Scalable real-time vision-based SLAM for planetary rovers

Robert Sim, Matt Griffin, Alex Shyr, and James J. Little
Laboratory for Computational Intelligence
University of British Columbia
Vancouver, BC, V6T 1Z4, Canada
{*simra,mgriffin,shyr,little*}@*cs.ubc.ca*

Abstract

Simultaneous localization and mapping is an important problem for autonomous planetary rovers and other space vehicles. While many authors have addressed the SLAM problem, few have done so in the context of producing large-scale maps in real time using vision. This paper is concerned primarily with the issues presented by the large numbers of candidate features obtained from vision sensors, and the implications for data association. We present a Rao-Blackwellised particle filter (RBPF) SLAM implementation that employs a stereo camera and the SIFT feature detector, and demonstrate that we can build maps of unknown environments with an extensive number of visual landmarks.

1 Introduction

Autonomous planetary rovers that rely on vision sensing require the ability to construct dense visual representations of their environment for the purposes of both navigation and data collection. A central problem to constructing these representations is that as a mobile rover explores it accumulates error in its pose estimate, and subsequently its acquired map becomes inaccurate. This problem, generally referred to as simultaneous localization and mapping, or SLAM, has been widely studied and a variety of solutions have been proposed (for example [3,5,8,9]). However, there are a limited number of vision-based solutions that address real-time mapping, and representations that can scale up to thousands of mapped features. This paper presents an approach and experimental results for achieving SLAM solutions in real-time over long trajectories (73m or more), resulting in maps consisting of many thousands of landmarks.

Our approach to solving the SLAM problem with a vision sensor is to combine a Rao-Blackwellised particle filter (RBPF)-based approach to mapping [8], coupled with efficient data structures developed by Montemerlo, *et. al.* for representing a distribution over maps (referred to as Fast-SLAM [7]), and fast data association techniques for matching the relatively unambiguous feature descriptors obtained using the SIFT feature detector [6].

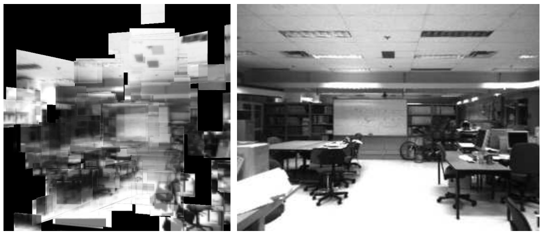


Figure 1. Left: A rendering of the map from a sensor's-eye view. Right: an image of the lab from a nearby position.

RBPF-based SLAM solutions operate by maintaining multiple map hypotheses, each associated with a stochastically sampled trajectory through the environment. The complete set of sampled trajectories and inferred maps approximates the probability distribution of maps conditioned on the vehicle's actions and observations, $p(M|A, Z)$, where $M = \{m_1, m_2, \ldots, m_n\}$ is the set of maps, each consisting of a set of probability distributions describing landmark positions, $A = \{u_1, u_2, \ldots, u_m\}$ are the control inputs to the vehicle (that is, the vehicle's actions), and $Z = \{z_1, z_2, \ldots, z_m\}$ are the vehicle's observations of the world (for brevity, we assume actions and observations are interleaved). One of the important contributions of the Fast-SLAM algorithm is the data structure it employs to share information between trajectory samples with common history. This facilitates real-time performance of the algorithm as the trajectory length grows.

As an exploratory vehicle moves through the environment, the number of landmarks in its map can grow to number in the hundreds of thousands. This is especially true for vision-based mapping, where feature detectors might typically return 500 feature observations in a single image. This poses a difficult problem for solving the data association problem, where a single feature observation might require comparison with all of the landmarks in the map. Such an extensive comparison might be required when extracted features are generic, without any uniquely defining characteristics (such as those typically employed in mapping algorithms employing laser range sensors). Furthermore, the computed data association is rarely unique, and often highly ambiguous. While FastSLAM allows for multiple data association hypotheses, these can reduce the robustness of the particle filter and potentially lead to sample starvation.

