Robust and Efficient Semantic SLAM with Semantic Keypoints

Sean L. Bowman*, **Kostas Daniilidis, George J. Pappas**
University of Pennsylvania
472 Levine Hall
3330 Walnut Street
Philadelphia, PA, USA, 19104

Abstract

Although SLAM has traditionally relied on representing the environment as low-level geometric features such as points, lines, and planes, recent advances in object recognition capabilities as well as demand for higher-level environment representations to facilitate higher-level autonomy have motivated object based environment representations, or Semantic SLAM. We present a Semantic SLAM algorithm that directly incorporates a sparse and deformable representation of objects into a factor graph SLAM system, resulting in a system that is efficient, robust to varying object shapes and environments, and easy to incorporate into an existing SLAM system. This keypoint-based representation facilitates robust detection in varying conditions and intraclass shape variation as well as computational efficiency. The performance of our algorithm is demonstrated in two different SLAM systems and in varying environments.

1 Introduction

This paper addresses the task of semantic simultaneous localization and mapping (SLAM), i.e. estimating the six degree of freedom pose of objects in an unknown environment while localizing the mobile robot within the map. While SLAM is an extensively studied problem, until recently most methods focused solely on creating a map of low-level geometric features in the environment such as corners [Hesch et al., 2014], lines [Kottas and Roumeliotis, 2013], and surface patches [Henry et al., 2012]. In contrast, high-level autonomy in unknown environments requires more meaningful maps of objects with *semantic* content, such as windows, tables, and chairs.

Traditional approaches to SLAM were based on Kalman filter methods in which only the most recent robot pose is estimated [Durrant-Whyte and Bailey, 2006]. This approach is computationally efficient, however the inability to modify previous pose estimates and relinearize previous measurement functions frequently causes errors to compound [Hesch et al., 2014]. More recently, great success has been seen with batch or pose-graph methods that optimize over the robot's entire trajectory rather than simply the most recent pose [Kaess et al., 2012, Mur-Artal and Tardós, 2016]. Our approach follows this formulation, viewing the estimation problem as a set of nodes in a graph (a "pose graph") where each node corresponds to an estimation variable (e.g. a robot or object pose). Two robot-pose nodes in the graph are linked by an edge if there is an odometry measurement available between them (e.g. LIDAR scan match that was performed between two subsequent

*seanbow@seas.upenn.edu

measurements), while a robot-pose node and an object-pose node share an edge if the object was observed from the corresponding robot pose.

Rapidly advancing methods of object recognition within images [Ren et al., 2015, Redmon and Farhadi, 2017] have led to many works exploring the inclusion of semantic information within these pose graph SLAM methods. Focusing on the localization problem only, [Atanasov et al., 2014] incorporated semantic observations in the metric optimization via a set-based Bayes filter. Many other approaches [Civera et al., 2011, Pronobis, 2011, Stückler et al., 2013, Vineet et al., 2015, Pillai and Leonard, 2015] extract both metric and semantic information but typically the two processes are carried out separately and the results are merged afterwards. [Lianos et al., 2018] uses semantic information to improve frame-to-frame feature matching in visual odometry systems. In [Bowman et al., 2017], the authors integrated semantic information directly into the SLAM estimation problem as we do here, however the only object variable included in the estimation was its position rather than full six degree-of-freedom pose.

Our approach directly integrates the 3D shape of semantic objects into a SLAM system by representing it as a sparse set of semantically meaningful keypoints [Pavlakos et al., 2017]. These keypoints can be reliably detected on different instances of an object class and from varying viewpoints and viewing conditions, and localizing them and including them in the map is no more computationally intensive than traditional geometric SLAM. A learned deformable object structure relates these keypoints to the full object pose, and is used to both obtain the object pose as well as better localize the keypoints themselves by constraining them to known object shapes. This approach results in a system that is efficient in both representation and computation, and can be easily integrated into existing pose graph-based SLAM systems. We demonstrate its effectiveness by applying it to two different SLAM algorithms and demonstrating its performance in varying environments.

2 Problem Formulation

In the classical simultaneous localization and mapping problem, 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. A mathematical statement of the SLAM problem then solves the following MAP estimation problem:

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

Most modern SLAM algorithms solve this by formulating it as a pose graph optimization. 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. 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

$$\hat{\mathcal{V}} = \arg\min_{\mathcal{V}} \sum_{f \in \mathcal{F}} f(\mathcal{V}). \tag{2}$$

For example, consider a simple case of a mobile ground robot equipped with wheel encoders. Along its trajectory, between each pair of poses $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$, the integrated wheel encoders report a pose difference $\mathbf{z}_i = \mathbf{x}_{i+1} - \mathbf{x}_i + w_i$, where $w_i \sim \mathcal{N}(0, \mathbf{R}_i)$ is some Gaussian noise. It is then easy to see that the solution for the estimation in (1) (assuming a uniform prior on $p(\mathcal{X}, \mathcal{L})$) is given by

$$\hat{\mathbf{x}}_{1:T} = \arg\max_{\mathbf{x}} \log p(\mathbf{z}_{1:T-1} | \mathbf{x}_{1:T}). \tag{3}$$

Assuming conditional independence of measurements given the trajectory and using the known distribution of $\mathbf{z}$, this can be written as

$$\hat{\mathbf{x}}_{1:T} = \arg\min_{\mathbf{x}} \sum_{i=1}^{T-1} \|\mathbf{z}_i - (\mathbf{x}_{i+1} - \mathbf{x}_i)\|^2_{\mathbf{R}_i} \tag{4}$$

which we see is a factor formulation as in (2) with $f(\mathbf{x}_i, \mathbf{x}_{i+1}) = \|\mathbf{z}_i - (\mathbf{x}_{i+1} - \mathbf{x}_i)\|^2_{\mathbf{R}_i}$.

