# Object-based World Modeling for Mobile-Manipulation Robots

Lawson L.S. Wong
Brown University
Providence, RI 02912
lsw@brown.edu

***Abstract*—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 relevant aspects of their spatial surroundings. However, service robots typically have little prior information about their environment. Mobile-manipulation robots therefore need to continuously perform the task of state estimation, using perceptual information to maintain a representation of the state, and its uncertainty, of task-relevant aspects of the world. Because indoor tasks frequently require interacting with objects, objects should be given critical emphasis in spatial representations for service robots. This paper proposes a spatial representation based on objects, their 'semantic' attributes (task-relevant properties such as type and pose), and their geometric realizations in the physical world.**

## I. Introduction

Indoor tasks frequently require interacting with objects, hence objects should be given critical emphasis in spatial representations for service robots. Compared to occupancy grids and feature-based maps that have been used traditionally in navigation and mapping, object-based representations for mobile-manipulation robots are still in their infancy. 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).

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.

In this paper, I describe a world model based on objects, their 'semantic' attributes (task-relevant properties such as type and pose), and their geometric realizations in the physical world. High-level results for estimating this 'world state' are shown; details on problem formulations, inference algorithms and derivations, and other experimental results can be found in the author's doctoral dissertation [30].

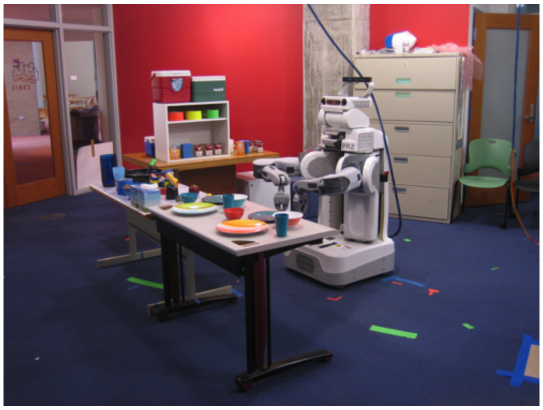


Fig. 1: A Willow Garage PR2 mobile-manipulation service robot operating in a typical laboratory environment. What does it need to know about the world in order to perform tasks intelligently?

## II. Background and Context

Understanding the mobile robot's spatial environment, by deriving a world model from its sensors, has long been a problem of interest to the robotics community [4, 2]. Early work typically focused on using ultrasonic range sensors, tracking low-level linear, planar, and corner features as landmarks in a map [3]. The field of simultaneous localization and mapping (SLAM) soon took off, producing metric maps for mobile robot navigation, and SLAM modules are currently widely available on all mobile robot platforms [9]. However, since most robots did not have manipulators, their main task was navigation, which in the indoors setting mainly requires knowledge of obstacles such as walls and furniture. Objects were typically thought of as being nuisance entities that should be removed from the map during post-processing [27].

It was clear eventually that in addition to features and occupancy information, human-centric concepts were also necessary in maps since in many applications humans need to interact with the maps. For example, the navigation task of "go to the kitchen, then to Alice's office" requires knowledge of which place in the map is the kitchen, which places are offices, and also who occupies which office. These human-centric concepts were referred to as "semantic knowledge", and thus the field of semantic mapping was born [12]. Most work in this area built off of the work in SLAM, and were conceived as providing labels for entities in existing metric/topological maps [6]. Occasionally, these labels themselves had additional abstractions on top to encode higher-level knowledge, such as the fact that kitchens and offices are rooms. Finding the appropriate labels for metric regions / topological nodes is often formulated as a classification problem [22]. The most common classification task for indoor semantic mapping is to recognize a certain region is a room, and to determine its type (kitchen, office, etc.). Kostavelis and Gasteratos [11] provides a recent survey on semantic mapping.

One intuitive cue for recognizing places is objects. For example, detecting a computer keyboard in a room indicates that the room type is most likely "office" and not "kitchen". This was first recognized by Ekvall et al. [5], and was subsequently incorporated in many semantic mapping works [e.g., 23, 28, 35, 19, 21]. In these applications, images (typically 2-D), and potentially the objects found by object detectors applied to the images, acted as the labels for the underlying metric/topological map. As the authors of those and numerous other works have found, objects are a very useful cue for determining the place category, to the point that Ranganathan and Dellaert [23] suggested we can basically use objects as a basic unit of representation to model indoor places (instead of mid-level geometric or visual features).

