Living and Searching in the World: Object-Based State Estimation for Mobile Robots

Lawson L. S. Wong
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
lsw@csail.mit.edu

Mobile-manipulation robots performing service tasks in human-centric indoor environments has long been a dream for developers of autonomous agents. Tasks such as cooking and cleaning require interaction with the environment, hence robots need to know relevant aspects of their spatial surroundings. However, unlike the structured settings that industrial robots operate in, service robots typically have little prior information about their environment. Even if this information was given, due to the involvement of many other agents (e.g., humans moving objects), uncertainty in the complete state of the world is inevitable over time. Additionally, most information about the world is irrelevant to any particular task at hand. Mobile manipulation robots therefore need to continuously perform the task of *state estimation*, using perceptual information to maintain the state, and its uncertainty, of task-relevant aspects of the world.

Because indoor tasks frequently require the use of objects, objects should be given critical emphasis in spatial representations for service robots. Compared to occupancy grids and feature-based maps often used in navigation and SLAM, object-based representations are arguably still in their infancy. In my thesis, I propose a representation framework based on objects, their 'semantic' attributes, and their geometric realizations in the physical world.

Completed Work

In previous work, I have outlined methods for maintaining the type, pose, and occupancy of objects in the world from noisy perception. Within the space of object-based state estimation tasks, perhaps the most basic one is: what objects did the robot perceive, and where are they located in the world? These two properties (type and pose) are examples of object *attributes* that an estimator should track.

Semantic World Modeling from Partial Views (Wong, Kaelbling, and Lozano-Pérez 2013a)

The 'what and where' problem, when considered abstractly on the level of objects and attributes, has a natural generalization: given detections of object attributes only (without knowing which objects generated them), estimate the objects that are present (including their number) and their attributes. I assume the existence of off-the-shelf black-box attribute detectors, such as object recognition and pose estimation modules. Because the information returned from such modules is typically very sparse (at most one detection per object from a single viewpoint), aggregating detections across multiple viewpoints is necessary.

However, this introduces data association issues, because it is unclear which measurements correspond to the same object across different views. I proposed a Bayesian-nonparametric batch-clustering approach, inspired by the observation that 'objects' are essentially clusters in joint attribute space. Given attribute detections from multiple viewpoints, this algorithm outputs samples from the distribution over hypotheses of object states, where a hypothesis consists of a list of objects and their attribute value distributions.

Combining Object and Metric Spatial Information (Wong, Kaelbling, and Lozano-Pérez 2014)

Alas, not all things in the world are objects and attributes. One concept lacking in the above work was the notion that objects occupy physical regions of space. The concept of free space, regions that no object overlaps, was also only implicitly represented. It is therefore difficult, in the object-attribute representation, to incorporate absence/'negative' observations, most prominently that observing a region of free space should suggest that no object overlaps that region. On the other hand, this information is handled very naturally in an occupancy grid (Moravec and Elfes 1985), but grids cannot incorporate the concept of 'objects'.

The complementary advantages of these two representations inspired a search for a way to maintain filters of both object and metric information. Because filtering in the joint state involves complex dependencies and is intractable, I instead adopted the strategy of filtering *separately* in the object and metric spaces by using the existing filters (in particular, the algorithm described in the previous section and occupancy grids respectively). To compensate for the lost dependencies between objects and their geometric realizations, I then developed a way to *merge* the filters on demand as queries about either posterior distribution are made.

I have demonstrated the above framework on a Willow Garage PR2 robot, mounted with a Microsoft Kinect sensor providing 3-D RGB-D images. Objects are recognized by a separate black-box system (Glover and Popovic 2013).

Work in Progress

I am currently investigating along two orthogonal directions: **expressiveness** and **scalability**. The former refers to the class of object attributes, and possibly other non-object spatial information, that the state estimator is capable of maintaining. The latter refers to the complexity of the estimator, both as a function of the number of objects / volume of space that a robot has explored, and of the time that it has been online. Naturally, maximizing expressiveness of the state estimator while maintaining scalability is desirable. Formulating characterizations of this trade-off and demonstrating efficient and sound estimation in real-world scenarios is the primary objective of the remainder of my thesis work.

Expressiveness

So far, I have only considered object type, object pose, and metric occupancy as attributes. There is nothing fundamental in the current approach that limits attention to these three – it is only a matter of integrating additional attribute perception modules. Additional attributes will be considered when integrating the estimator in demonstration tasks such as object search (Wong, Kaelbling, and Lozano-Pérez 2013b).