More generally, suppose a robot receives several different classes of measurements $\mathcal{Z}_1, \ldots, \mathcal{Z}_N$, e.g. odometry, GPS, visual, etc. Assuming measurements are conditionally independent given the trajectory and map, and a uniform prior on $\mathcal{Z}$[1], we can write (1) as

$$\hat{\mathcal{X}}, \hat{\mathcal{L}} = \arg\max_{\mathcal{X},\mathcal{L}} \log p(\mathcal{Z}|\mathcal{X},\mathcal{L}) p(\mathcal{X},\mathcal{L}) \tag{5}$$

$$= \arg\max_{\mathcal{X},\mathcal{L}} \left[\sum_{i=1}^{N} \log p(\mathcal{Z}_i|\mathcal{X},\mathcal{L}) + \log p(\mathcal{X},\mathcal{L}) \right] \tag{6}$$

$$= \arg\min_{\mathcal{X},\mathcal{L}} \left[\sum_{i=1}^{N} - \log p(\mathcal{Z}_i|\mathcal{X},\mathcal{L}) - \log p(\mathcal{X},\mathcal{L}) \right], \tag{7}$$

and so we see that negative measurement log-likelihoods correspond exactly to the factors in (2). Additionally, we see the inherent modularity in the factor graph formulation; new information or measurement types results in only another additive term to the optimization. In the following section, we describe the specific form of our semantic representation and measurement that allow it to be included in systems of the form in (2) or (7).

3 Semantic Factors

Here, we focus on a particular formulation of the SLAM problem that incorporates semantic objects. An individual object is represented as its overall pose $o_j \in SO(3)$ along with a set of *semantic keypoints* $\ell_i \in \mathbb{R}^3$. These semantic keypoints consist of semantically meaningful points on the object that can be reliably found across different instances of the object class and meaningfully located in space. For example, the object class *car* may have among its semantic keypoints those of "front left wheel" and "rear right headlight." Using the methods of [Pavlakos et al., 2017], an object's semantic keypoints are able to be reliably detected and identified across various viewpoints. For example, in Figure 1 various semantic keypoint detections for the object classes bicycle, bus, car, and chair are shown.

An object class is represented as a deformable shape model consisting of the mean shape position of each of its p semantic keypoints relative to its own pose o, along with directions of possible shape variability to account for intraclass variation of keypoint locations. More specifically, let $\mathbf{S} \in \mathbb{R}^{3\times p}$ be a matrix consisting of an object's p keypoints represented in the object's own frame stacked horizontally. We then have

$$\mathbf{S}(c) = \mathbf{B}_0 + \sum_{i=1}^{k} c_i \mathbf{B}_i, \tag{8}$$

where $\mathbf{B}_0$ is the object class's mean shape and $\mathbf{B}_1, \ldots, \mathbf{B}_k$ are modes of possible shape variability, computed offline by Principal Component Analysis [Pavlakos et al., 2017], written as a function of the deformation coefficients $c \in \mathbb{R}^k$.

Intuitively, repeated observations of a keypoint ℓ_j are used to triangulate it in space; the deformable shape model of the known object class is then used to indirectly estimate both the deformation coefficients c and

[1]Most methods additionally assume a uniform prior on $p(\mathcal{X},\mathcal{L})$ and perform a maximum likelihood estimation, however later in section 3.1 we will use this term to capture semantic object structure