However, in the current generation of semantic maps, there is no fundamental *representation* of the objects, in the sense that one could not ask, for example, "How many keyboards are there in the room?". Consider a room where there were actually two keyboards, each of which were detected in five images taken in the room (i.e, ten detections in total). Current semantic maps would take that as strong evidence that the room is of type "office". However, they may interpret the detections as indicating that there are ten keyboards in the room, or, more typically, make no interpretation at all. The reason for this is that such interpretation is *unnecessary* for navigation and place recognition, as evidenced by recent work involving object information, but without explicit recognition.

Similarly, the recent success of dense 3-D reconstruction has led to the suggestion that dense surface maps and point clouds are also a viable representation of space [e.g., 18, 29]. In these dense maps, each point in a point cloud or surface element in a surface reconstruction is endowed with a semantic label. The resulting reconstructions are very visually-appealing and have superior resolution. However, they are limited to visual sensors, require smooth frame transitions, and often are computationally intensive to process. Nevertheless, they have great potential as a fine representation of space.

*Regardless of the map's representation, a map annotated with object detections does not equate to object-level understanding.* In mobile-manipulation tasks, we need to understand the objects themselves. There is a fundamental difference between obtaining ten keyboard detections and reporting "office", versus identifying that there are two keyboards, determining which one to pick up, localizing the target to sufficient accuracy for robust manipulation, and truly understanding the functional properties of a keyboard. I argue that semantic maps today are not sufficient for mobile-manipulation tasks, which require precise knowledge about *object states*, including information that may not be visible (but can be inferred from other object-based knowledge). *I cannot cook with a reconstructed cloud of points labeled "wok"; I cook with a wok.* Ultimately, *recognition* must be part of the pipeline.

## III. Object-based World Modeling

Object state estimation and world modeling considers the acquisition and maintenance of knowledge beyond the point of individual object detections. 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. Additionally, the geometric shape models of objects are tracked as special attributes, used to determine their physical occupancy and realization in the world, thus providing information about feasible motions. In this paper, I consider the problem of estimating, filtering, and combining both forms of information.

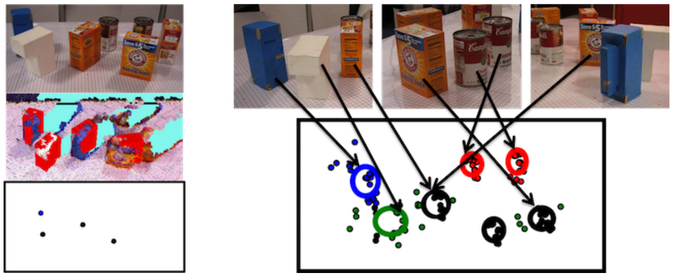
To measure object states, we rely on attribute detectors, particularly ones operating on 3-D visual data. Object recognition and pose estimation has received widespread attention from the computer vision and robotics communities. With the recent advances in RGB-D cameras, several systems have been developed to detect object types/instances and their 6-D poses from 3-D point clouds [e.g, 25, 8, 13, 1, 16]. This paper uses the one in Glover and Popovic [7] as the black-box attribute detector, but the methods developed in this paper are agnostic to the detector used. Also, although I mainly focus on object type-and-pose estimation, this was only chosen as a concrete and familiar proof-of-concept application; the approach can generalize to other semantic attributes and tasks.

A basic world model could simply use a detector's output on a single image as a representation of the world. However, doing so suffers from many sources of error: sensor measurement noise, object occlusion, and modeling and approximation errors in the detection algorithms. Aggregating measurements across different viewpoints, as illustrated in Figure 2, can improve coverage and help reduce estimation error.

The primary challenge in aggregating object detections across multiple views of the world is *identity management*, induced by the fact that measurements often cannot be uniquely mapped to an underlying object. Tackling this data association problem in static scenes is described next. From there, I consider how to aggregate detections across time as well, with the added difficulty that the world may change over time. Finally, I consider how to integrate this object-based representation with traditional spatial representations such as occupancy grids, thereby aggregating information across different sensing modalities and representations.

