# Learning the State of the World: Object-based World Modeling for Mobile Manipulation Robots

**Lawson L.S. Wong** (Ph.D. 2016, MIT; now at Brown University; lsw@brown.edu)
DOI: 10.1145/3054837.3054845

Mobile-manipulation robots performing service tasks in human-centric indoor environments have long been a dream for developers of autonomous agents. Tasks such as cooking and cleaning involve interaction with the environment, hence robots need to know about their spatial surroundings. However, service robots operate in environments that are relatively unstructured and dynamic. Mobile-manipulation robots therefore need to continuously perform *state estimation*, using perceptual information to maintain a representation of the state, and its uncertainty, of the world.

By definition, mobile-manipulation robots are capable of moving in and interacting with the world. Hence, at the very least, such robots need to know about the physical occupancy of space and potential targets of interaction (i.e., objects). For the former, there is a long history of representations in the field of navigation and mapping; occupancy grids (Moravec & Elfes, 1985) are a widely-used example. In contrast, object-based representations for robotics are still in their infancy. In my dissertation, I propose a representation based on objects, their 'semantic' attributes (properties such as type and pose), and their geometric realizations in the physical world.

Objects are challenging to keep track of because there is significant *uncertainty* in their states. Object detection and recognition is still far from solved within classical computer vision, and even less so from a robotic vision standpoint. Objects can also be inherently ambiguous because they have the same values for some, or even all, attributes. Besides detection noise, other agents may manipulate objects as well and change object states without informing robots. Compounded over multitudes of objects (thousands or more) and long temporal horizons (days or longer), the above sources of uncertainty give rise to a large and difficult estimation problem.



Figure 1: Mobile-manipulation robots operating in human-centric environments must know about, and be able to model, the world in terms of objects.

### Data Association for Semantic World Modeling from Partial Views

A basic world model could simply use an object detector's output on a single image as a representation of the world. However, doing so suffers from errors such as sensor measurement noise, object occlusion, and detection algorithm approximations. Aggregating measurements across different viewpoints can reduce estimation error. The key challenge in this strategy is *identity management*, induced by measurements that often cannot be uniquely mapped to an underlying object.

I proposed a Bayesian nonparametric clustering approach to data association, inspired by the observation that 'objects' are essentially clusters in joint attribute space. Building on top of Dirichlet-process mixture models (Antoniak, 1974), I incorporated crucial domain assumptions, and used the new model to cluster similar attribute measurements in static scenes. 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 (Wong, Kaelbling, & Lozano-Pérez, 2015). In recent work, I extended the model to a dynamic clustering setting to handle objects states that change over time (Wong, Kurutach, Lozano-Pérez, & Kaelbling, 2016). Figure 2 illustrates the full world modeling problem.

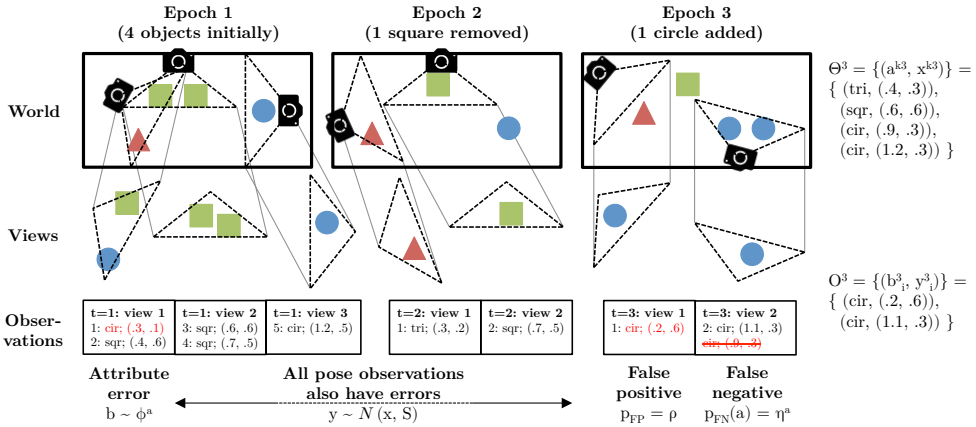


Figure 2: An illustration of the world modeling problem. An unknown number of objects exist in the world (top row), and change in pose and number over discrete epochs. In each epoch, partial views of the world are captured, as depicted by the triangular viewcones. Objects within the viewcones have detectable attributes; in this example, the attributes are shape type (discrete) and 2-D location. The observations are noisy, as depicted by the perturbed versions of viewcones in the middle row. Uncertainty exists both in the attribute values and the existence of objects, as detections may include false positives and negatives (e.g., $t = 3$). The actual attribute detection values obtained from the views are shown in the bottom row; this is the format of input data. Given these noisy measurements as input, the goal is to determine which objects were in existence at each epoch, their attribute values (e.g., $\Theta^3$ in top right), and their progression over time.

### Combining Object and Metric Information

One concept lacking in the above work is 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. This information is handled very naturally in conventional occupancy grids, but grids cannot handle objects elegantly.

The complementary advantages of these two representations inspired a search for a way to maintain estimates of both object and metric information. Because filtering in the joint state is often intractable, I instead adopted the strategy of filtering *separately* in the object and metric spaces by using the previous section's model and occupancy grids. To compensate for the lost dependencies, I then developed a way to *fuse* the filters on demand as queries about either posterior distribution are made (Wong, Kaelbling, & Lozano-Pérez, 2014). Our results suggest that maintaining simple, disparate, and aggressively-factored estimators is potentially superior to keeping a complex joint estimate.

## References

Antoniak, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, *2*(6), 1152-1174.

Moravec, H., & Elfes, A. E. (1985). High resolution maps from wide angle sonar. In *ICRA*.

Wong, L., Kaelbling, L., & Lozano-Pérez, T. (2014). Not seeing is also believing: Combining object and metric spatial information. In *ICRA*.

Wong, L., Kaelbling, L., & Lozano-Pérez, T. (2015). Data association for semantic world modeling from partial views. *The International Journal of Robotics Research*, *34*(7), 1064–1082.

Wong, L., Kurutach, T., Lozano-Pérez, T., & Kaelbling, L. (2016). Object-based world modeling in semi-static environments with dependent Dirichlet-process mixtures. In *IJCAI*.

**Lawson L.S. Wong** is a postdoctoral fellow at Brown University, working with Stefanie Tellex. He completed his Ph.D. in Jan. 2016, advised by Leslie Pack Kaelbling and Tomás Lozano-Pérez. He is currently supported by a Croucher Foundation Fellowship for Postdoctoral Research.