Self-Organizing Visual Maps

Robert Sim
Department of Computer Science
University of British Columbia
2366 Main Mall
Vancouver, BC, V6T 1Z4

Gregory Dudek
Centre for Intelligent Machines
McGill University
3480 University St. Suite 409
Montreal, QC, H3A 2A7

Abstract

This paper deals with automatically learning the spatial distribution of a set of images. That is, given a sequence of images acquired from well-separated locations, how can they be arranged to best explain their genesis? The solution to this problem can be viewed as an instance of robot mapping although it can also be used in other contexts. We examine the problem where only limited prior odometric information is available, employing a feature-based method derived from a probabilistic pose estimation framework. Initially, a set of visual features is selected from the images and correspondences are found across the ensemble. The images are then localized by first assembling the small subset of images for which odometric confidence is high, and sequentially inserting the remaining images, localizing each against the previous estimates, and taking advantage of any priors that are available. We present experimental results validating the approach, and demonstrating metrically and topologically accurate results over two large image ensembles. Finally, we discuss the results, their relationship to the autonomous exploration of an unknown environment, and their utility for robot localization and navigation.

Introduction

This paper addresses the problem of building a map of an unknown environment from an ensemble of observations and limited pose information. We examine the extent to which we can organize a set of measurements from an unknown environment to produce a visual map of that environment with little or no knowledge of where in the environment the measurements were obtained. In particular, we are interested in taking a set of snapshots of the environment using an uncalibrated monocular camera, and organizing them to quantitatively or qualitatively indicate where they were taken which, in turn, allows us to construct a visual map. We assume that, at most, we have limited prior trajectory information, so as to bootstrap the process– the source of this information might be from the first few odometry readings along a trajectory, the general shape of the trajectory, information from an observer, or from a localization method that is expensive to operate, and hence is only applied to a small subset of the observation poses. While metric accuracy is of interest, our primary aim is to recover the topology of the ensemble. That is, to assure that metrically adjacent poses in the world are topologically adjacent in the resulting map.

The problem of automated robotic mapping is of substantial pragmatic interest for the development of mobile robot systems. The question of how we bootstrap a spatial representation, particularly a vision-based one, also appears to be relevant to other research areas such as computer vision and even ethology. Several authors have considered the use of self-organization in robot navigation (Takahashi *et al.* 2001; Beni & Wang 1991; Deneubourg *et al.* 1989; Selfridge 1962), often with impressive results. We believe this paper is among the first to demonstrate how to build a complete map of a real (non-simulated) unknown environment using monocular vision. We present quantitative data to substantiate this.

We approach the problem in the context of probabilistic robot localization using learned image-domain features (as opposed to features of the 3D environment) (Sim & Dudek 2001). To achieve this there are two steps involved: first, reliable features are selected and correspondences are found across the image ensemble. Subsequently, the quantitative behaviours of the features as functions of pose are exploited in order to compute a maximum-likelihood pose for each image in the ensemble. While other batch-oriented mapping approaches are iterative in nature (Thrun, Fox, & Burghard 1998; Kohonen 1984), we demonstrate that if accurate pose information is provided for a small subset of images, the remaining images in the ensemble can be localized without the need for further iteration and, in some cases, without regard for the order in which the images are localized.

Outline

In the following section, we consider prior work related to our problem; in particular, approaches to self-organizing maps, and the simultaneous localization and mapping problem. We then proceed to present our approach, providing an overview of our feature-based localization framework, followed by the details of how we apply the framework to organize the input ensemble. Finally, we present experimental results on a variety of ensembles, demonstrating the accuracy and robustness of the approach.

Previous Work

The construction of self-organizing spatial maps (SOM's) has a substantial history in computer science. Kohonen developed a number of algorithms for covering an input space (Kohonen 1984; 1995). While spatial coverage has been used as a metaphor, the problem of representing a data space in terms of self-organizing features has numerous applications ranging from text searching to audition. The problem of spanning an input space with feature detectors or local basis functions has found wide application in machine learning, neural nets, and allied areas. In much of this algorithmic work, the key contributions have related to convergence and complexity issues.

