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

Lawson L.S. Wong

## 1 Introduction

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 typically involve interaction with the environment, hence robots need to know relevant aspects of their spatial surroundings. However, service robots typically have little prior information about their environment, unlike industrial robots in structured environments. Even if this information was given initially, due to the involvement of other agents (e.g., humans adding/moving/removing 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 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. Compared to occupancy grids and feature-based maps that have been used traditionally in navigation and mapping, object-based representations 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). In this thesis, I propose a representation based on objects, their 'semantic' attributes (task-relevant 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.

In previous investigations, I have developed estimators for the types, poses, and occupancy of objects in a static world from noisy perception. In the remainder of the thesis, I will explore:

- Extensions to increase the representation's *expressiveness*. Object attributes that change over time, possibly without the robot's control or constant observation, need to be tracked. Estimation under state constraints (e.g., contact, collision) will also be incorporated.

- Approaches that make the representation *scalable*. Aggressive independence assumptions and factoring approximations are needed to maintain tractability of the state estimator. Even this may not be sufficient; an 'attention mechanism' that focuses on task-relevant aspects of the state, and ignores or delays processing for other features, will be explored.

Finally, a state estimator with these characteristics will be demonstrated on a PR2 robot.

# 2 Background

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 (Crowley, 1985). Early work typically focused on using ultrasonic range sensors, tracking low-level planar and corner features as landmarks in a map (Cox and Leonard, 1994). The resulting geometric maps were useful for mobile robot navigation, but the features are too low-level for modeling objects.

For mobile-manipulation robots that operate on objects, the world model must contain information about object states. With the advent of more effective visual sensors, image features, and object detectors, world models are now capable of supporting richer representations of objects. For example, Ranganathan and Dellaert (2007) demonstrated that it is useful and natural to model indoor places using objects as the basic unit of representation. However, like much of the related work in semantic mapping (Vasudevan et al., 2007; Zender et al., 2008; Nüchter and Hertzberg, 2008; Pronobis and Jensfelt, 2012), the ultimate goal was place recognition for navigation tasks.

Recently, there has also been growing interest in world modeling involving object information, but without explicit recognition. Anati et al. (2012) showed that object-based robot localization is still possible even if 'soft' heatmaps of local image features are used instead of explicit object poses. The recent success of dense 3-D reconstruction (Newcombe et al., 2011; Whelan et al., 2012, 2014) has also led to dense surface maps being a viable representation of space.

I argue that these representations are not sufficient for mobile-manipulation tasks, which require precise knowledge about *object states*, including information that may not be visible. I cannot cook with a collection of localized SIFT features or a reconstructed cloud of points; I cook with a wok. Ultimately, *recognition* must be part of the pipeline. Object state estimation, the focus of this thesis, considers the acquisition and maintenance of knowledge beyond the point of recognition.

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 (Rusu et al., 2010; Glover et al., 2011; Lai et al., 2012; Aldoma et al., 2013; Marton et al., 2014). I have used one such detector (Glover and Popovic, 2013) as the black-box attribute detector, but the methods developed in this thesis are agnostic to the detector used.