### A. Semantic World Modeling from Partial Views

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



(a) Single viewpoint

(b) Aggregation of object detections from multiple viewpoints

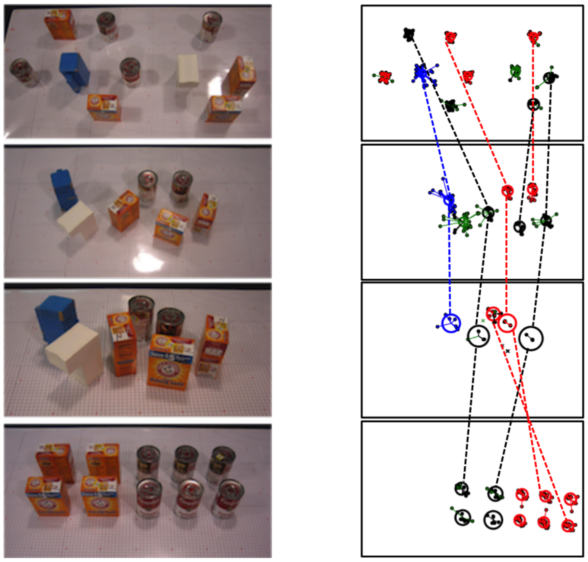
Fig. 2: **(a)** Given a tabletop scene (top), we want to estimate the types and poses of objects in the scene using a black-box object detector. From a single RGB-D image, however, objects may be occluded or erroneously classified. In the rendered image (middle; detections superimposed in red), three objects are missing due to occlusion, and the bottom two objects have been misidentified. The semantic attributes that result in our representation are very sparse (bottom; dot location is measured 2-D pose, color represents type). A single viewpoint is insufficient to identify all objects in a scene correctly. **(b)** Aggregation of measurements from many different viewpoints (top) is therefore needed to construct good estimates. However, this introduces data association issues of the type addressed in this work, especially when multiple instances of the same object type are present. From all the object detection data, as shown (bottom) by dots (each dot is one detection), our goal is to estimate the object types and poses in the scene (shown as thick ellipses centered around location estimate; color represents type, ellipse size reflects uncertainty). The estimate above identifies all types correctly with minimal pose error.

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 (see Figure 2).

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 points in joint attribute space, and observations across different viewpoints form clusters centered at these points [33]. Given attribute detections from multiple viewpoints, this algorithm outputs a distribution (in the form of samples) over hypotheses of object states, where a hypothesis consists of a list of objects and (distributions of) their attribute values. The number of existing objects is not known *a priori* and is also estimated by the Bayesian nonparametric model.

Figures 2 and 3 show the result of this approach, where the thick ellipses are the hypothesized clusters (objects), and they are typically centered at groups of small dots, which show the actual noisy measurements returned from the object detector.

### *B. World Modeling in Semi-Static Environments with Dependent Dirichlet Process Mixtures*



(a) Sequence of static scenes from Wong et al. [33]

(b) Most-likely object configurations and trajectories

Fig. 3: Inferring object states across different viewpoints *and* times. A concatenated sequence of scenes (epochs) is shown from top to bottom. In each scene, the world is assumed to be static, and detections from multiple viewpoints are aggregated using methods from Section III-A. Between scenes, possible changes in object states are considered using methods from Section III-B. The most-likely inferred clusters and tracks are shown on the right.

Our operational definition of an “object” is an entity that can be subject to manipulation, and a typical effect of manipulation is some change in the object state (e.g., its pose). Hence the semantic world model for static worlds from the previous section is insufficient. At any given time, however, most objects are not being manipulated (e.g., books on a bookshelf, a home when its occupants are at work). We view indoor environments as being static at most times, changing only at discrete events. An example of this may be a cleaning robot that is turned on every time a home’s occupants leave. The state of the world will likely have changed from the previous day, but during the operational period of the robot, the world is essentially static (possibly except for the robot’s own manipulation actions, which it can track more easily). We denote such environments as being “semi-static”. Figure 4 illustrates the semi-static semantic world modeling problem, allowing for objects to change state over time (epochs).

The way I have posed the data association problem is reminiscent of multiple object tracking (MOT) problems [14], which has been well-studied in the computer vision and target-tracking communities. Indeed, conventional approaches such as multiple hypothesis tracking (MHT) [24] and more recent batch methods such as Markov-chain Monte Carlo data association (MCMCDA) [20] can be applied directly to our problem. However, we can exploit the (semi-)static nature of our data to reap great computational gains, as we demonstrate against MHT in the static case [33]. Moreover, the semi-static nature of the data degrades the performance of MCMCDA.