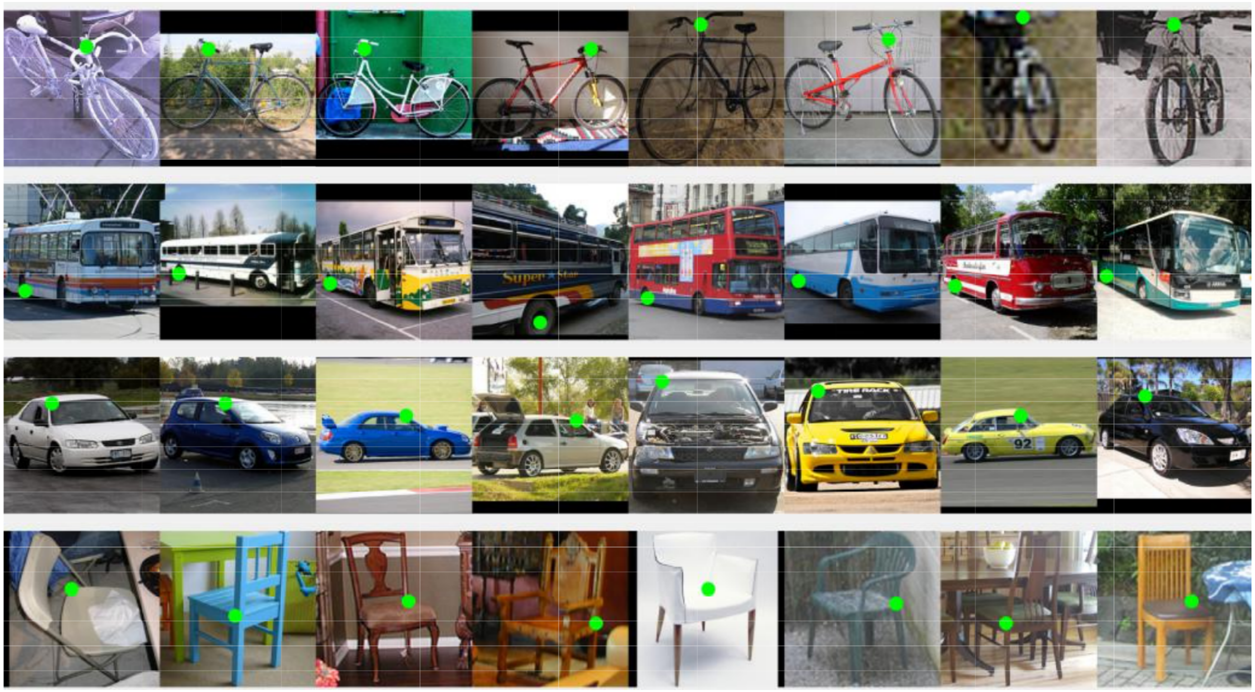

Figure 1: Detected semantic keypoints for the object classes bicycle, bus, car, and chair

the overall object pose o. See Figure 2 for an example of a car being observed from two camera poses. The semantic keypoints, denoted by colored circles and their associated image patches, are constrained in space by the corresponding image observations, denoted by red lines drawn to the camera positions. The object pose, represented by the axis in the middle of the car, is then constrained by the deformable object structure, denoted by the purple lines drawn to the keypoints.

3.1 Semantic Measurement Model

Formally, an object in the map consists of four elements: its class o^C (assumed to be known from the object detector, see [Bowman et al., 2017] for a probabilistic treatment), its pose $o \in SO(3)$, the positions of its keypoints $\ell_i \in \mathbb{R}^3$, $i = 1, \ldots, p$, and its deformation coefficients $c \in \mathbb{R}^k$.

When a camera $\mathbf{x}$ observes this object o, the measurement $h(\mathbf{x}, o)$ consists of projections of each of the object's semantic keypoints onto the image plane:

$$h(\mathbf{x}, o) = \left[h_\pi(\mathbf{x}, \ell_1)^T \quad \cdots \quad h_\pi(\mathbf{x}, \ell_p)^T\right]^T, \tag{9}$$

where $h_\pi(\mathbf{x}, \ell)$ is the standard perspective projection of a point at ℓ onto a camera at pose $\mathbf{x}$.

The probability of a semantic measurement $\mathbf{z} = [z_1^T \ \cdots \ z_p^T]^T$ is given as

$$p(\mathbf{x}, o, \ell, c|\mathbf{z}) = p(o, c|\mathbf{x}, \ell, \mathbf{z})p(\mathbf{x}, \ell|\mathbf{z}). \tag{10}$$

Note that as the actual measurement $\mathbf{z}$ observes only the semantic keypoints ℓ, we have $p(o, c|\mathbf{x}, \ell, \mathbf{z}) = p(o, c|\ell)$, and thus

$$\begin{aligned} p(\mathbf{x}, o, \ell, c|\mathbf{z}) &= p(o, c|\ell)p(\mathbf{x}, \ell|\mathbf{z}) && (11) \\ &= \frac{p(\ell|o, c)p(o, c)}{p(\ell)} \frac{p(\mathbf{z}|\mathbf{x}, \ell)p(\mathbf{x}, \ell)}{p(\mathbf{z})} && (12) \\ &\propto p(\mathbf{z}|\mathbf{x}, \ell)p(\ell|o, c)p(o, c), && (13) \end{aligned}$$

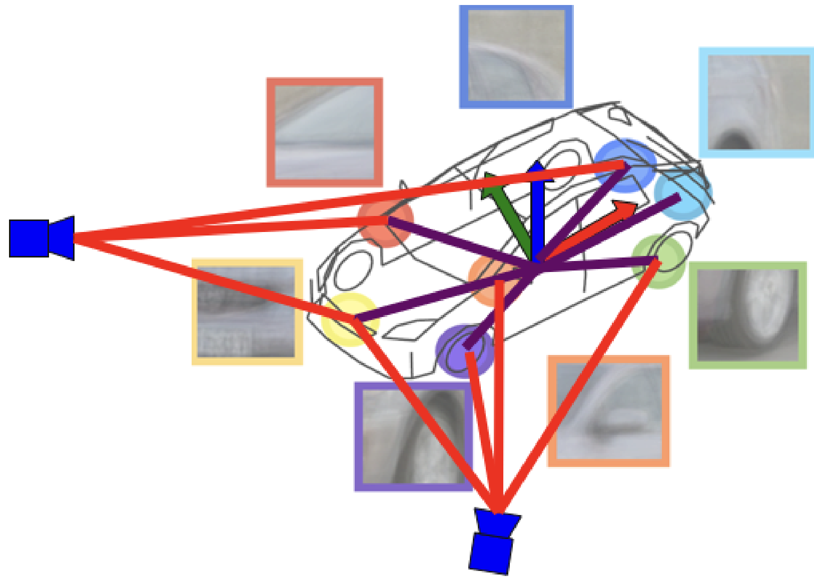

Figure 2: Example factor structure for a car object observed from two camera poses

where we assume uniform priors $p(\ell)$, $p(\mathbf{x}, \ell)$, and $p(\mathbf{z})$.