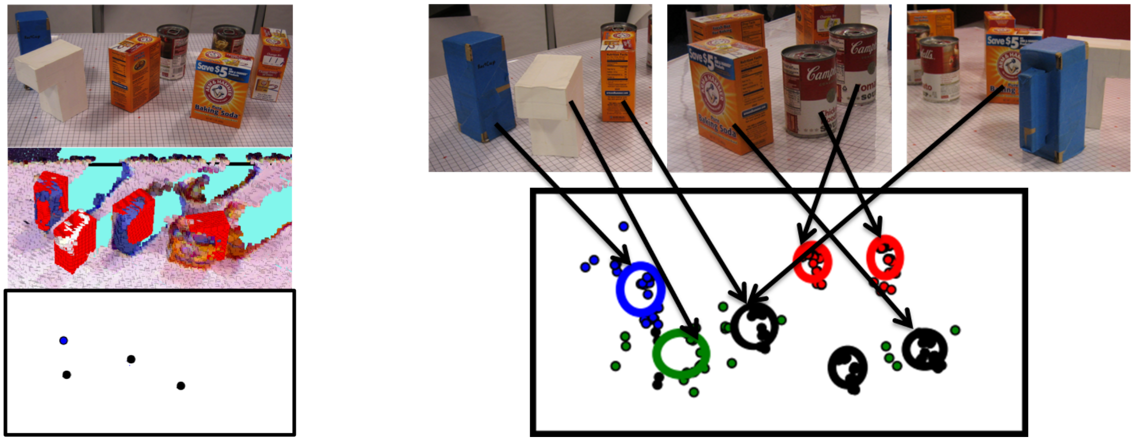
A basic world model could simply use a detector's output on a single image as a representation of the world. However, doing so suffers from many sources of error: sensor measurement noise, object occlusion, and modeling and approximation errors in the detection algorithms. Aggregating measurements across different viewpoints can help reduce estimation error. For example, Hager and Wegbreit (2011) demonstrate the utility of considering a prior 3-D scene model and its potential evolution over scenes. Using this observation as a premise, active perception approaches (e.g., Eidenberger and Scharinger, 2010; Velez et al., 2012; Atanasov et al., 2013) seek the next best view (camera pose) where previously-occluded objects may be visible, typically by formulating the problem as a partially-observable Markov decision process. Because the focus is on planning instead of estimation, active perception is complementary to the world modeling problem, which considers estimation using measurements from an arbitrary collection of camera poses.

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 forms the first part of my thesis. The second part of my thesis considers how to integrate this object-based representation with traditional spatial representations such as occupancy grids. Finally, I propose to investigate mechanisms that will make the developed estimation approaches scalable to large domains over long time horizons.

## 3 Completed investigations

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.

### 3.1 Semantic World Modeling from Partial Views (Wong et al., 2013b)



**(a)** Single viewpoint

**(b)** Aggregation of object detections from multiple viewpoints

**Figure 1:** **(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). **(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 error in pose.

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 (see Figure 1).

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 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.

### 3.1.1 Related work

The data association problem was historically motivated by target tracking; Bar-Shalom and Fortmann (1988) provide a comprehensive overview of the foundations, as well as coverage of greedy nearest-neighbor methods and an approximate Bayesian filter, the joint probabilistic data association filter (JPDAF). Apart from being a suboptimal approximation, the JPDAF is also limited by its assumption of a fixed number of tracked targets (objects), which is not valid for our problem.

A more principled approach when the number of tracks is unknown is multiple hypothesis tracking (MHT) (Reid, 1979). In principle, MHT considers the tree of all possible association hypotheses, branching on the possible tracks that each measurement can correspond to. However, due to the number of measurements involved, maintaining the entire tree (and hence the exact posterior distribution) is exponentially expensive and intractable for any non-trivial branching factor. As a result, practical implementations of MHTs must use one of many proposed heuristics (e.g., Kurien, 1990; Cox and Hingorani, 1996), typically pruning away all but the few most-likely branches in the association tree. Aggressive pruning potentially removes correct associations that happen to appear unlikely at the moment. Although this problem is somewhat mitigated by postponing ambiguous associations through delayed filtering, the window for resolving issues is short because of computational limitations. In the context of semantic world modeling, Elfring et al. (2013) identified the presence of data association issues, and applied an MHT approach.

The MHT pruning heuristics were necessitated by the combinatorial complexity of MHT, which in turn is due to the enumeration of all possible association histories. Instead of attempting to evaluate every point in this large space, most of which contains little probability mass, efficient sampling techniques have been proposed that try to only explore high-probability regions. Markov-chain Monte Carlo (MCMC) methods for sampling association matchings and tracks have been explored by Dellaert et al. (2003) for structure-from-motion and by Pasula et al. (1999) for traffic surveillance. More recently, Oh et al. (2009) generalized the latter work by considering a wider class of transition moves during sampling, and provided theoretical bounds on the mixing (convergence) time of their sampling algorithm, MCMCDA. Because only a small space of likely associations is frequently sampled, and all measurement associations are repeatedly considered (unlike MHT with pruning), MCMCDA empirically outperforms MHT both in efficiency and accuracy, especially in environments with heavy detection noise.

### 3.1.2 Contributions

In the context of previous work, I view my approach as building on the semantic world modeling problem formulation of Elfring et al. (2013) and the data association techniques of Oh et al. (2009). As argued above and by Oh et al. (2009), MHT has various drawbacks, which are directly inherited by the approach of Elfring et al. (2013). However, instead of directly applying MCMCDA to world modeling, I will introduce more domain assumptions to make inference more efficient.

