Seminar Report: Probabilistic Data Association for Semantic SLAM

Damian Bogunowicz

Faculty of Informatics - Technical University of Munich

Abstract

The paper “Probabilistic Data Association for Semantic Slam” was released by the team of researchers from the GRASP Lab from University of Pennsylvania, Philadelphia. It was published and submitted to the ICRA (International Conference on Robotics and Automation) conference in 2017. The publication has been awarded for the Best Conference Paper. The proposed method addresses two fundamental issues related to SLAM: loop closure and data association. The work shows, that those two problems can be efficiently tackled by including semantic information about landmarks in the optimization framework. Loop closure recognition which utilizes semantic information has proven to be much more robust than the traditional approaches, which rely solely on low-level geometric features. Regarding data association, the authors propose a formulation of an optimization problem, which involves a shift from “hard decision” on data association (computing maximum likelihood) to “soft decision” (considering the whole distribution). This conception helps the agent to relate it’s measurements to correct landmarks.

1 Introduction

Simultaneous localization and mapping (SLAM) is a problem of mapping an unknown environment while estimating a robot’s pose within it. This is a crucial ability for agent to have in many domains of robotics. Such tasks as navigation, object manipulation or autonomous surveillance require accurate knowledge of the robot’s pose and the surrounding environment. The difficulty of the problem comes from the fact, that it is not clear which of the two sub-problems (localization or mapping) should be considered the cause and which should be considered the effect. This can be conceptually viewed as a chicken and egg problem.

The first important idea of the paper is considering semantic information in the optimization algorithm. Traditionally, most of the SLAM approaches use the low-level geometric features such as corners [5] or surface patches [4] to reconstruct the metric structure of the environment. Additionally, the authors consider semantic information extracted from the scene. This can be obtained through recent object detection methods such as [2] or [9]. Apart from those two information sources (metric and semantic), one could additionally incorporate inertial measurements e.g. collected by inertial measurement unit (IMU) into the optimization network. The authors prove that such a coupling of observation sources improves the ability of an agent to perform loop closing. In SLAM, loop closure recognition is a problem of detecting visited locations to correct errors accumulated over time.

The second contribution is tackling the problem of data association - finding the correct relationship between sensor measurements and landmarks. The incorporation of semantic information allows for formulation of an optimization problem in the form of an expectation minimization (EM) algorithm. This enables using the whole distribution of data associations during optimization. Such an approach reduces the probability of incorporating false positives into the map.

2 Probabilistic Data Association In SLAM

The SLAM problem can be formulated in the following way: the robot moves in an unknown environment modelled as a collection of M static landmarks $\mathcal{L} = \{l_m\}_{m=1}^{M}$. The agent should estimate the landmark positions $\mathcal{L}$ and a sequence of poses (sensor trajectory) $\mathcal{X} = \{\mathbf{x}_t\}_{t=1}^{T}$ given a set of sensor measurements $\mathcal{Z} = \{\mathbf{z}_t\}_{k=1}^{K}$. The robot relates measurements with the landmarks through data association $\mathcal{D} = \{(\alpha_k, \beta_k)\}_{k=1}^{K}$. This notation indicates that a measurement z_k of landmark $l_{\beta_k}$was obtained from a particular sensor state x_{α_k}.

A complete statement of a SLAM problem involves iterating over two optimization steps:

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

$$\mathcal{X}^{i+1}, \mathcal{L}^{i+1} = \arg\max_{\mathcal{X},\mathcal{L}} \log p(\mathcal{Z}|\mathcal{X}, \mathcal{L}, \mathcal{D}^{i+1}) \tag{2}$$

First, given prior estimates $\mathcal{X}^i$ and $\mathcal{L}^i$, we compute a new maximum likelihood estimate of data association $\mathcal{D}^{i+1}$ (e.g. via Hungarian algorithm [8]). Secondly, given $\mathcal{D}^{i+1}$, we may find new estimates of most likely set of landmarks and poses. This can be computed using filtering [7] or pose-graph optimization [6].