Let us first examine the first term in (13), $p(\mathbf{z}|\mathbf{x}, \ell)$, and begin to compute log-probabilities as required in (7). As the measurements $\mathbf{z}$ are simply perspective projections of the keypoints onto an image plane with some additive (Gaussian) measurement noise, we have

$$\log p(\mathbf{z}|\mathbf{x}, \ell) \propto \log \prod_{i=1}^{p} p(z_i|\mathbf{x}, \ell_i) \tag{14}$$

$$\propto -\sum_{i=1}^{p} \|z_i - h_\pi(\mathbf{x}, \ell_i)\|^2_{\mathbf{R}}, \tag{15}$$

where $\mathbf{R} \in \mathbb{R}^{2\times 2}$ is the image measurement covariance matrix.

Next, let us examine the second term $p(\ell|o, c)$. This probability relates to the deformable object structure, and describes how likely a given object configuration is given the learned object basis structure. Let ${}^G\bar{q}_O$ and ${}^G p_O$ be the rotation and position, respectively, of the object with respect to the global frame. Following equation (8), we have

$$\ell_i = R({}^G\bar{q}_O)\left(\mathbf{b}_0^i + \sum_{j=1}^{k} c_j \mathbf{b}_j^i\right) + {}^G p_i, \quad i = 1, \ldots, p \tag{16}$$

$$= R({}^G\bar{q}_O)\mathbf{s}_i(c) + {}^G p_i, \quad i = 1, \ldots, p, \tag{17}$$

where $\mathbf{b}_j^i$ is the ith column of $\mathbf{B}_j$, and $\mathbf{s}_i(c)$ is the ith structure-determined keypoint position in the local frame with deformation coefficients c.

Because the deformable shape model may not perfectly capture all intraclass variation, and because keypoint positions will not be estimated perfectly due to image noise and state uncertainty, we allow for estimated keypoints ℓ to vary from their structure $\mathbf{s}(c)$ by introducing a gaussian noise term $w_{st} \sim \mathcal{N}(0, \mathbf{R}_{struct})$, writing a probabilistic expression for ℓ_i as

$$\ell_i = R({}^G\bar{q}_O)\mathbf{s}_i(c) + {}^G p_i + w_{st}, \quad i = 1, \ldots, p. \tag{18}$$

We can now write the desired log-probability as

$$\log p(\ell|o,c) = \log \prod_{i=1}^{p} p(\ell_i|o,c) \tag{19}$$

$$\propto -\sum_{i=1}^{p} \|\ell_i - R({}^G\bar{q}_O)\mathbf{s}_i(c) - {}^Gp_i\|^2_{\mathbf{R}_{struct}}. \tag{20}$$

Finally, let us examine the term $p(o,c)$. We assume that the deformation coefficients are independent of the object pose and that the pose prior $p(o)$ is uniform, so we have $p(o,c) \propto p(c)$. As in [Pavlakos et al., 2017], we use the term $p(c)$ as a simple regularizer on the coefficients c:

$$\log p(c) \propto -\lambda \|c\|_2^2, \tag{21}$$

where λ is a chosen regularization parameter.

Combining equations (13), (15), (20), and (21), we can now write the expression for the full semantic measurement log-probability,

$$-\log p(\mathbf{x},o,\ell,c|\mathbf{z}) \propto \sum_{i=1}^{p} \|z_i - h_\pi(\mathbf{x},\ell_i)\|^2_{\mathbf{R}} + \sum_{i=1}^{p} \|\ell_i - R({}^G\bar{q}_O)\mathbf{s}_i(c) - {}^Gp_i\|^2_{\mathbf{R}_{struct}} + \lambda\|c\|_2^2. \tag{22}$$

In practice, a single object is necessarily observed from multiple different camera poses. While each observation alters the measurement probability (equation (15)) associated with the object, the structure probabilities (equations (20) and (21)) remain the same. Suppose an object is observed by a set of measurements $\{\mathbf{z}_i\}_{i=1}^{K}$. We can write the full log-probability associated with this object as

$$-\log p(\mathbf{x},o,\ell,c|\mathbf{z}_{1:K}) \propto \sum_{k=1}^{K}\sum_{i=1}^{p} \|[\mathbf{z}_k]_i - h_\pi(\mathbf{x},\ell_i)\|^2_{\mathbf{R}} + \sum_{i=1}^{p} \|\ell_i - R({}^G\bar{q}_O)\mathbf{s}_i(c) - {}^Gp_i\|^2_{\mathbf{R}_{struct}} + \lambda\|c\|_2^2, \tag{23}$$

where $[\mathbf{z}_k]_i$ is the ith keypoint measurement in measurement $\mathbf{z}_k$.

4 System Architecture and Experiments

To thoroughly demonstrate the effectiveness of our semantic SLAM method, we performed experiments on two separate platforms. In both systems, we created a semantic front-end that selects every 10th camera frame as a semantic keyframe. First, our front-end applies to each key image the Faster R-CNN object detector [Ren et al., 2015] to detect object bounding boxes. To each detected bounding box, we applied the semantic keypoint detector from [Pavlakos et al., 2017] to detect each object's semantic keypoints. Next, the Mahalanobis distance between each measurement and each object in the map of the same class is computed, and a simple maximum likelihood data association is performed with the Hungarian algorithm [Munkres, 1957]. The resulting keypoint measurements and their data associations were then used in custom GTSAM [Dellaert, 2012] factors that implement equation (23) for inclusion in the larger factor graph SLAM architecture.