I extended the clustering-based approach for static semantic world modeling to allow clusters to change over time,

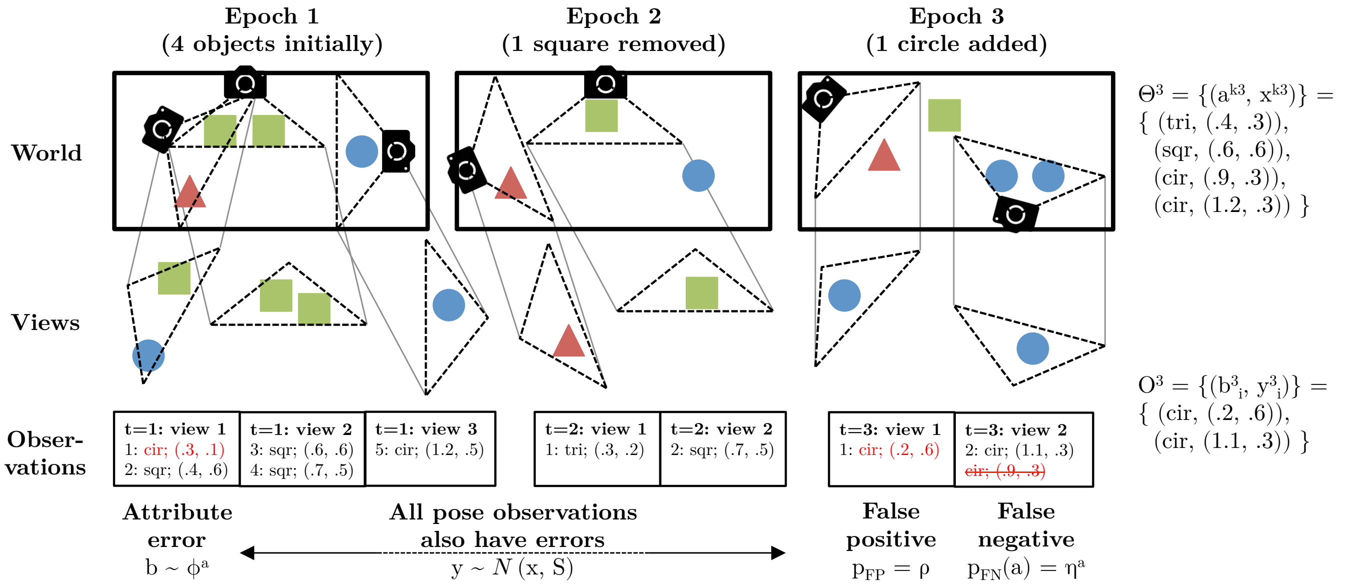


Fig. 4: An illustration of the world modeling problem. An unknown number of objects exist in the world (top row), and change in pose and number over time (world at each epoch enclosed in box). At each epoch, limited views of the world are captured, as depicted by the triangular viewcones. Within these viewcones, objects and their attributes are detected using black-box perception modules (e.g., off-the-shelf object detectors). 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 ("Observations"); 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.

and developed novel inference algorithms that can efficiently achieve superior performance in semi-static environments [34]. In the static case, we used a Dirichlet process mixture model (DPMM) to cluster observations into a list of object attributes. To capture changes across epochs, we now use a *dependent* DPMM, which is a natural generalization of DPMMs to the time-varying case [15]. An example inferred sequence of world states over multiple epochs is shown in Figure 3.

### C. Combining Object and Metric Spatial Information

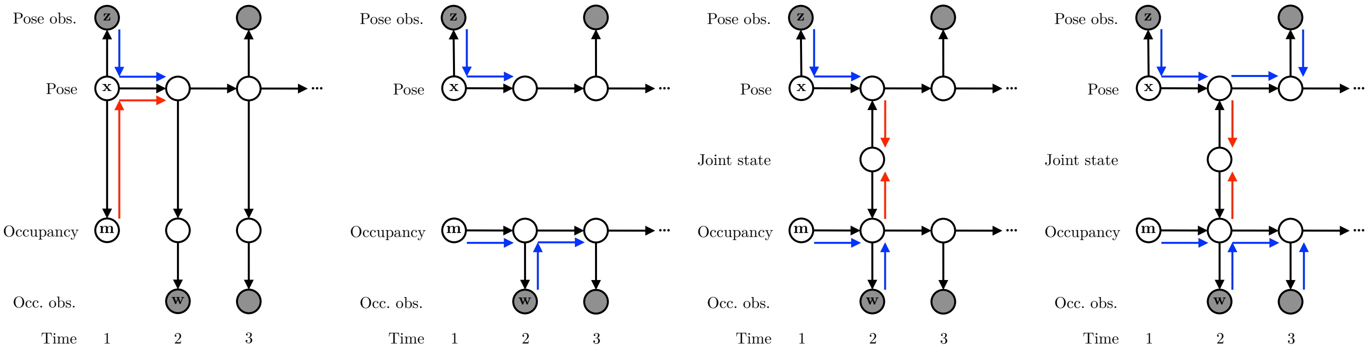
Alas, not all things in the world are objects and attributes. One concept that was lacking in the above work was the notion that objects occupy physical regions of space. The concept of free space, regions that no object overlaps, was also only implicitly represented. It is therefore difficult, in the 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, but grids cannot incorporate the concept of 'objects' (besides representing them as a collection of cells).