In vision, however, there is usually a great deal of contextual information associated with a feature that can constrain data association, and reduce the cost of matching. In our work, we employ the SIFT feature descriptor, which provides descriptions of feature observations that have been shown to be very robust for feature correspondence. In addition, we apply a kd-tree over the space of SIFT features to facilitate approximate nearest-neighbor lookups in time logarithmic in the number of visually distinct landmarks.

The main contributions of this paper are two-fold. First, we present an implementation of FastSLAM which is based on vision-based sensing, rather than traditional range sensing with a laser. Second, we present methods for performing rapid data association of hundreds of landmark observations in a single image against a database of tens of thousands of mapped landmarks. These results leverage the strengths of particle filter-based approaches for uncertainty estimation (such as the possibility of multi-modal and non-Gaussian estimates), with data association techniques that were previously only applied to Kalman-filter based estimators (for example, [9]). Furthermore, where previous implementations of the FastSLAM algorithm have generally employed sensors with a wide field of view, our experimentation demonstrates the performance of the algorithm using sensors with a comparatively narrow field of view. Finally, we demonstrate experimentally that robust SLAM solutions can be achieved in real-time over long trajectories (more than 70m).

The remainder of this paper is structured to provide a coverage of the strengths and weaknesses of current methods, elaborate on the details of our implementation, present and discuss experimental results, and finally discuss planned improvements.

2 Related Work

There is a significant body of literature on SLAM using the Extended Kalman Filter and its inverse, the Extended Information Filter [3, 5, 12]. These approaches model the posterior distribution over maps as a unimodal Gaussian distribution. Of particular interest is the view based approach of Eustice, *et. al.* [3], which enables constant-time filter updating without significant sparsification approximations. However, a significant difficulty with a view-based approach is that the resulting map does not lend itself well to evaluation or human inspection, a strong prerequisite for an exploratory vehicle.

Two map representations are popular in the literature, landmark based [7, 9, 11] and occupancy grid based [2, 4]. Occupancy grids are effective for dense but ambiguous information while landmarks are more suited to sparse but distinguishable features. Very impressive occupancy grids have been produced online by recent scan matching techniques which also use particle filters for pose estimation [4] [2]. Landmark and vision-based approaches have also performed well in the past, as in [9]. In the latter case, a reasonably small environment was successfully mapped by using a Kalman Filter and assuming independence between landmark and pose estimates. For large environments, this approach is likely to be overconfident and lead to filter divergence.

In related work, we have also applied our approach to mapping where control and odometry information is unknown. In Sim, *et. al.*, we demonstrated an approach to solving the SLAM problem using the approach outlined here, coupled with a visual odometry estimate for motion estimation [10].

3 Simultaneous Localization and Mapping

This paper represents map estimation as the evolution of a Rao-Blackwellised particle filter [8]. In this context, the trajectory and landmark distribution is modeled as a dynamic Bayes network, where trajectories are instantiated as samples, and the landmark distribution can be expressed analytically for each trajectory. At time t, let s_t denote the vehicle pose, m_t the map learned thus far and $x_t = \{s_t, m_t\}$ be the complete *state*. Also, let u_t denote a control signal or a measurement of the vehicle's motion from time $t-1$ to time t and z_t be the current observation. The set of observations and controls from time 0 to t are denoted as z^t and u^t respectively. Our goal is to estimate the density

$$p(s_t, m_t|z^t, u^t) = p(x_t|z^t, u^t) \quad (1)$$

It has been demonstrated elsewhere that $p(s_t, m_t|z^t, u^t)$ can be approximated by factoring the distribution in terms of sampled trajectories s_t, and independent landmark distributions conditioned on the sampled trajectories [8]:

$$p(s_t, m_t|z^t, u^t) \approx p(s_t|z^t, u^t) \prod_k p(m(k)|s^t, z^t, u^t) \quad (2)$$

where $m(k)$ denotes the $k-th$ landmark in the map. That is, we instantiate a set of samples s_t, propagate them according to u^t, and construct maps for each according to z^t.