The authors reformulate equation (1) in order to improve agent's ability to perform correct data association. Initially, $\mathcal{D}^{i+1}$ is being chosen as a mode of $p(\mathcal{D}|\mathcal{X}^i, \mathcal{L}^i, \mathcal{Z})$. This way hard decisions on data associations are considered. In the refined formulation, when estimating new landmarks and poses, the entire density of $\mathcal{D}^{i+1}$ is being computed. This is being reflected in the expectation maximization (EM) algorithm in the following way:

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

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

where $\mathbb{D}$ is a space of all possible values of $\mathcal{D}$. The new formulation has the advantage that it "averages" over all possible associations - no hard decisions are computed anymore. We can further rewrite (3) as follows:

$$\mathcal{X}^{i+1}, \mathcal{L}^{i+1} = \arg\max_{\mathcal{X},\mathcal{L}} \sum_{k=1}^{K} \sum_{j=1}^{M} w_{kj}^i \log p(\mathbf{z}_k|\mathbf{x}_{\alpha_k}, l_j) \tag{4}$$

where weight w_{kj}^i quantifies the influence of the "soft" data association between measurement k and landmark j.

As result, the EM formulation of SLAM consists of two steps:

1. Expectation step - instead of computing 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_{kj}^i.
2. Maximization step - maximize the expected measurement log likelihood over the previously computed distribution.

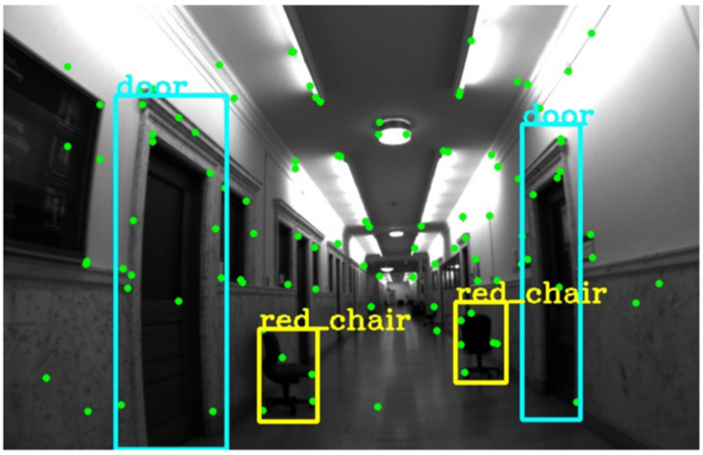

Figure 1: Sample keyframe image overlaid with ORB features (green points) and objects detections (bounding boxes and semantic classes). Source: [1]

3 Semantic SLAM

While in traditional SLAM, the landmarks are usually estimated solely from low-level features, our approach additionally considers semantic information. That means that each landmark l contains two pieces of information - position $l^p \in \mathbb{R}^3$ and a class label $\mathbb{C}$. This implies that we need to gather more data from the environment. As mentioned in chapter one, there are three pieces of information, which allow the robot to successfully tackle the SLAM problem:

- Inertial information - the robot is equipped with a sensor package which consists of an IMU and one monocular camera. For every sensor state x_t (where t is a time step), the agent observes the environment using keyframes from the camera and IMU readings (6-D pose, velocity and IMU bias values) related to this keyframe. The IMU measurements computed between timestep t and $t+1$ (such as linear acceleration and rotational velocity) are denoted as $\mathcal{I}_t$, the inertial information.

- Geometric information - additionally, in every timestep, our agent takes a keyframe and extracts ORB features from the image. Those features, denoted as geometric information $\mathcal{Y}_t$, are then tracked forward to the subsequent keyframe.

- Semantic information - the last type of measurement are semantic observations extracted from every keyframe. This information can be obtained by taking a keyframe and extracting object detections using e.g. deformable parts model (DPM) detector [6, 10, 2]. From every keyframe we obtain a set of object detections $\mathcal{S}_t$. A single object detection $\mathbf{s}_k = (s_k^c, s_k^s, s_k^b) \in \mathcal{S}_t$ consists of a detected class s_k^c, a score quantifying the detection confidence s_k^s and a bounding box s_k^b.

While the inertial and geometric measurements are used to track the sensor trajectory locally, the semantic measurements are utilized to construct a map of objects that can be used to perform the loop closure. While closure based solely on low-level features is often viewpoint-dependent and subject to failure in ambiguous or repetitive environment, addition of the object detections makes the method more robust.

Having discussed not only the mathematical foundations behind the optimizations steps (chapter two), but also types of measurements and their role in the algorithm (chapter three), we can formulate the final task of the algorithm: given inertial, geometric and semantic measurements, estimate the sensor state trajectory and the positions and classes of the objects in the environment.