4.1 OmniMapper System

The first SLAM architecture is based on the OmniMapper [Trevor et al., 2014] system. The OmniMapper is a factor-graph based SLAM that builds an odometry spine by ICP matching LiDAR scans, including loop closure constraints on recognizing previously visited locations. In addition to relative pose measurements between subsequent poses based on scan matching, the OmniMapper classifies LiDAR scans as ground or

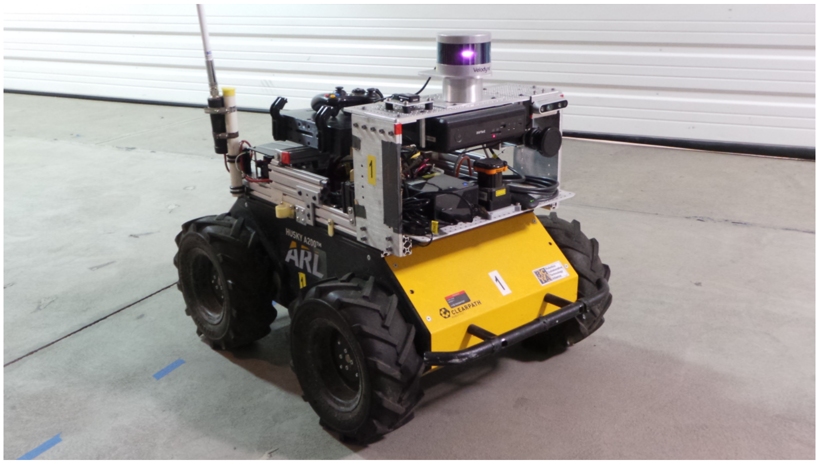

Figure 3: Clearpath Husky robot used in first series of experiments

Figure 4: Example image collected from Husky robot along with semantic keypoint detections

obstacle hits based on their vertical position and uses this classification to update a log-odds occupancy grid that can be used for navigation and obstacle avoidance.

The robot platform used in this experiment is the Clearpath Husky, shown in Figure 3. LiDAR and camera data was collected from trajectories in an urban environment and processed offline. See Figure 4 for an example image collected along the trajectory along with semantic keypoints detected on a window, and see Figure 5 for the system's estimate of the robot trajectory and map at the time the picture in Figure 4 was taken. Note the one detected and localized window shown in the estimate, along with the four semantic keypoints that correspond to the window corners.

Continuing the trajectory Figure 6 shows a later point in the experiment after a longer path through the urban environment, showing the estimated trajectory, occupancy grid, and several estimated window objects.

Our method is also able to perform well at single-object localization up close, with applications of manipulation or other interaction where precise pose estimates are necessary. A robot was driven on a straight line trajectory towards a black crate placed on the ground and images were continuously taken of the crate. See Figure 7 for an example of an image as the robot nears the crate, and Figure 8 for the estimated trajectory and crate pose along with the position of the crate's semantic keypoints. Note how the keypoints line up directly over the occupancy grid-shown obstacle that the crate represents, as well as the subjective quality of the keypoint

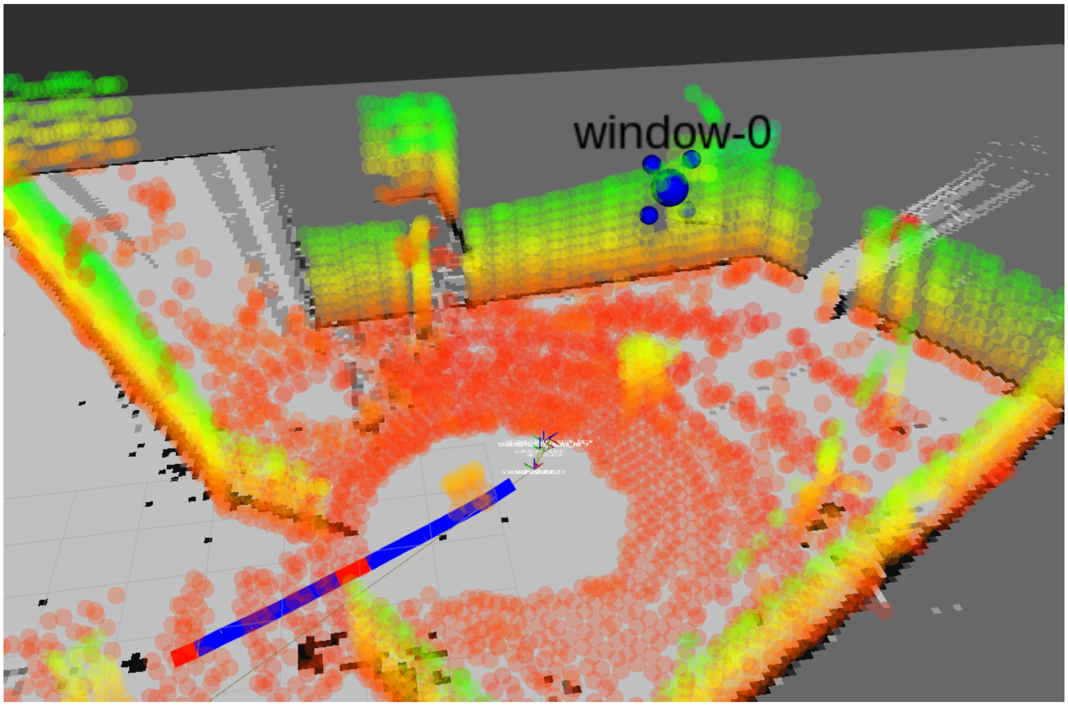

Figure 5: Estimate of the robot trajectory, map, and detected window object at the time at which the image in Figure 4 was taken

localization relative to their displayed positions on the crate in Figure 7.