The issue of automated mapping has also been addressed in the robotics community. One approach to fully automated robot mapping is to interleave the map synthesis and position estimation phases of robot navigation (sometimes known as SLAM: simultaneous localization and mapping). As it is generally applied, this entails incrementally building a map based on geometric measurements (e.g. from a laser rangefinder, sonar or stereo camera) and intermittently using the map to correct the robot's position as it moves (Leonard & Durrant-Whyte 1991b; Yamauchi, Schultz, & Adams 1998; Davison & Kita 2001). When the motion of a robot can only be roughly estimated, a topological representation becomes very attractive. Early work by Kuipers and Byun used repeated observation of a previously observed landmark to instantiate cycles in a topological map of an environment during the mapping process (Kuipers & Byun 1987; 1991) . The idea of performing SLAM in a topological context was also been examined theoretically (Deng & Mirzaian 1996). The probabilistic fusion of uncertain motion estimates has been examined by several authors (cf, (Smith & Cheeseman 1986)) and the use of Expectation Maximization has recently proven quite successful although it still depends on estimates of successive robot motions (Shatkay & Kaelbling 1997; Thrun 1998; Choset & Nagatani 2001).

A closely related problem in computer vision is that of close-range photogrammetry, or structure-from-motion (SFM) (Longuet-Higgins 1981; Hartley & Zisserman 2000; Davison 2003), which involves recovering the ensemble of camera positions, as well as the full three-dimensional geometry of the scene that is imaged. In the case where a para-perspective camera model is assumed, the problem is linear and the solution can be computed directly.

The key difference between the SFM problem and inferring pose with a visual map is that a solution to the SFM problem is dependent on explicit assumptions about the optical geometry of the imaging apparatus. In the visual mapping framework we have avoided committing to any such assumptions, and as such the self-organizing behaviour exhibited in the experimental results is equally applicable to exotic imaging hardware, such as an omnidirectional camera.

Visual Map Framework

Our approach employs an adaptation of the visual map framework described in (Sim & Dudek 2001). We review it here in brief and refer the reader to the cited work for further details.

The key idea is to learn visual features, parametrically describe them so that they can be used to estimate one's position (that is, they can be used for localization). The features are pre-screened using an attention operator that efficiently detects statistically anomalous parts of an image and robust, useful features are recorded along with an estimate of their individual utility.

In the localization context, assume for the moment that we have collected an ensemble of training images with ground-truth position information associated with each image. The learning framework operates by first selecting a set of local features from the images using a measure of visual attention, tracking those features across the ensemble of images by maximizing the correlation of the local image intensity of the feature, and subsequently parameterizing the set of observed features in terms of their behaviour as a function of the known positions of the robot.

The resulting *tracked features* can be applied in a Bayesian framework to solve the localization problem. Specifically, given an observation image $\mathbf{z}$, the probability that the robot is at pose $\mathbf{q}$ is proportional to the probability of the observation conditioned on the pose:

$$p(\mathbf{q}|\mathbf{z}) = \frac{p(\mathbf{z}|\mathbf{q})p(\mathbf{q})}{p(\mathbf{z})} \tag{1}$$

where $p(\mathbf{q})$ is the prior on $\mathbf{q}$ and $p(\mathbf{z})$ is a normalization constant. For a feature-based approach, we express the probability of the observation conditioned on the pose as a mixture model of probability distributions derived from the individual features:

$$p(\mathbf{z}|\mathbf{q}) = k \sum_{\mathbf{l}_i \in z} p(\mathbf{l}_i|\mathbf{q}) \tag{2}$$

where $\mathbf{l}_i$ is a detected observation of feature i in the image and k is a normalizing constant.

The individual feature models are generative in nature. That is, given the proposed pose $\mathbf{q}$, an expected observation $\mathbf{l}_i^*$ is generated by learning a parameterization $\mathbf{l}_i^* = F_i(\mathbf{q})$ of the feature, and the observation probability is determined by a Gaussian distribution centered at the expected observation and with covariance determined by cross-validation over the training observations. Whereas in prior work the parameterization was computed using radial basis function networks, in this work we construct the interpolants using bilinear interpolation of the observations associated with the nearest neighbour training poses, as determined by the Delaunay triangulation of the training poses. In this work, the feature vector $\mathbf{l}_i$ is defined as the position of the feature in the image:

$$\mathbf{l}_i = [x_i \ y_i] \tag{3}$$

Other scenarios might call for a different choice of feature vector.