Unlike target tracking, for which most data association algorithms are designed, semantic world modeling has three distinguishing domain characteristics:

- Objects can have attributes besides location, and hence are distinguishable from each other in general (which likely makes data association easier). Some data association methods can be readily generalized to this case (as was done by Elfring et al., 2013), but it excludes some from consideration, such as the probability hypothesis density (PHD) filter by Mahler (2007).

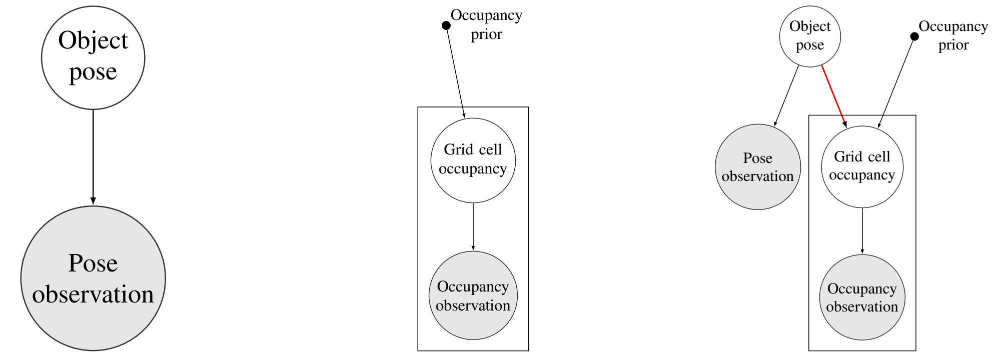
- Only a small region of the world is visible from any viewpoint. Most data association methods operate in regimes where all targets are sensed (possibly with noise/failure) at each time point.

- Most object states do not change over short periods of time.

In light of the final point, I studied the semantic world modeling problem under the stringent assumption that the world is static, i.e., object states do not change. This does not trivialize the data association problem, since it is still necessary to determine measurement-to-object correspondences (and is exacerbated by the limited field of view). However, target-tracking algorithms no longer seem most appropriate, since time is no longer an essential dimension. Instead, the problem becomes more akin to *clustering*, where objects are represented by points in the joint attribute (product) space, and measurements form clusters around these points.

A useful model for performing clustering with an unbounded number of clusters is the Dirichlet process mixture model (DPMM) (Antoniak, 1974; Neal, 2000), a Bayesian nonparametric approach that can be viewed as an elegant extension to finite mixture models. I applied this method to world modeling and derived a Gibbs sampling algorithm to perform inference. The sampling candidate proposals in this algorithm can be viewed as a subset of those considered by Oh et al. (2009). However, clustering ignores a crucial assumption in data association, which led me to introduce modifications and approximations to address this issue.