4.2 Visual Odometry-based SLAM on KITTI

The second SLAM system is based on the algorithm presented in [Bowman et al., 2017]. We implement a GTSAM [Dellaert, 2012]-based algorithm based on stereo visual odometry from the VISO [Geiger et al., 2011] visual odometry algorithm. In addition to relative pose factors from visual odometry, we include geometric visual feature factors as described in [Bowman et al., 2017], and custom semantic object factors as described above that implement equation (23). We applied our algorithm to trajectory 05 in the KITTI [Geiger et al., 2012] outdoor dataset. The KITTI dataset consists of a vehicle equipped with several sensors driving through an urban environment, and parked cars were used as the estimated semantic objects.

See Figure 9 for an example image taken from early in the KITTI dataset trajectory 05, showing a challenging example of a line of parked cars with numerous occlusions and that is traversed at relatively high speed. Our algorithm's estimate of the trajectory along with the estimated cars is shown in Figure 10. Although some detections were missed, the detections and estimated poses and keypoints are very accurate given the conditions.

In Figure 11, our algorithm's trajectory and map estimate after a longer trajectory is shown. Even in long trajectories with numerous objects in the map and several loop closure situations, our algorithm is able to localize not only the camera's position along the trajectory, but also the position and orientation of parked cars along the path.

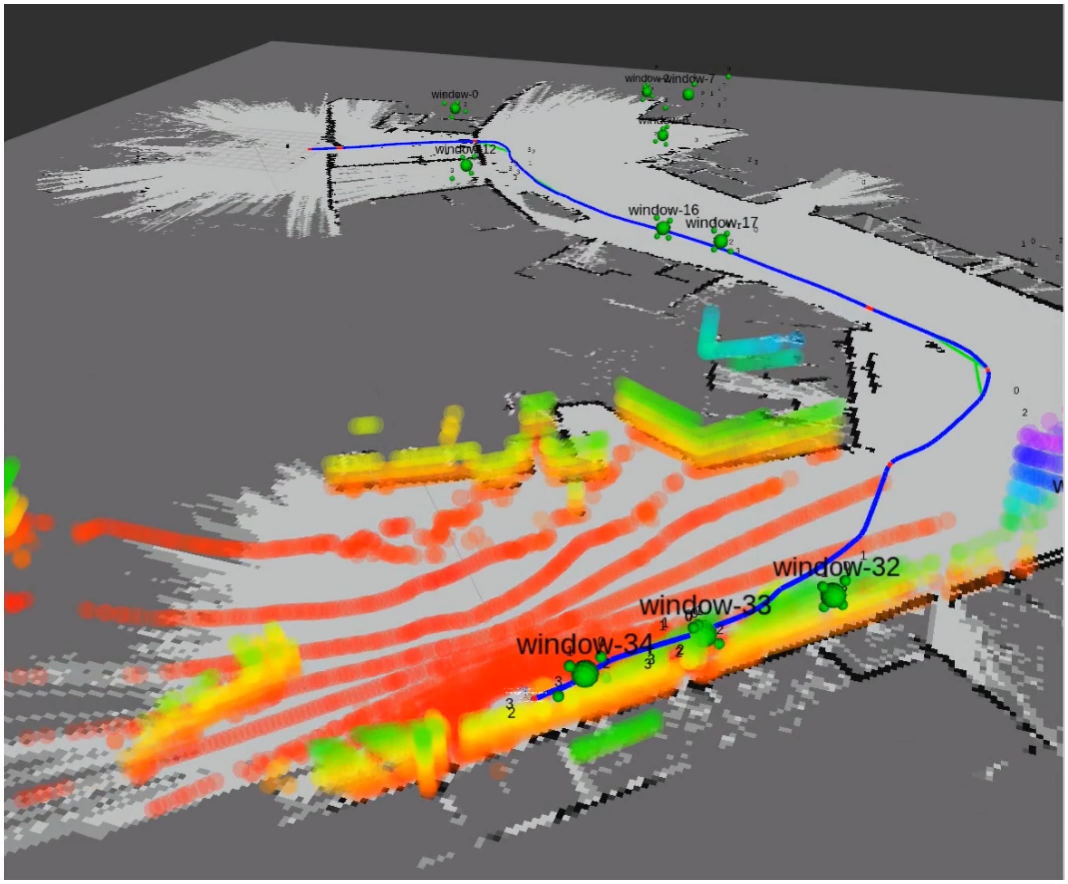

Figure 6: Estimate of the trajectory after a longer path through the same environment as seen in Figures 4 and 5.