4 Semantic SLAM Using EM

Following the observations from chapter two, we may write down the detailed formulation of the expectation maximization algorithm. As previously said, this formulation considers the whole density of data association.

$$w_{kj}^{t,(i)} = \sum_{l^c \in \mathcal{C}} \sum_{\mathcal{D}_t \in \mathbb{D}_t(k,j)} \frac{p(\mathcal{S}_t | \mathcal{X}^{(i)}, \mathcal{L}^{(i)}, \mathcal{D}_t)}{\sum_{l^c} \sum_{\mathcal{D}_t \in \mathbb{D}_t} p(\mathcal{S}_t | \mathcal{X}^{(i)}, \mathcal{L}^{(i)}, \mathcal{D}_t)}, \forall t, k, j \tag{5}$$

$$\mathcal{X}^{(i+1)}, l_{1:M}^{p,(i+1)} = \operatorname*{arg\,max}_{\mathcal{X}, l_{1:M}^{p}} \sum_{t=1}^{T} \sum_{\mathbf{s}_k \in \mathcal{S}_t} \sum_{j=1}^{M} w_{kj}^{t,(i)} \log p(\mathbf{s}_k | \mathbf{x}_t, l_j) + \log p(\mathcal{Y}|\mathcal{X}) + \log p(\mathcal{I}|\mathcal{X}) \tag{6}$$

In the expectation step we compute all data association weights for the given timestep. Then we use the computed weights to calculate new estimations of sensor states and landmark positions (maximization step). The optimization equation (6) contains three terms: semantic observation term, geometric observation term and inertial term. Let's investigate each of those elements individually.|

4.1 Semantic Term

$$w_{kj}^{t,(i)} \log p(\mathbf{s}_k | \mathbf{x}_t, l_j) = ||h_\pi(\mathbf{x}_t, l_j) - s_k^b||_{\mathbf{R}^2_{s/w_{k,j}^{t,(i)}}} \tag{7}$$

where $h_\pi(\mathbf{x}_t, l_j)$ is standard perspective projection of a landmark $l_j$onto camera at pose $\mathbf{x}_t$. The camera measurement is Gaussian distributed with mean $h_\pi(\mathbf{x}_t, l_j)$ and covariance $\mathbf{R_s}$. The semantic term (due to the re-observation of a previously seen landmark) is the method's source of loop closure constraints.

4.2 Geometric Term

$$\log p(\mathcal{Y}|\mathcal{X}) = \sum_{i=1}^{N_y} \sum_{k:\beta_k^y = i} ||h(\mathbf{x}_{\alpha_k^y}, \rho_i) - \mathbf{y}_k||^2_{\mathbf{R}_y} \tag{8}$$

where N_y is the total number of observed physical geometric landmarks, and ρ_i is a 3D position of a landmark, that generated measurement $\mathbf{y}_k$, in the global frame of reference. The projection has Gaussian pixel noise with covariance $\mathbf{R}_y$. The geometric term introduces geometric measurements as structureless constraints between the camera poses that observed them.

4.3 Inertial Term

$$\log p(\mathcal{I}_{ij}|\mathcal{X}) = -||\mathbf{r}_{\mathcal{I}_{ij}}||^2_{\sum_{ij}} \tag{9}$$

where $\mathbf{r}_{\mathcal{I}_{ij}}$ is a vector containing inertial residuals on the rotation, velocity and position differences between two successive keyframes $\mathbf{x}_i$ and $\mathbf{x}_j$. $\sum_{ij}$ is a noise covariance of the residuals.

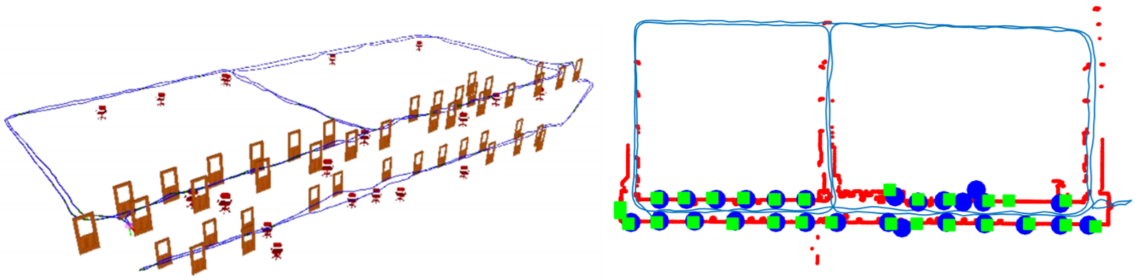

Figure 2: Estimated trajectory (blue line) in a sample office experiment from the algorithm. On the left hand side: visualisation of landmark position and classes. On the right hand side: estimated door landmark positions (blue circles) are compared with their ground truth positions (green squares). Red points are a partial ground truth map. Source: [1]