## 3.2 Combining Object and Metric Spatial Information (Wong et al., 2014a)

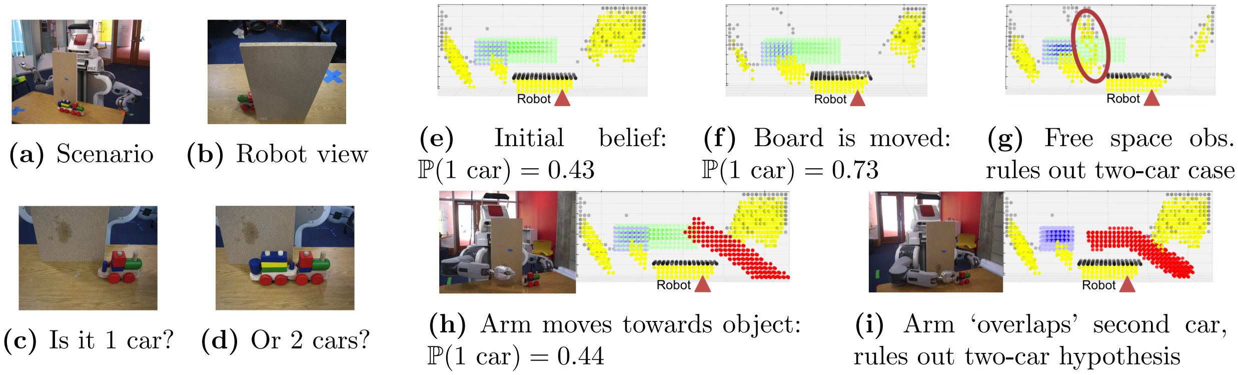


**(a)** Object-level **(b)** Metric-level **(c)** Information fusion

**Figure 2:** Graphical model depiction of object and metric spatial information. **(a)** Object attributes such as pose are detected and tracked using the methods described in Section 3.1. **(b)** Occupancy observations are used separately to maintain a standard occupancy grid (Thrun et al., 2005). **(c)** Combining these two sources of information is useful (see Figure 3); I derived an inference procedure to link the two (red arrow).

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).

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.



**(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

**Figure 3:** 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`

Since Moravec and Elfes (1985) 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 (Thrun, 1998) suggests the utility of going beyond metric representations. These hybrid representations have been successfully applied in tasks such as navigation (Konolige et al., 2011). In the related field of semantic mapping (e.g., Kuipers, 2000; Ekvall et al., 2007; Pronobis and Jensfelt, 2012; Liu and von Wichert, 2014), topological information is typically extracted from metric layers (occupancy grids). As described in Section 2, some works in semantic mapping do place greater emphasis on the detailed modeling of objects (e.g., Ranganathan and Dellaert, 2007; Wurm et al., 2011; Mason and Marthi, 2012). Wurm et al. (2011) modeled objects as local grid models within a hierarchy of octrees, where the hierarchy was informed by support relations. Mason and Marthi (2012) describe a system that maintains world models by detecting protrusions atop planes and considering them as 'objects'. However, as with the hybrid mapping community, object-based information is rarely propagated back down to the metric level.

Combining object-level and metric-level information is useful, as depicted in the example in Figure 3. In particular, I identified two ways in which fusion is particularly informative: free space detections strongly indicate that objects cannot be positioned in such regions (Figure 3(g)), 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.

# 4 Proposed Directions

I propose to continue investigating 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. Characterizing this trade-off and demonstrating efficient and sound estimation in real-world scenarios is the primary objective of this proposed thesis.

## 4.1 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 that are needed should surface when identifying suitable demonstration tasks later.

Besides aggregating static properties, there are at least two aspects of object state that will 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?

The second extension is the incorporation of known state constraints. Examples of constraints include object-object non-interpenetration (Grundmann et al., 2010; Wong et al., 2012), support/containment relationships, and stability/contact. State estimation with hard constraints is challenging because of the additional coupling of (typically 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.

## 4.2 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 the work described in Section 3.2, where object and metric information were 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. However, there are no guarantees on when such an approach is sound; a theoretical characterization is desirable.

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. More concretely, I hope to devise a scheme that takes as input a list of state variables and a description of the task, and prioritizes the variables such that state estimation resources can be devoted to an appropriately-small, high-priority subset.

# 5 Potential Developments

In line with the directions proposed in the previous section, I have identified two thrusts that are currently in the early stages of development. The first is on incorporating temporal dynamics in object attributes, the second is on task-relevant estimation by aggressively ignoring information.

## 5.1 Temporal Dynamics in Object Attributes via a Mathematical Connection

For the work described in Section 3.1 on semantic world modeling, I developed a DPMM-based approach to resolve the data association problem, with additional constraints that are implied by common data association assumptions. Under this picture of clustering, objects are analogous to clusters in the product space of attributes. Hence, temporal dynamics in object attributes are analogous to evolution in clusters. Clustering problems where clusters change over time have been studied recently within different communities (e.g., Chakrabarti et al., 2006; Blei and Lafferty, 2006). Within the Bayesian nonparametrics community, the dependent Dirichlet process (DDP) (MacEachern, 1999, 2000) is a general method for modeling dependencies between a *collection* of clusters/distributions, such as over a spatial region or, in our case, a temporal interval.

