incremental Smoothing and Mapping. *IEEE T-RO*, 24(6):1365–1378, 2008.

[Kaess *et al.*, 2012] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. Leonard, and F. Dellaert. iSAM2: Incremental Smoothing and Mapping Using the Bayes Tree. *IJRR*, 31(2):216–235, 2012.

[Koller and Friedman, 2009] D. Koller and N. Friedman. *Probabilistic Graphical Models*. MIT press, 2009.

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

[Krizhevsky *et al.*, 2012] A. Krizhevsky, I. Sutskever, and Geoffrey H. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*, pages 1097–1105. 2012.

[Kümmerle *et al.*, 2011] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard. g2o: A General Framework for Graph Optimization. In *IEEE ICRA*, pages 3607–3613, 2011.

[Ma *et al.*, 2012] Y. Ma, S. Soatto, J. Kosecka, and S. Sastry. *An Invitation to 3-D Vision: from Images to Geometric Models*, volume 26. Springer Science & Business Media, 2012.

[Mourikis and Roumeliotis, 2006] A. Mourikis and S. Roumeliotis. A Multi-State Constraint Kalman Filter for Vision-aided Inertial Navigation. Technical report, University of Minnesota, 2006.

[Munkres, 1957] J. Munkres. Algorithms for the Assignment and Transportation Problems. *Journal of SIAM*, 5(1):32–38, 1957.

[Mur-Artal and Tardós, 2017] R. Mur-Artal and J. Tardós. ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras. *IEEE T-RO*, 33(5):1255–1262, 2017.

[Murthy *et al.*, 2017] J. Murthy, S. Sharma, and K. Krishna. Shape priors for real-time monocular object localization in dynamic environments. In *IEEE/RSJ IROS*, pages 1768–1774, 2017.

[Neira and Tardós, 2001] J. Neira and J. Tardós. Data Association in Stochastic Mapping Using the Joint Compatibility Test. *IEEE T-RO*, 17(6):890–897, 2001.

[Pavlakos *et al.*, 2017] G. Pavlakos, X. Zhou, A. Chan, K. G. Derpanis, and K. Daniilidis. 6-DoF object pose from semantic keypoints. In *IEEE ICRA*, pages 2011–2018, 2017.

[Qin *et al.*, 2017] T. Qin, P. Li, and S. Shen. VINS-Mono: A robust and versatile monocular visual-inertial state estimator. In *arXiv preprint:1708.03852*, 2017.

[Redmon and Farhadi, 2016] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. *arXiv preprint:1612.08242*, 2016.

[Ren *et al.*, 2015] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*, pages 91–99. 2015.

[Rublee *et al.*, 2011] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. ORB: An efficient alternative to SIFT or SURF. In *ICCV*, pages 2564–2571, 2011.

[Salas-Moreno *et al.*, 2013] R. Salas-Moreno, R. Newcombe, H. Strasdat, P. Kelly, and A. Davison. SLAM++: Simultaneous Localisation and Mapping at the Level of Objects. In *IEEE CVPR*, 2013.

[Simonyan and Zisserman, 2014] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556*, 2014.

[Strasdat *et al.*, 2010] H. Strasdat, J. Montiel, and A. Davison. Real-time monocular SLAM: Why filter? In *IEEE ICRA*, 2010.

[Thrun *et al.*, 2005] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. MIT Press Cambridge, 2005.

[Whelan *et al.*, 2016] T. Whelan, R. Salas-Moreno, B., A. Davison, and S. Leutenegger. ElasticFusion: Real-Time Dense SLAM and Light Source Estimation. *IJRR*, 35(14):1697–1716, 2016.