A simplistic approach to running an RBPF for SLAM would yield an update complexity of $O(MN)$, where M is the number of particles at each step and N is the number of landmarks. However, Montemerlo et al. introduced in their FastSLAM work a tree-based structure which refines this complexity to $O(M \log N)$ by sharing landmark estimates between samples [7]. Each sample in the filter will share unaltered landmark estimates with its ancestor particles (those landmarks that have not been observed from the time of the ancestor to the present). Each landmark observation results in a landmark being copied from its parent and updated but the rest of the tree remains the same.

3.1 Data Association

In an unmodified natural environment, landmarks are difficult to uniquely identify. This problem is known as data association or correspondence and incorrectly matching observations to landmarks can lead to inconsistencies in a map. Stereo vision can quickly provide 3D information and when coupled with a scale-invariant feature transform (SIFT) detector [6] it can provide distinct landmarks. SIFT features are desirable as landmarks because they are somewhat invariant to image scale, rotation and translation as well as to illumination changes and affine or 3D projection. This combination can result in many viable landmarks from an unaltered environment.

4 Implementation

4.1 State Representation

We describe samples of the vehicle's pose with the vector $s_t = [x, y, \theta]$, situated in a plane [1].

At each time step, the N pose samples are propagated according to the motion model $p(s_t|s_{t-1}, u_t)$, which is user-defined. Over time the distribution of samples can become non-Gaussian, and even multi-modal. The noise model we apply is designed to take a conservative approach to estimating the possible drift in the robot's pose over time, while keeping in mind that noisier models require more particles to prevent starvation as the underlying pose distribution disperses.

The specific action sequence is dependent on the robot's exploration policy. For this paper, we drive the robot by hand, and infer the robot's actions from odometry measurements between observations. For each action the filter produces a new generation of particles that inherit properties from the previous generation. The set of hypothetical trajectories executed by the robot is represented by this tree of particles. After taking an observation (described in the next section), each particle in the current generation of particles is weighted according to the probability of the current observation z_t, conditioned on that particle's trajectory:

$$\begin{aligned} w_i &= p(z_t|s_{i,t}, m_{i,t}) & (3)\\ &= k \exp(-0.5 \Delta z^T \Sigma^{-1} \Delta z) & (4) \end{aligned}$$

where $\Delta z = h(s_{i,t}) - z_t$, $h(\cdot)$ is a generative model of the observations as a function of pose, Σ is the sum of the observation covariance and observed landmark covariance. The particle is weighted by how well the current observation is consistent with the map constructed from that particle's trajectory. The weights are subsequently normalized across the population of samples, and then sampled probabilistically with replacement to produce the next generation of particles. Should any particle not be chosen for advancement it is pruned and all particles with no children are then recursively removed from the tree. If an insufficient number of particles are used, or resampling takes place too frequently, this can lead to starvation as hypotheses are pruned.

4.2 Sensing and Data Association

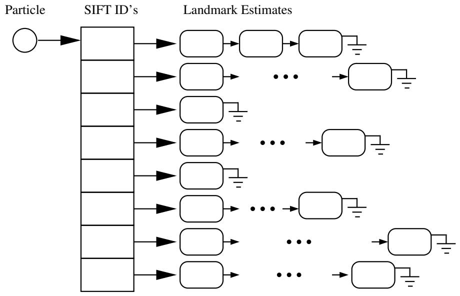


Figure 2. Each particle has an associated map, organized by SIFT descriptor. Similarly, each SIFT descriptor might have multiple landmark estimates, each spatially distinct.

We employ a data structure similar to that described in [7] as a map representation. Each particle has associated with it a set of landmark estimates, described by Gaussian distributions. However, in the vision-based case, we take advantage of the descriptive power of the SIFT transform (described below), enabling us improve the quality of data association. In this formulation, each particle maintains a list of SIFT IDs, and these IDs in turn point to a linked list of one or more 3D landmark estimates (Figure 2). Note that one SIFT ID can point to multiple landmarks– landmarks that have similar appearance but are spatially distinct.