A pose estimate is obtained by finding the pose $\mathbf{q}^*$ that maximizes Equation 1. It should be noted that the framework requires no commitment as to how uncertainty is represented or the optimization is performed. It should be noted, however, that the probability density for $\mathbf{q}$ might be multi-modal, and, as is the case for the problem at hand, weak priors on $\mathbf{q}$ might require a global search for the correct pose. For this work, we employ a multi-resolution grid decomposition of the environment, first approximating $p(\mathbf{q}|\mathbf{z})$ at a coarse scale and computing increasingly higher resolution grids in the neighbourhood of $\mathbf{q}^*$ as it is determined at each resolution.

Self-Organization

We now turn to the problem of inferring the poses of the training images when ground truth is unavailable, or only partially available. The self-organization process involves two steps. In the first step, image features are selected and tracked, and in the second step the set of images are localized.

Tracking

Tracking proceeds by considering the images in an arbitrary order (possibly, but not necessarily, according to distance along the robot's trajectory). An attention operator is applied to the first image $\mathbf{z}$ in the set[1], and each detected feature initializes a *tracking set* $T_i \in T$. The image itself is added to the ensemble set E. For each subsequent image $\mathbf{z}$, the following algorithm is performed:

1. A search is conducted over the image for matches to each tracking set in T, and successful matches are added to their respective tracking sets T_i. Call the set of successful matches M.
2. The attention operator is then applied to the image and the set of detected features S is determined.
3. If the cardinality of M is less than the cardinality of S, new tracked sets T_i are initialized by elements selected from S. The elements are selected first on the basis of their response to the attention operator, and second on the basis of their distance from the nearest image position in M. In this way, features in S which are close to prior matches are omitted, and regions of the image where features exist but matching failed receive continued attention. Call this new set of tracking sets T_S.
4. A search for matches to the new tracking sets in T_S is conducted over each image in E (that is, the previously examined images), and the successful matches are added to their respective tracking set.
5. $T = T \cup T_S$
6. $E = E \cup \mathbf{z}$

The template used for by any particular tracking set is defined as the local appearance image of the initial feature in the set. We use local windows of 33 pixels in width and height. Matching is considered successful when the normalized correlation of the template with the local image under consideration exceeds a user-defined threshold.

When tracking is completed, we have a set of feature correspondences across the ensemble of images. The process is $O(kn)$ where k is the final number of tracked sets, and n is the number of images.

Localization

Once tracking is complete, the next step is to determine the position of each image in the ensemble. For the moment, consider the problem when there is a single feature that was tracked reliably across all of the images. If we assume that the image feature is derived from a fixed 3D point in space, the motion of the feature through the image will be according to a monotonic mapping as a function of camera pose and the camera's intrinsic parameters. As such, the topology of a set of observation poses is preserved in the mapping from pose-space to image-space. While the mapping itself is nonlinear (due to perspective projection), it can be approximated by associating actual poses with a small set of the observations and determining the local mappings of the remaining unknown poses by constructing an interpolant over the known poses. Such an algorithm would proceed as follows:

1. Initialize $S = \{(\mathbf{q}, \mathbf{z})\}$, the set of (pose, observation) pairs for which the pose is known. Compute D, the parameterization of S as defined by the feature learning framework.
2. For each observation $\mathbf{z}$ with unknown pose,
   (a) Use D as an interpolant to find the pose $\mathbf{q}^*$ that maximizes the probability that $\mathbf{q}^*$ produces observation $\mathbf{z}$.
   (b) Add $(\mathbf{q}^*, \mathbf{z})$ to S and update D accordingly.

For a parameterization model based on a Delaunay Triangulation interpolant, updating the D takes $O(\log n)$ amortized time, where n is the number of observations in the model. The cost of updating the covariance associated with each model is $O(k \log n)$, where k is the number of samples omitted during cross-validation.

In addition, the cost of finding the maximally likely pose with D is $O(m \log n)$, where m corresponds to the number of poses that are evaluated (finding a face in the triangulation that contains a point $\mathbf{q}$ can be performed in $\log n$ time.). Given n total observations, the entire

[1] We select local-maxima of edge-density.

Figure 1: The LAB scene, a 2.0m by 2.0m pose space.

algorithm takes $O(n(m+k+1)\log n)$ time. Both m and k can be bounded by constants, although in practice we typically bound k by n.

In practice, of course, there is more than one feature detected in the image ensemble. Furthermore, in a suitably small environment, some might span the whole set of images, but in most environments, most are only visible in a subset of images. Finally, matching failures might introduce a significant number of outliers to individual tracking sets. Multiple features, and the presence of outlier observations are addressed by the localization framework we have presented; the maximum likelihood pose is computed by maximizing Equation 2, and the effects of outliers in a tracked set are reduced by their contribution to the covariance associated with that set.