5 Conclusion

We have presented a method that is able to reliably and efficiently integrate semantic objects into a SLAM system, not only improving the performance of robot localization, but also providing an accurate semantic map usable for higher level autonomy. The representation of objects as a sparse and deformable skeleton of keypoints allows for efficient solution of the SLAM problem, allowing for real-time solutions even after lengthy trajectories. Our experiments demonstrated our method's ability to integrate into different factor graph-based SLAM systems and its performance in varying environments, providing robust semantic information across varying platforms.

6 Acknowledgements

We would like to thank Karl Schmeckpeper and Jon Fink for assistance in experimentation and data collection. We would also like to thank the RCTA leadership team including Stuart Young, Dava Baran, Dilip Patel, and Matthew Weiker for support and guidance throughout the program.

References

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

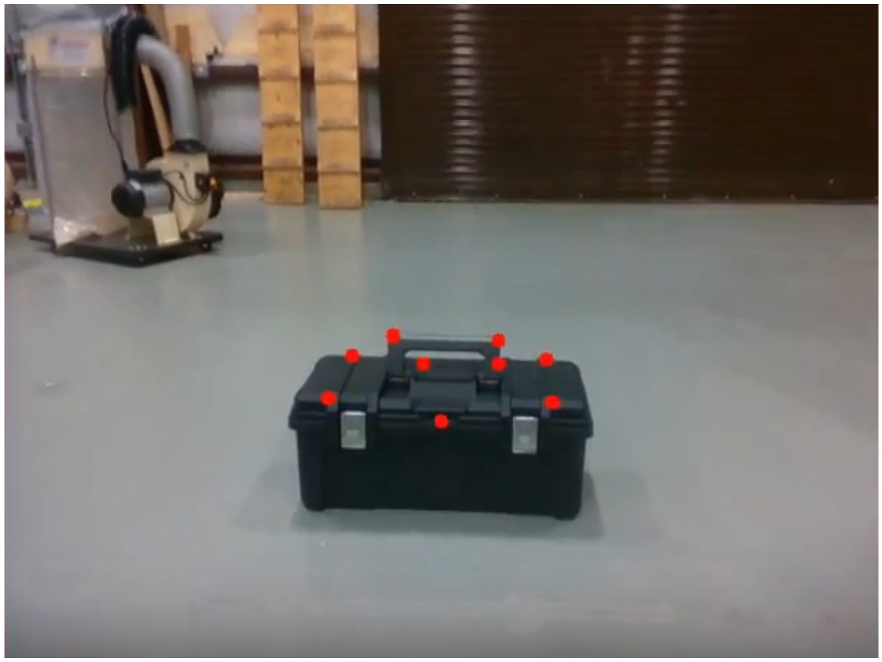

Figure 7: Image as a robot nears a black crate placed on the ground along with detected semantic keypoints.

Bowman, S. L., Atanasov, N., Daniilidis, K., and Pappas, G. J. (2017). Probabilistic data association for semantic slam. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1722–1729.

Civera, J., Galvez-Lopez, D., Riazuelo, L., Tardos, J., and Montiel, J. (2011). Towards Semantic SLAM Using a Monocular Camera. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 1277–1284.

Dellaert, F. (2012). Factor graphs and gtsam: A hands-on introduction. Technical Report GT-RIM-CP&R-2012-002, GT RIM.

Durrant-Whyte, H. and Bailey, T. (2006). Simultaneous localization and mapping: part i. *Robotics Automation Magazine, IEEE*, 13(2):99–110.

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

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

Henry, P., Krainin, M., Herbst, E., Ren, X., and Fox, D. (2012). RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments. *The International Journal of Robotics Research (IJRR)*, 31(5):647–663.

Hesch, J. A., Kottas, D. G., Bowman, S. L., and Roumeliotis, S. I. (2014). Consistency Analysis and Improvement of Vision-aided Inertial Navigation. *IEEE Trans. on Robotics (TRO)*, 30(1):158–176.

Kaess, M., Johannsson, H., Roberts, R., Ila, V., Leonard, J., and Dellaert, F. (2012). iSAM2: Incremental Smoothing and Mapping Using the Bayes Tree. *The International Journal of Robotics Research (IJRR)*, 31(2):216–235.