[1] Our current work is aimed at extending our results to a full 6-DOF motion model.

We are using a Point Grey Research BumbleBee stereo camera for our primary sensor, and extract SIFT features using a difference of Gaussian detector [6]. The features' partial invariance to image scale, rotation, translation and 3D or affine projection are what make them desirable landmarks. Each SIFT feature has a 36 dimension identifier, or key, associated with it and this matching is based on finding a suitably distinct match. We perform a linear search of the keys in the left image for each key in the right. The two keys with the smallest Euclidean distances from our target key are found and if the ratio of best and second best distances is below a set threshold (currently 0.6) it is considered a good match. That is, for keys k_{l1},, k_{l2} and k_r, in the left and right images, according to subscripts, a successful match of k_{l1} to k_r satisfies the property

$$\frac{k_{l1} - k_r}{k_{l2} - k_r} < 0.6 \tag{5}$$

Once a 3D feature is extracted from the stereo pair, we determine if this feature corresponds to one we have seen before. Our approach to data association is depicted in Figure 3. To efficiently store and access what can quickly become a large number of keys we use a kd-tree. The kd-tree facilitates nearest-neighbor matching in time logarithmic in the size of the tree, and has been demonstrated to be reliable for object recognition tasks [1]. The disadvantage of using a kd-tree is that it can sometimes produce not the nearest match but a close match. We maintain a single tree for the sensor and associate an arbitrary integer ID with each SIFT identifier we add. New keys are considered to be *candidate keys* and are not passed as an observation to the particle filter until they have been observed for a sufficient number of frames. Since we do not currently use negative information to remove erroneous landmarks from the maps this is an effort to limit their number.

Each particle's map is indexed by a set of IDs associated with SIFT descriptors and each node contains a linked list of landmarks sharing that descriptor. Multiple data associations can be entertained by the filter because each particle determines the specific landmark to which an observation corresponds. The number of landmarks associated with an ID is typically quite small as shown by Table 1. A particle's weight is updated for a given landmark observation according to Equation 4 by first selecting from the linked list for the matched landmark ID the landmark estimate that is closest to the observed point in the global frame of reference. The maximum distance threshold for this comparison is based on an approximation of the camera's error and if multiple landmarks fall within this range the closest is chosen. Clearly since the filter is initiated without a map any observation with an unknown ID or a 3D position which does not match is treated as a new landmark.

In the following section we describe our experimental results.

5 Experimental Results

For the purposes of our experiments, we used an RWI B14 robot with a BumbleBee stereo head from Point Grey Research. The robot was driven through a laboratory environment, and the robot collected 5000 images along a trajectory of approximately 74m. We ran the system using 100 particles, which enabled a frame rate of approximately 2.1Hz (Figure 5). Table 1 describes at approximately 1000 time-step intervals the average number of landmarks associated with each map, the total distance traveled, the total number of SIFT id's, the time step in history at which the filter converges to a single map, and the total number of landmark instances in the system (these can outnumber the product of samples and mean landmarks as many instances are not promoted to full landmarks until they have been observed at least 3 times).