Since Moravec and Elfes [17] pioneered the occupancy grid model of space, occupancy grids have been used extensively in robotics, most notably in mapping. These maps have paved the way for tasks such as navigation and motion planning, in which knowledge of free and occupied spaces is sufficient for success. However, as we move to tasks that require richer interaction with the world, such as locating and manipulating objects, occupancy information alone is insufficient.

In the mapping community, there has been recognition that using metric representations only is insufficient. In particular, the rise of topological mapping, and the combination of the two in hybrid metric-topological mapping [26] suggests the utility of going beyond metric representations. These hybrid representations have been successfully applied in tasks such as navigation [10]. In the related field of semantic mapping topological information is typically extracted from metric layers (occupancy grids). These existing examples suggest that a hybrid object-metric representation is worth exploring.

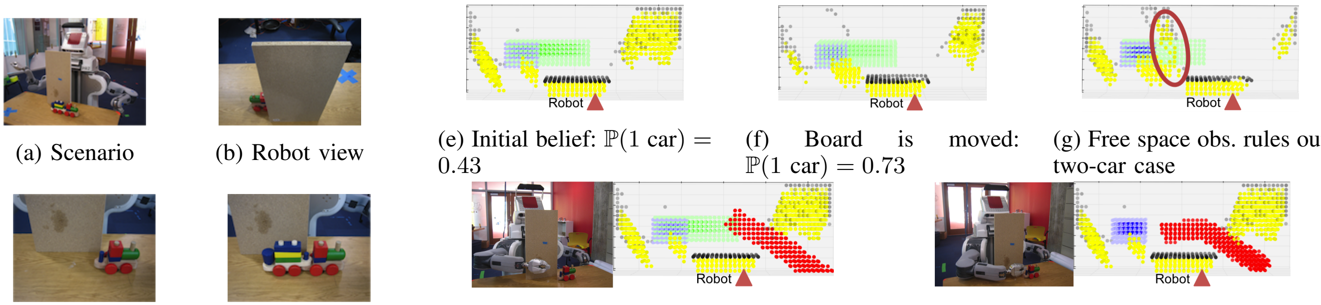
Returning to object attributes and occupancy grids, the complementary advantages of these two representations inspired a search for a way to maintain filters of both object and metric information. Because filtering in the joint state involves complex dependencies and is intractable, I instead adopted the strategy of filtering *separately* in the object and metric spaces by using the existing filters. 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. After the query, the estimate is *discarded* and filtering proceeds in the separate spaces to maintain tractability [32]. Figure 5 illustrates this *factor-fuse-forget* filtering framework.

Combining object-level and metric-level information is useful, as depicted in the example in Figure 6. In particular, I



(a) Standard filter (b) Factor (c) Fuse (d) Forget

Fig. 5: The *factor-fuse-forget* filtering framework, applied to estimating a hybrid object-metric representation of the world state over time.
**(a)** Consider the problem of filtering objects' poses ($x$), given observations of both object poses ($z$) and occupancy ($w$). If both object poses and their observations have normally-distributed errors, then filtering is efficient, as annotated by the blue arrows. Occupancy observations can provide additional evidence for an object's presence (if the overlapping cells are occupied) or absence (if free space is detected). However, because the occupancy state ($m$) and its observations are not conjugate to the object pose, incorporating them during filtering is computationally expensive (e.g, requiring nonparametric / particle-based methods); such inefficiency is indicated by the red arrows.
**(b)** Instead of forcing all observations into a single representation, we should leverage the complementary advantages of each. Poses with Gaussian observation errors can be tracked with a Kalman filter. Occupancy states with binary observation values can be tracked with a (dynamic) occupancy grid. We propose to keep *both* filters around and keep them factored/separated to maintain tractability.
**(c)** Occasionally, we would like to combine both sources of information. Fusion of the two filtered estimates into a single joint state (posterior distribution) is allowed on-demand, and simply applies Bayes' rule, but is computationally expensive in general due to non-conjugacy.
**(d)** After obtaining a fused joint state estimate, one could continue filtering in the joint space. However, this reduces to the original problem (a), and is undesirable. Instead, we choose to forget the joint estimate and continue filtering in the *factored* spaces to maintain tractability.