When it cannot be assumed that the environment is small enough such that one or more feature spans it, we must rely on stronger priors to bootstrap the process.

In the following section we present experimental results on two image ensembles.

Experimental Results

A Small Scene

For our first experiment, we demonstrate the procedure on a relatively compact scene. The LAB scene consists of an ensemble of 121 images of the scene depicted in Figure 1. The images were collected over a 2m by 2m environment, at 20cm intervals. Ground truth was measured by hand, accurate to 0.5cm.

Given the ensemble, the images were sorted at random and tracking was performed as described above, resulting in 91 useful tracked features. (A tracked feature was considered useful if it contained at least 4 observations). The localization stage proceeded by first providing the ground truth information to four images selected at random. The remaining images were again sorted at random and added, without any prior information about their pose, according to the methodology described in the previous section. Figure 2 depicts the set of inferred poses versus their ground truth positions. The four 'holes' in the data set at $(20, 200)$, $(40, 140)$, $(60, 0)$, and $(140, 0)$ correspond to the four initial poses for which ground truth was supplied. For the purposes of visualization, Figure 3 plots the original grid of poses,

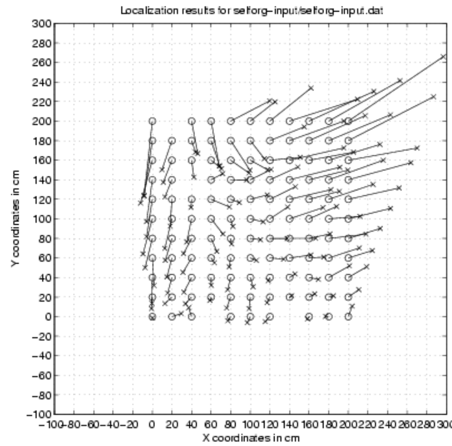

Figure 2: Self-organizing pose estimates plotted versus ground truth for the LAB scene.

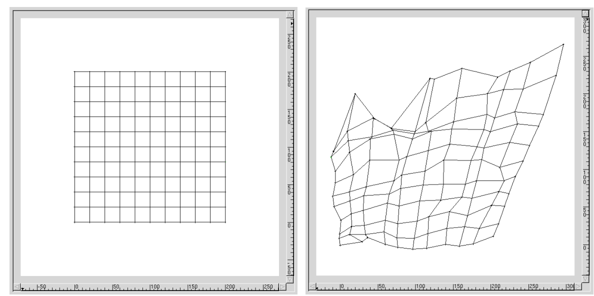

Figure 3: Ground truth, and the map resulting from the self-organizing process for the LAB scene.

and beside it the same grid imposed upon the set of pose estimates computed for the ensemble.

In order to quantify the distortion in the resulting map, the lengths of the mapped line segments corresponding to the original grid were measured, and the average and standard deviation in the segment lengths was recorded. For the ground-truth mesh, the average and standard deviation segment length was 20cm and 0cm, respectively (assuming perfect ground truth). In the inferred map, the mean segment length was 24.2cm and the standard deviation was 11.5cm. These results demonstrate that the resulting map was slightly dilated and with variation in the segment lengths of about 11.5cm or 58% of 20cm on average.

While there is clearly some warping in the mesh, for the most part the topology of the poses is preserved.

A Larger Scene

For our second experiment, we examine a larger pose space, 3.0m in width and 5.5m in depth, depicted in Figure 4. For this experiment, 252 images were collected at 25cm intervals using a pair of robots, one of which used a laser range-finder to measure the 'ground-truth' pose of the moving robot(Rekleitis *et al.* 2001).

Figure 4: The second scene, a 3.0m by 5.5m pose space.

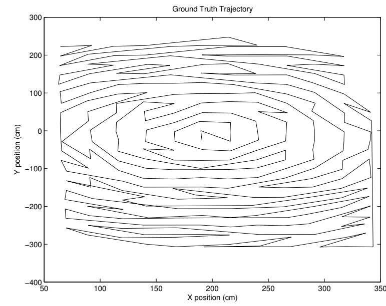

Figure 5: Ground truth simulated trajectory.