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

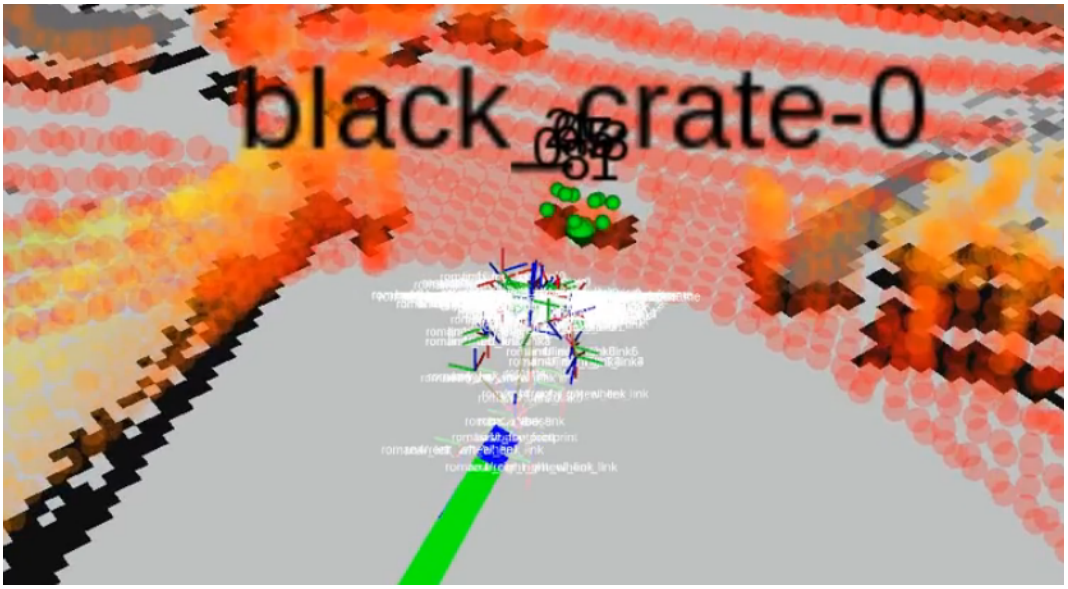

Figure 8: Estimate of the trajectory as the robot approaches the crate along with its detected pose and 3D keypoint positions.

Figure 9: Image from early in the KITTI dataset trajectory 05, showing a line of parked cars.

Lianos, N., Schönberger, J. L., Pollefeys, M., and Sattler, T. (2018). VSO: Visual Semantic Odometry. In *European Conference on Computer Vision (ECCV)*.

Munkres, J. (1957). Algorithms for the Assignment and Transportation Problems. *Journal of the Society for Industrial & Applied Mathematics (SIAM)*, 5(1):32–38.

Mur-Artal, R. and Tardós, J. D. (2016). ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras. *CoRR*, abs/1610.06475.

Pavlakos, G., Zhou, X., Chan, A., Derpanis, K. G., and Daniilidis, K. (2017). 6-dof object pose from semantic keypoints. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2011–2018.

Pillai, S. and Leonard, J. (2015). Monocular slam supported object recognition. In *Proceedings of Robotics: Science and Systems (RSS)*, Rome, Italy.

Pronobis, A. (2011). *Semantic Mapping with Mobile Robots.* dissertation, KTH Royal Institute of Technology.

Redmon, J. and Farhadi, A. (2017). Yolo9000: Better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

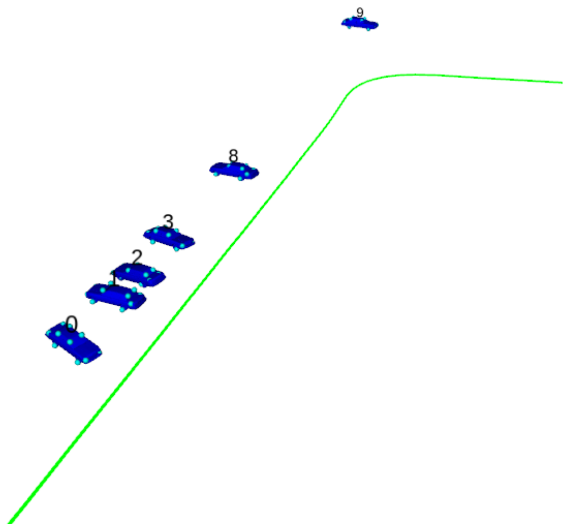

Figure 10: Image from early in the KITTI dataset trajectory 05, showing a line of parked cars.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc.

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

Trevor, A. J. B., Rogers, J. G., and Christensen, H. I. (2014). Omnimapper: A modular multimodal mapping framework. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1983–1990.

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

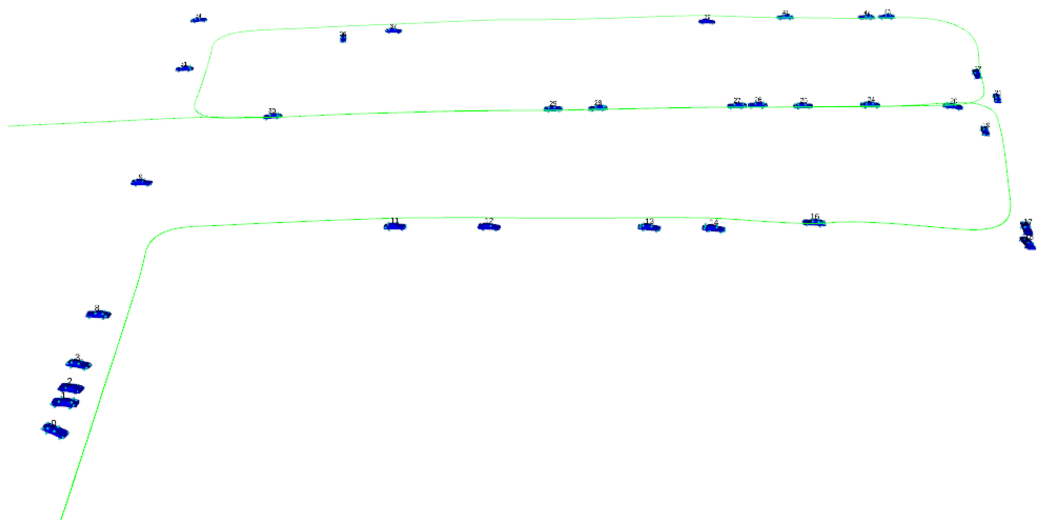

Figure 11: Estimated trajectory and objects from KITTI trajectory 05 after a longer duration