(a) Scenario (b) Robot view (c) Is it 1 car? (d) Or 2 cars?

(e) Initial belief: $\mathbb{P}(1 \text{ car}) = 0.43$ (f) Board is moved: $\mathbb{P}(1 \text{ car}) = 0.73$ (g) Free space obs. rules out two-car case

(h) Arm moves towards object: $\mathbb{P}(1 \text{ car}) = 0.44$ (i) Arm 'overlaps' second car, rules out two-car hypothesis

Fig. 6: A 3-D demonstration on a PR2 robot. Plots show occupancy grids with 1m × 0.4m × 0.2m volume, containing $10^4$ cubes of side length 2cm, with the final (vertical) dimension projected onto the table. Colors depict occupancy type/source: Yellow = free space observation; Black = occupancy observation; Blue = inferred occupancy from one-car train; Green = inferred occupancy from two-car train; Red = occupied by robot in its current state. In this projection, the robot is situated at the bottom center of the plot, facing 'upwards'; the black line observed near the bottom corresponds to the board.
**(a)-(b)** A toy train is on a table, but only part of the front is visible to the robot.
**(c)-(d)** This is indicative of two possible scenarios: the train has one car or two cars; there is in fact only one car.
**(e)-(g)** One way to determine the answer is to move the occluding board away. This reveals free space where the second car would have been (circled in (e)), hence ruling out the two-car case.
**(h)-(i)** Another way is to use the robot arm. If the arm successfully sweeps through cells without detecting collision, the cells must have originally been free and are now occupied by the arm. Sweeping through where the second car would have been therefore eliminates the possibility of the train being there. Video: http://lis.csail.mit.edu/movies/ICRA14_1678_VI_fi.mp4

identified two ways in which fusion is particularly informative: free space detections strongly indicate that objects cannot be positioned in such regions (Figure 6g), and object pose detections can be used to infer that overlapping occupancy cells must be occupied. By considering the hypothetical occupancy induced by objects and observing occupancy information that is inconsistent, hypotheses about objects' attributes can be ruled out, and uncertainty reduced.

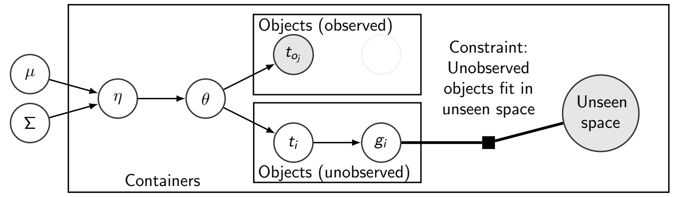


Fig. 7: A probabilistic model for inferring locations of unseen objects, using object-object co-occurrence information and capacity constraints [31]. Instead of having to custom-build such models and inference procedures, the world model described in this paper can provide the same information automatically.

## IV. Discussion

I have argued that object-based world models are needed for mobile-manipulation robots operating in typical indoor environments, and that existing spatial representations such as feature-based SLAM and semantic maps will not be sufficient for mobile manipulators that wish to reason and act intelligently in the world. I presented three aspects of the world modeling problem, where the common technical thread was that information had to be aggregated in some manner; respectively, across space, time, and sensor modalities. The key challenge in accomplishing this fusion is a data association problem, and a potential mismatch between different state and belief representations. The key contributions were a new clustering-based batch data association paradigm using Bayesian nonparametric models, and an unconventional factor-fuse-forget filtering framework.

### *A. What can we do now?*

Many robotics tasks require inferring posterior distributions over object states, both in terms of attributes and geometry. The world model is a first step towards performing such probabilistic models and inference algorithms generically. For example, in previous work on object search, I developed a custom graphical model for inferring potential locations of unseen objects, using object-object co-occurrence information and capacity constraints (see Figure 7). This model was created based on two intuitions: objects of similar types tend to be co-located, and large unseen objects cannot fit in small unexplored regions. This same model and inference procedure can naturally come out of our current general world model. In particular, using our world model, we should be able to track the poses of many objects over time, and from this data infer the likely locations and possibly co-occurrence statistics of object pairs. Then, together with an occupancy grid representing our current explored regions, we should be able to fuse this information with object pose predictions to automatically deduce capacity constraints. These inferences can then guide the robot to take useful manipulation actions, possibly even predicting what might appear after an object is moved away and unexplored space is revealed.