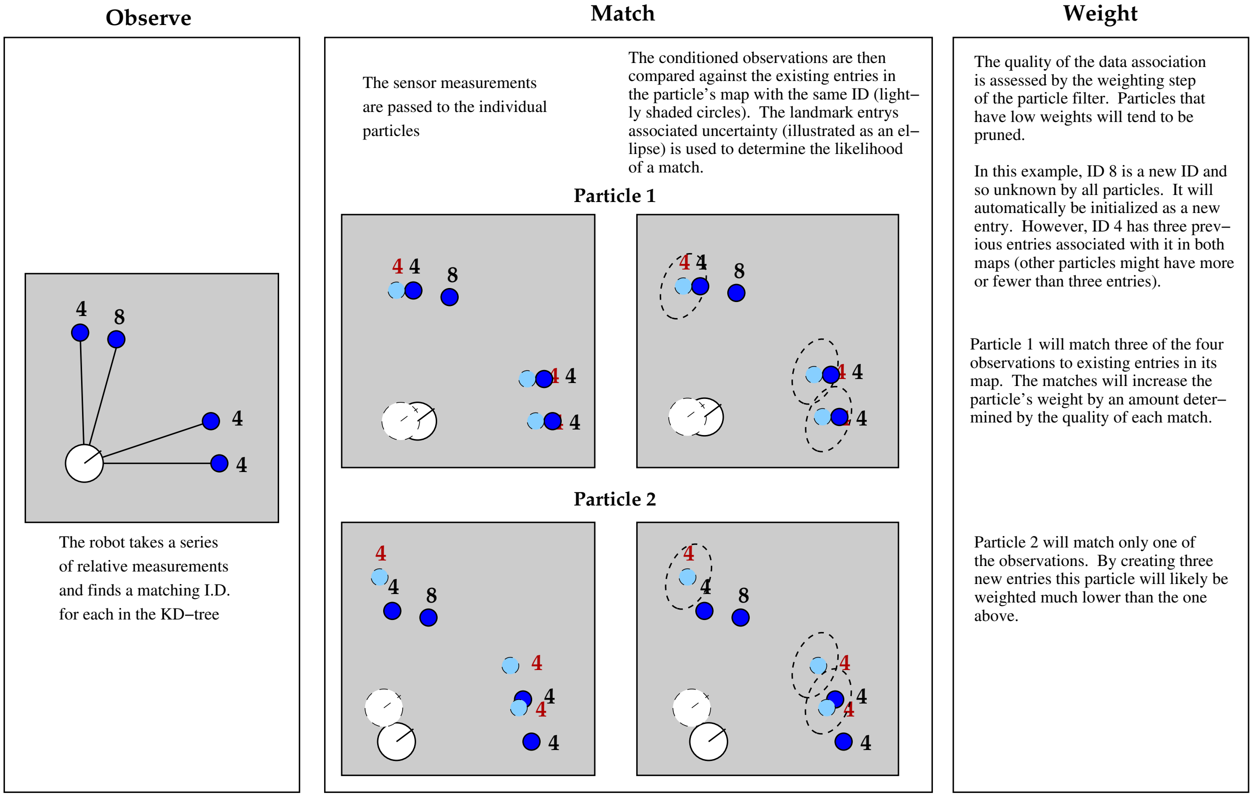
Table 1. Map summary (see text for details).

Time	Avg. LMs per Samp.	Dist. traveled (m)	Tracked Sift Features	Filter convergence	Total LMs
5043	31577	72.64	29462	4847	108636
4009	25300	59.66	23766	4000	84146
2987	18830	42.02	17826	2889	70369
2043	12840	27.05	12338	1989	60871
1021	6083	14.04	5922	919	23688

Figure 4 depicts the map constructed for the maximum-likelihood particle at the end of exploration, beside a map computed using dead reckoning alone. The filter-based map is clearly more accurate in that it correctly captures the rectilinear alignment of the three rooms traversed by the robot.

6 Discussion

Among the key observations from our experiments is that we are able to successfully map a large environment in real-time. At the end of map construction, we are matching 29,462 SIFT features, and each map consists of more than 31,000 landmarks, with a total of only 109,000 landmarks shared across all the maps. As the maps grow in



The robot takes a series of relative measurements and finds a matching I.D. for each in the KD−tree

The sensor measurements are passed to the individual particles

The conditioned observations are then compared against the existing entries in the particle's map with the same ID (lightly shaded circles). The landmark entrys associated uncertainty (illustrated as an ellipse) is used to determine the likelihood of a match.

The quality of the data association is assessed by the weighting step of the particle filter. Particles that have low weights will tend to be pruned.

In this example, ID 8 is a new ID and so unknown by all particles. It will automatically be initialized as a new entry. However, ID 4 has three previous entries associated with it in both maps (other particles might have more or fewer than three entries).