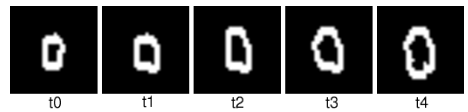

Figure 6: Evolution of the annular prior $p(\mathbf{q})$ over the first few input images. Each thumbnail illustrates $p(\mathbf{q})$ over the 2D pose space at time t_i.

As in the previous experiment, tracking was performed over the image ensemble and a set of 49 useful tracked features were extracted. In this instance, the larger interval between images, some illumination variation in the scene and the larger number of input images presented significant challenges for the tracker, resulting in the smaller number of tracked features.

Given the size of the environment, no one feature spanned the entire pose space. As a result, it was necessary to impose constraints on the *a priori* known ground-truth poses, and the order in which the input images were considered for localization. In addition, a weak prior (i.e. having only a slight effect) $p(\mathbf{q})$ was applied as each image was added in order to control the distortion in the mesh.

Rather than select the initial ground-truth images at random, ground truth was supplied to the four images closest to the centre of the environment. The remainder of the images were sorted by simulating a spiral trajectory of the robot through the environment, intersecting each image pose, and adding the images as they were encountered along the trajectory. Figure 5 illustrates the simulated ground-truth trajectory through the ensemble. Finally, given the sort order, as images were added it was assumed that their pose fell on an annular ring surrounding the previously estimated poses. The radius and width of the ring was defined in terms of the interval used to collect the images. The computed *a priori* distributions over the first few images input into the map are depicted in Figure 6. The intent of using these priors was to simulate a robot exploring the environment along trajectories of increasing radius from a home position.

As in the previous section, Figure 7 plots the original grid of poses, and beside it the same grid imposed upon the set of pose estimates computed for the ensemble. Again, the positive y-axis corresponds to looming forward in the image, and as such the mesh distorts as features accelerate in image space as the camera approaches them. Note however, that as in the first experiment, the topology of the embedded poses is preserved for most of the grid.

Discussion

We have demonstrated an approach to spatially organizing images of an unknown environment using little or no positional prior knowledge. The repeated occurrences of learned visual features in the images allows us to accomplish this. The visual map of the environment that is produced appears to be topologically correct and also demonstrates a substantial degree of metric accuracy and can be described as a locally conformal mapping of the environment. This representation can then be readily used for path execution, trajectory planning and other spatial tasks.

While several authors have considered systems that interleave mapping and position estimation, we believe ours is among the first to do this based on monocular image data. In addition, unlike prior work which typically uses odometry to constrain the localization process, we can accomplish this with essentially no prior estimate of the position the measurements are collected from. On the other hand, if some positional prior is available we can readily exploit it. In the second example shown in this paper we exploited such a prior. Even in this example, it should be noted that the data acquisition trajectory was one that did not include cycles. In general, cyclic trajectories (ones that re-visit previously seen locations via another route) will greatly improve the quality of the results; in fact they are prerequisite for many existing mapping and localization techniques, both topological and metric ones.

We believe that absence of a requirement of a position prior (i.e. odometry) makes this approach suitable for unconventional mapping applications, such as the integration of data from walking robots or from manually collected video sequences. Our ability to do this depends on the repeated occurrence of visual features

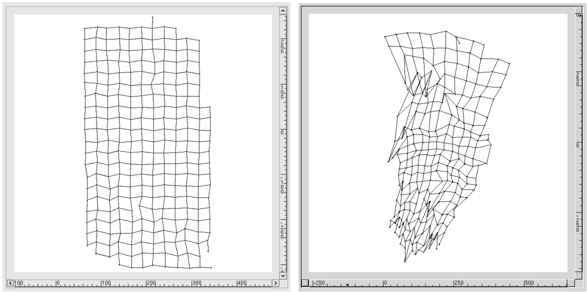

Figure 7: Ground truth, and the map resulting from the self-organizing process for the environment depicted in Figure 1.

in images from adjacent positions. This implies that successfully mapping depends on images being taken at sufficiently small intervals to assure common elements between successive measurements.

References

Beni, G., and Wang, J. 1991. Theoretical problems for the realization of distributed robotic system. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 1914–1920. Sacramento, CA: IEEE Press.

Choset, H., and Nagatani, K. 2001. Topological simultaneous localization and mapping (SLAM): Toward exact localization without explicit localization. *IEEE Transactions on Robotics and Automation* 17(2):125–137.

Davison, A. J., and Kita, N. 2001. 3D simultaneous localisation and map-building using active vision for a robot moving on undulating terrain. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 384–391.