More generally, the world model estimates the state of the world and tracks our uncertainty (with respect to object attributes and occupancy). Since this is a type of belief about the world, we should be able to use it to guide actions; connecting to an action selection strategy / planning algorithm is clearly a next step. For example, we can perform the usual next-best-view information gathering actions.

Another interesting direction is to use the world model to track our state of *ignorance* in the world's objects – for example, many occupied cells in an occupancy grid may not be overlapping any known objects in the world model. This may indicate that a collection of unidentified cells may correspond to an unknown object type, and the robot can then attempt to *learn* about novel objects. For any world model to be practical in human environments, I envision that it must have the capability to adapt and grow its representation.

### *B. How might we go further?*

The point above raises an interesting counterpoint: world models, and estimators in general, cannot *just* grow in size. Even with the proof-of-concept domains shown in this dissertation, inference was already non-trivial. Consider the scenario shown in Figure 1: there are many more objects, additional attributes of interest, large regions of space, compounded over long time horizons. Attempting to directly scale up our current model to a problem of this size is hopeless. Instead, I argue that world models and estimators must also have the capability to *compress* its representation, by aggressively 'forgetting' information, or simply ignoring it.

Aggressively forgetting is in the same vein as heavy pruning in methods such as the multiple hypothesis tracker (MHT). However, we also know that simply discarding information permanently without maintaining a sufficient statistic eventually leads to errors. Since computational tractability demands a pruning strategy, perhaps what we need instead is a *recovery* mechanism, triggered by a *fault diagnosis/identification* process. For example, for online purposes, we filter incoming information using an aggressively-pruned MHT, but also store a significantly longer historical snapshot. If, through some inference procedure or failed-assertions triggers, a *meta-estimator* comes to believe that the MHT has diverged from the true state, then the stored history is consulted and used to reset the filter, possibly by using more robust but slower solutions such as clustering-based batch data association discussed in Sections III-A and III-B.

Perhaps even harder, but certainly necessary eventually, is allowing estimators to learn to ignore certain spaces of information. This is based on the recognition that for any given task, most of the world state is typically irrelevant. For example, we can function properly in my office without worrying about the state of things in my home. Fundamentally:

**Estimators, including world models,**
**must be tied to the task.**

In the long run, since the task changes, what we need is a method for agents to automatically construct, adapt, and reconfigure estimators that can be used in different tasks. Choice of state representation and estimator, conventionally a privilege and responsibility that is solely granted to human system designers, should be made available to all intelligent agents, natural or artificial.

# References

[1] A. Aldoma, F. Tombari, J. Prankl, A. Richtsfeld, L. Di Stefano, and M. Vincze. Multimodal cue integration through hypotheses verification for RGB-D object recognition and 6DOF pose estimation. In *IEEE International Conference on on Robotics and Automation*, 2013.

[2] W. Burgard and M. Hebert. World modeling. In B. Siciliano and O. Khatib, editors, *Springer Handbook of Robotics*, pages 853–869. Springer-Verlag Berlin Heidelberg, 2008.

[3] I.J. Cox and J.J. Leonard. Modeling a dynamic environment using a Bayesian multiple hypothesis approach. *Artificial Intelligence*, 66(2):311–344, 1994.

[4] J.L. Crowley. Dynamic world modeling for an intelligent mobile robot using a rotating ultra-sonic ranging device. In *IEEE International Conference on Robotics and Automation*, 1985.

[5] S. Ekvall, D. Kragic, and P. Jensfelt. Object detection and mapping for service robot tasks. *Robotica*, 25(2):175–187, 2007.

[6] C. Galindo, A. Saffiotti, S. Coradeschi, P. Buschka, J.A. Fernández-Madrigál, and J. González. Multi-hierarchical semantic maps for mobile robotics. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2005.

[7] J. Glover and S. Popovic. Bingham Procrustean alignment for object detection in clutter. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013.

[8] J. Glover, R.B. Rusu, and G. Bradski. Monte Carlo pose estimation with quaternion kernels and the Bingham distribution. In *Robotics: Science and Systems*, 2011.

[9] E. Herbst and D. Fox. Mapping as a service. In *AAAI Conference on Artificial Intelligence Workshop on Intelligent Robotic Systems*, 2013.