Particle 1 will match three of the four observations to existing entries in its map. The matches will increase the particle's weight by an amount determined by the quality of each match.

Particle 2 will match only one of the observations. By creating three new entries this particle will likely be weighted much lower than the one above.

Figure 3. SIFT-based data association.

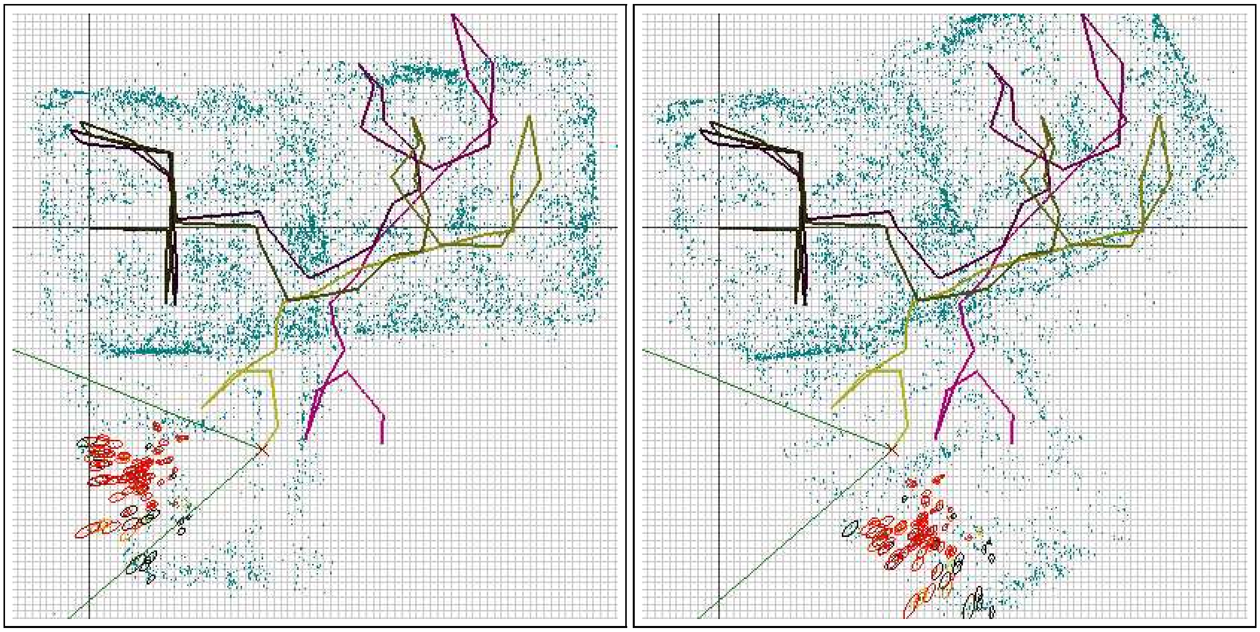


Figure 4. Left: The constructed map for the best sample at the end of exploration. Yellow: maximum weight trajectory. Pink: Dead reckoning trajectory. Grid lines indicate 25cm intervals. Right: the constructed map using dead reckoning.

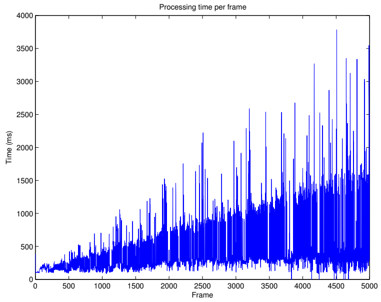


Figure 5. Per-frame computation time in milliseconds. The mean frame rate is 2.1Hz on a 3.2GHz Intel Xeon Processor

size, only a small slow-down in computation time is exhibited. More importantly, relative to dead-reckoning, we have demonstrated that our resulting map is more accurate. This can be observed in the fact that the three rooms observed by the robot are correctly aligned, as well as the fact that when the robot returns to the first room it correctly locates the position of the door into the third room (near the center of the constructed map).