Davison, A. 2003. Real-time simultaneous localisation and mapping with a single camera. In *Proceedings of the IEEE International Conference on Computer Vision*.

Deneubourg, J.; Goss, S.; Pasteels, J.; Fresneau, D.; and Lachaud, J. 1989. Self-organization mechanisms in ant societies (ii): Learning in foraging and division of labor. In Pasteels, J., and Deneubourg, J., eds., *From Individual to Collective Behavior in Social Insects, Experientia Supplementum*, volume 54. Birkhauser Verlag. 177–196.

Deng, X., and Mirzaian, A. 1996. Competitive robot mapping with homogeneous markers. *IEEE Transactions on Robotics and Automation* 12(4):532–542.

Gross, H.; Stephan, V.; and Bohme, H. 1996. Sensory-based robot navigation using self-organizing networks and q-learning. In *Proceedings of the World Congress on Neural Networks*, 94–99. Lawrence Erlbaum Associates, Inc.

Hartley, R., and Zisserman, A. 2000. *Multiple View Geometry in Computer Vision.* Cambridge, United Kingdom: Cambridge University Press.

Kohonen, T. 1984. *Self-Organization and Associative Memory.* New York: Springer-Verlag.

Kohonen, T. 1995. *Self-organizing maps.* Berlin; Heidelberg; New-York: Springer.

Kuipers, B., and Beeson, P. 2002. Bootstrap learning for place recognition. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, 174–180.

Kuipers, B. J., and Byun, Y. T. 1987. A qualitative approach to robot exploration and map-learning. In *Proceedings of the IEEE workshop on spatial reasoning and multi-sensor fusion*, 390–404. Los Altos, CA: IEEE.

Kuipers, B., and Byun, Y.-T. 1991. A robot exploration and mapping strategy based on a semantic hierachy of spatial representations. *Robotics and Autonomous Systems* 8:46–63.

Leonard, J. J., and Durrant-Whyte, H. F. 1991a. Simultaneous map building and localization for an autonomous mobile robot. In *Proceedings of the IEEE Int. Workshop on Intelligent Robots and Systems*, 1442–1447.

Leonard, J. J., and Durrant-Whyte, H. F. 1991b. Mobile robot localization by tracking geometric beacons. *IEEE Transactions on Robotics and Automation* 7(3):376–382.

Leonard, J. J., and Feder, H. J. S. 2000. A computationally efficient method for large-scale concurrent mapping and localization. In Hollerbach, J., and Koditschek, D., eds., *Robotics Research: The Ninth International Symposium*, 169–176. London: Springer-Verlag.

Longuet-Higgins, J. 1981. A computer algorithm for reconstructing a scene from two projections. *Nature* 293:133–135.

Rekleitis, I.; Sim, R.; Dudek, G.; and Milios, E. 2001. Collaborative exploration for the construction of visual maps. In *IEEE/RSJ/ International Conference on Intelligent Robots and Systems*, volume 3, 1269–1274. Maui, HI: IEEE/RSJ.

Selfridge, O. 1962. The organization of organization. In Yovits, M.; Jacobi, G.; and Goldstein, G., eds., *Self-Organizing Systems*. Washington D.C.: McGregor & Werner. 1–7.

Shatkay, H., and Kaelbling, L. P. 1997. Learning topological maps with weak local odometric information. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 920–929. Nagoya, Japan: Morgan Kaufmann.

Sim, R., and Dudek, G. 2001. Learning generative models of scene features. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*. Lihue, HI: IEEE Press.

Smith, R., and Cheeseman, P. 1986. On the representation and estimation of spatial uncertainty. *International Journal of Robotics Research* 5(4):56–68.

Takahashi, T.; Tanaka, T.; Nishida, K.; and Kurita, T. 2001. Self-organization of place cells and reward-based navigation for a mobile robot. *Proceedings of the 8th International Conference on Neural Information Processing, Shanghai (China)* 1164–1169.

Thrun, S.; Fox, D.; and Burghard, W. 1998. A probabilistic approach to concurrent mapping and localization for mobile robots. *Autonomous Robots* 5:253–271.

Thrun, S. 1998. Finding landmarks for mobile robot navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 958–963.

Yamauchi, B.; Schultz, A.; and Adams, W. 1998. Mobile robot exploration and map building with continuous localization. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 3715–2720. Leuven, Belgium: IEEE Press.