[10] K. Konolige, E. Marder-Eppstein, and B. Marthi. Navigation in hybrid metric-topological maps. In *IEEE International Conference on Robotics and Automation*, 2011.

[11] I. Kostavelis and A. Gasteratos. Semantic mapping for mobile robotics tasks: A survey. *Robotics and Autonomous Systems*, 66:86–103, 2015.

[12] B. Kuipers. The spatial semantic hierarchy. *Artificial Intelligence*, 119(1–2):191–233, 2000.

[13] K. Lai, L. Bo, X. Ren, and D. Fox. Detection-based object labeling in 3D scenes. In *IEEE International Conference on on Robotics and Automation*, 2012.

[14] W. Luo, J. Xing, X. Zhang, X. Zhao, and T.-K. Kim. Multiple object tracking: A literature review. *ArXiv e-prints*, 2014.

[15] S.N. MacEachern. Dependent Dirichlet processes. Technical report, Ohio State University, 2000.

[16] Z.-C. Marton, F. Balint-Benczedi, O.M. Mozos, N. Blodow, A. Kanezaki, L.C. Goron, D. Pangercic, and M. Beetz. Part-based geometric categorization and object reconstruction in cluttered table-top scenes. *Journal of Intelligent & Robotic Systems*, pages 1–22, 2014.

[17] H. Moravec and A.E. Elfes. High resolution maps from wide angle sonar. In *IEEE International Conference on Robotics and Automation*, 1985.

[18] R.A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A.J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. In *IEEE International Symposium on Mixed and Augmented Reality*, 2011.

[19] A. Nüchter and J. Hertzberg. Towards semantic maps for mobile robots. *Robotics and Autonomous Systems*, 56(11):915–926, 2008.

[20] S. Oh, S. Russell, and S. Sastry. Markov chain Monte Carlo data association for multi-target tracking. *IEEE Transactions on Automatic Control*, 54(3):481–497, 2009.

[21] A. Pronobis and P. Jensfelt. Large-scale semantic mapping and reasoning with heterogeneous modalities. In *IEEE International Conference on Robotics and Automation*, 2012.

[22] A. Pronobis, O.M. Mozos, B. Caputo, and P. Jensfelt. Multi-modal semantic place classification. *The International Journal of Robotics Research*, 29(2–3):298–320, 2010.

[23] A. Ranganathan and F. Dellaert. Semantic modeling of places using objects. In *Robotics: Science and Systems*, 2007.

[24] D.B. Reid. An algorithm for tracking multiple targets. *IEEE Transactions on Automatic Control*, 24(6):843–854, 1979.

[25] R.B. Rusu, G. Bradski, R. Thibaux, and J. Hsu. Fast 3D recognition and pose using the viewpoint feature histogram. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010.

[26] S. Thrun. Learning metric-topological maps for indoor mobile robot navigation. *Artificial Intelligence*, 99(1):21–71, 1998.

[27] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. MIT Press, 2005.

[28] S. Vasudevan, S. Gächter, V. Nguyen, and R. Siegwart. Cognitive maps for mobile robots – an object based approach. *Robotics and Autonomous Systems*, 55(5):359–371, 2007.

[29] T. Whelan, M. Kaess, H. Johannsson, M. Fallon, J.J. Leonard, and J. McDonald. Real-time large-scale dense RGB-D SLAM with volumetric fusion. *The International Journal of Robotics Research*, 34(4–5):598–626, 2015.

[30] L.L.S. Wong. *Object-based World Modeling for Mobile-Manipulation Robots*. PhD thesis, Massachusetts Institute of Technology, 2016.

[31] L.L.S. Wong, L.P. Kaelbling, and T. Lozano-Pérez. Manipulation-based active search for occluded objects. In *IEEE International Conference on Robotics and Automation*, 2013.

[32] L.L.S. Wong, L.P. Kaelbling, and T. Lozano-Pérez. Not seeing is also believing: Combining object and metric spatial information. In *IEEE International Conference on Robotics and Automation*, 2014.

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

[34] L.L.S. Wong, T. Kurutach, L.P. Kaelbling, and T. Lozano-Pérez. Object-based world modeling in semi-static environments with dependent Dirichlet-process mixtures. In *International Joint Conference on Artificial Intelligence*, 2016. To appear.

[35] H. Zender, O.M. Mozos, P. Jensfelt, G.-J.M. Kruijff, and W. Burgard. Conceptual spatial representations for indoor mobile robots. *Robotics and Autonomous Systems*, 56(6):493–502, 2008.