7 Conclusion

We have presented an implementation of an RBPF-based SLAM algorithm using a vision-based sensor. The key goal of our work is to facilitate scalable maps that incorporate large numbers of visual landmarks (on the order of hundreds of thousands). The primary contributions of this work are the facilitation of vision-based sensing in association with a particle filter that supports multiple data associations. We have also experimentally demonstrated the success of the system on a real robot.

Among the goals of our work are the problem of constructing visual representations of very large environments (on the order of 100m in diameter), and in particular in outdoor environments. We also hope to extend this work to full 6-DOF representations, which will be better suited to space vehicles and exploration over rough terrain. We believe that the data structures developed in this work can facilitate the construction of maps of this size, and our ongoing goal is to accomplish this task experimentally. One outstanding question for our work is how to reliably cull SIFT features that are rarely observed. This features tend to clutter the kd-tree and deletion requires costly re-balancing in the tree. An outstanding question is whether loop closure can be successful in very large environments, and how many particles might be required to successfully close the loop. Finally, we are interested in developing approaches to autonomous exploration based on particle-based representations of the environment.

Acknowledgements

The authors wish to thank Pantelis Elinas for providing assistance with key matching and stereo correspondence. This work was supported in part by the Canadian Space Agency/NSERC Postdoctoral Fellowship, and NSERC undergraduate summer research assistantships.

References

[1] J. S. Beis and D. G. Lowe. Shape indexing using approximate nearest-neighbour search in high-dimensional spaces. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1000–1006, Peurto Rico, June 1997. IEEE, IEEE Press.

[2] A. I. Eliazar and R. Parr. DP-slam 2.0. In *Proc. ICRA 2004*, New Orleans, LA, 2004. IEEE Press.

[3] R. Eustice, H. Singh, and J. Leonard. Exactly sparse delayed-state filters. In *Proc. of the 2005 IEEE Int. Conf. on Robotics and Automation*, pages 2428–2435, Barcelona, Spain, April 2005.

[4] D. Hähnel, D. Fox, W. Burgard, and S. Thrun. A highly efficient FastSLAM algorithm for generating cyclic maps of large-scale environments from raw laser range measurements. In *Proc. Conf. Intelligent Robots and Systems (IROS)*, 2003.

[5] J. J. Leonard and H. F. Durrant-Whyte. Simultaneous map building and localization for an autonomous mobile robot. In *Proc. of the IEEE Int. Workshop on Intelligent Robots and Systems*, pages 1442–1447, Osaka, Japan, November 1991.

[6] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Int. Conf. on Computer Vision*, pages 1150–1157, Corfu, Greece, September 1999. IEEE Press.

[7] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit. FastSLAM: A factored solution to the simultaneous localization and mapping problem. In *Proceedings of the AAAI National Conf. on Artificial Intelligence*, Edmonton, Canada, 2002. AAAI.

[8] K. Murphy. Bayesian map learning in dynamic environments. In *1999 Neural Information Processing Systems*, 1999.

[9] S. Se, D. G. Lowe, and J. Little. Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. *Int. J. Robotics Research*, 21(8):735–758, 2002.

[10] R. Sim, P. Elinas, M. Griffin, and J. J. Little. Vision-based SLAM using the Rao-Blackwellised particle filter. In *Proceedings of IJCAI Workshop on Reasoning with Uncertainty in Robotics*, Edinburgh, Scotland, 2005.

[11] R. Smith, M. Self, and P. Cheeseman. Estimating Uncertain Spatial Relationships in Robotics. In *Autonomous Robot Vehicles*, I.J. Cox and G.T. Wilfong, eds. 167–193, Springer-Verlag. 1990.

[12] S. Thrun, Y. Liu, D. Koller, A. Ng, Z. Ghahramani, and H. Durrant-Whyte. Simultaneous localization and mapping with sparse extended information filters. *Int. J. Robotics Research*, 23(7-8):693–716, 2004.