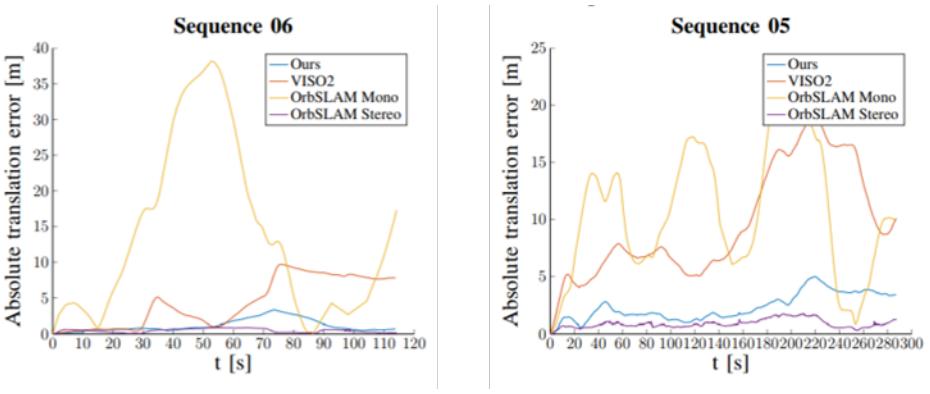

Figure 3: Norm of position error between estimate and ground truth for KITTI seq. 06 (left) and seq. 05 (right). Source: [1]

5 Experiments

The algorithm has been tested on several long-trajectory real indoor and outdoor datasets, which include odometry and visual measurements in cluttered scenes and varying lighting conditions. The first group of experiments were conducted inside an office building, while the second set of tests were conducted on KITTI dataset [3].

5.1 Office Building Experiments

The method is tested through several runs inside an office building. The landmarks used in this environment are different types of chairs and doors. Due to fact that the agent uses both semantic, geometric and inertial information, the algorithm correctly closes loops and obtains an accurate map of the office floor. It is shown to be superior to ORB-SLAM2, which is unable to capture the correct map in a repetitive environment such as office hallways.

5.2 KITTI Outdoor Dataset Experiments

KITTI is a dataset for computer vision research in the context of autonomous driving. The data has been recorded in and around the city of Karlsruhe, Germany. In this experiment, the agent was localizing itself using cars. As in the case of indoor environment, the method delivers very good result also on outdoor dataset. On KITTI sequence 05 or 06 it performs much better then ORBSLAM Mono or VISO2. In terms of absolute translational error our method can be compared to Stereo OrbSLAM.

6 Conclusions

The authors introduce semantic features into the optimization framework for improved localization performance and loop closure. This way, it is possible to reconstruct the full 6-D pose, as well as positions and classes of the objects contained in the environment. This way the algorithm handles complex and cluttered scenes without any noticeable increase of computational cost. The research confirms the fact, that using semantic information about environment can improve autonomous operation of robots.

References

[1] S. L. Bowman, N. Atanasov, K. Daniilidis, and G. J. Pappas. Probabilistic data association for semantic slam. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1722–1729, May 2017. 1, 2, 3

[2] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, Sep. 2010. 1, 3

[3] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, June 2012. 5

[4] Peter Henry, Michael Krainin, Evan Herbst, Xiaofeng Ren, and Dieter Fox. Rgb-d mapping: Using kinect-style depth cameras for dense 3d modeling of indoor environments. *The International Journal of Robotics Research*, 31(5):647–663, 2012. 1

[5] Joel A. Hesch, Dimitrios G. Kottas, Sean L. Bowman, and Stergios Roumeliotis. Consistency analysis and improvement of vision-aided inertial navigation. *IEEE Transactions on Robotics*, 30(1):158–176, 1 2014. 1

[6] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard. G2o: A general framework for graph optimization. In *2011 IEEE International Conference on Robotics and Automation*, pages 3607–3613, May 2011. 2, 3

[7] A. I. Mourikis and S. I. Roumeliotis. A multi-state constraint kalman filter for vision-aided inertial navigation. In *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pages 3565–3572, April 2007. 2

[8] J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, 1957. 2

[9] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, June 2017. 1

[10] Menglong Zhu, Nikolay Atanasov, George J. Pappas, and Kostas Daniilidis. Active deformable part models inference. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 281–296, Cham, 2014. Springer International Publishing. 3