Besides expanding on static properties, there are at least two aspects of object state that require non-trivial extensions to the existing representations. First is the temporal dynamics of object states, in particular considering changes that do not occur continuously over time, but rather at discrete events. The motivating case for this is intervention by another agent while the robot is away – when the robot returns, how much of the previous world state estimate can it retain? Is it possible/useful to tell if an object has been moved, or must estimates be obtained from scratch once a change has been detected? By modeling long-term object changes using Poisson processes, and appealing to a recently-identified construction of the dependent Dirichlet process (DDP) based on Poisson processes (Lin, Grimson, and Fisher 2010), object dynamics can be included as a natural generalization to my semantic world modeling work by using DDPs.

The second extension is the incorporation of known state constraints. Examples of constraints include object-object non-interpenetration (Wong, Kaelbling, and Lozano-Pérez 2012), support/containment relationships, and stability/contact. State estimation with hard constraints is challenging because they couple together many state variables. However, constraints also offer an avenue for estimation to be more efficient, since they can greatly reduce the feasible state space. For example, the vertical position of objects in a stack is essentially determined once it is recognized that each object must be resting stably on the one beneath it.

Scalability

As the spatial representation's expressiveness increases, the space of possible states grows combinatorially too, and maintaining distributions over all states is clearly intractable. There are at least two approaches for reducing the space that are worthy of further investigation: *factoring* the state (asserting a simpler model, i.e., more independence assumptions and fewer dependencies), and *ignoring* (or delaying evaluation in) certain subspaces of states.

One possibility for the former was already explored in my work on combining object and metric spatial information, where each was filtered independently, and only fused on demand at query time. This strategy offers computational advantages during filtering, while allowing dependencies to still be incorporated when more accurate answers are necessary. Ultimately though, even the most aggressive factorization cannot provide a good solution – there are just too many things one can keep track of in the world! Ideally, a state estimator should consider the task at hand as well, and ignore all information that is irrelevant in the present moment.

To achieve computationally-scalable models that are not restricted by strong simplifying assumptions, the idea is to allow the system to identify task-relevant variables that are not tracked well, and improve upon those models *locally*. More technically, this requires identifying when the estimator's model is mismatched (with respect to task performance), and progressively refining the model by model class expansion (for small task-relevant subsets of variables).

For the former, I am currently exploring execution monitoring techniques (Pettersson 2005), with thresholds automatically learned from task performance. For the latter, one possibility is to use grammars that generate increasingly-complex models (Grosse, Salakhutdinov, and Tenenbaum 2012); another recent approach involves partitioning a hierarchy of variables into groups of varying fineness at data-determined levels (Steinhardt and Liang 2014).

By combining all of the above, I hope to demonstrate a state estimator for robots operating in large indoor environments with many objects over long periods of time.

References

Glover, J., and Popovic, S. 2013. Bingham Procrustean alignment for object detection in clutter. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*.

Grosse, R. B.; Salakhutdinov, R.; and Tenenbaum, J. B. 2012. Exploiting compositionality to explore a large space of model structures. In *Uncertainty in Artificial Intelligence (UAI)*.

Lin, D.; Grimson, E.; and Fisher, J. 2010. Construction of dependent Dirichlet processes based on Poisson processes. In *Neural Information Processing Systems (NIPS)*.

Moravec, H., and Elfes, A. E. 1985. High resolution maps from wide angle sonar. In *IEEE Intl. Conf. on Robotics and Automation*.

Pettersson, O. 2005. Execution monitoring in robotics: A survey. *Robotics and Autonomous Systems* 53:73–88.

Steinhardt, J., and Liang, P. 2014. Filtering with abstract particles. In *Intl. Conf. Machine Learning (ICML)*.

Wong, L. L. S.; Kaelbling, L. P.; and Lozano-Pérez, T. 2012. Collision-free state estimation. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*.

Wong, L. L. S.; Kaelbling, L. P.; and Lozano-Pérez, T. 2013a. Constructing semantic world models from partial views. In *Intl. Symp. on Robotics Research (ISRR)*.

Wong, L. L. S.; Kaelbling, L. P.; and Lozano-Pérez, T. 2013b. Manipulation-based active search for occluded objects. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*.

Wong, L. L. S.; Kaelbling, L. P.; and Lozano-Pérez, T. 2014. Not seeing is also believing: Combining object and metric spatial information. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*.