The DDP can be considered as a formulation that builds on top of the stick-breaking construction for the DP (Sethuraman, 1994). To define the DDP prior on distributions at two time steps, take the (infinite) collection of sticks that define the distribution at the first time step, apply a given stochastic transition process that transforms the stick weights and atoms, the result of which forms the distribution at the second time. Lin et al. (2010) observed an elegant mathematical connection between DDPs and Poisson processes, which represented clusters across time as a Markov chain of DPs, that can in turn be represented using compound Poisson processes. This observation led to the development of the DDP mixture model (DDPMM), along with an elegant Gibbs sampling algorithm for performing inference in this model. Building on this, Campbell et al. (2013) recently introduced an efficient algorithm for performing maximum *a posteriori* (MAP) inference, via an analysis of the low-variance asymptotics of the Gibbs sampler (Kulis and Jordan, 2012).

The DDPMM is exactly the right tool for incorporating discrete object dynamics into the semantic world modeling work. Addition, transition, and removal of clusters are all supported in the DDPMM, which can be used to model analogous operations on objects (when considered as clusters of attributes). As in the DPMM case, the generic DDPMM likely performs poorly on the data association problem, because it ignores important information encoded in the assumptions/constraints of the problem. Work is needed to encode these constraints in the DDPMM model and to adapt the Gibbs sampling algorithm to respect such assumptions.

While the DDPMM provides an elegant foundation for extending the semantic world model across time, inference is computationally challenging. Even in the static case, Gibbs sampling is slow because it requires many iterations through measurement data. Constraints that couple variables together further slows down inference by an order of magnitude. If the dynamic possibilities of objects are also considered, inference via a modified Gibbs sampling algorithm may be prohibitive. For the static case, I modified the efficient MAP inference algorithm by Kulis and Jordan (2012), and achieved object hypotheses comparable to the Gibbs sampler (Wong et al., 2014b). For the DDPMM, adapting the method by Campbell et al. (2013) will likely produce an analogous result.

Of course, comparing a Gibbs sampler against MAP inference is unfair, since the Gibbs sampler produces samples from a distribution of data association hypotheses, whereas MAP inference essentially produces a deterministic answer. However, it begs the question: Do we need to reason over the big latent distribution? Can we perform efficient MAP inference by default, and only perform the more expensive Gibbs sampling when a better characterization of uncertainty is needed?

## 5.2 A Model Attention and Selection Framework for Large-Scale Estimation

To deal with different kinds of uncertainty, much effort has gone into developing various estimators, including my previous work described in Section 3. However, little attention has been placed on *why* the particular estimation problem arises. State estimation should not be treated as a black box; *it should be tied to the task.* Estimating large sets of variables is computationally costly; just because a technique exists to estimate the values of certain variables does not justify its application. For robots performing tasks, only variables relevant to successful completion should be estimated. For example, while cooking, a robot should not prioritize estimating cleanliness of its surroundings. Similarly, while cleaning a specific room, not only should a robot not be concerned with estimating variables used in the cooking task, it should not even estimate cleanliness of other rooms.

Of course, the selection of relevant variables is not so clear-cut in practice. Lack of cleanliness in the kitchen environment may lead to food contamination during cooking. Yet, we want to avoid estimating all uncertain variables at once. Instead, I propose to initially only track a minimal set of directly-relevant variables, and gradually increase the sophistication of models *on-demand*, in a *local* fashion. This estimator refinement process is triggered by violations in expectations of task success. With respect to state estimation, if observed empirical quantities differ significantly from the current probabilistic model, then this indicates the model must be improved.

In light of the object-based semantic world model and the object-metric fused grid described in Section 3, the latter estimator can be viewed as a *refinement* of the former, because it fuses extra observations with the former model. The drawback of doing so is computational complexity: because the method reasons over grids of space, its representation scales with the volume of space covered, which, under discretization, typically results in many more grid cells compared to the number of objects seen. Moreover, the number of observations that need to be handled differs greatly as well; for example, each image of a scene with several objects on a table will only result in several attribute detections, but each image pixel generates an occupancy observation (or more). Ideally, we would track only the coarse object-attribute estimates (and only objects with relevant attribute values), and if the estimate is not sufficiently accurate (e.g., too much uncertainty), *nearby* occupancy information is incorporated via the finer estimator.

The above behavior emerges from a *attention-mismatch-refinement-learning* framework, wherein a small subset of task-relevant variables are estimated, and only upon differing from expected task outcomes (e.g., success) is the estimator incrementally refined by expanding the model class (with finer models and/or including more variables). The framework has the following components:

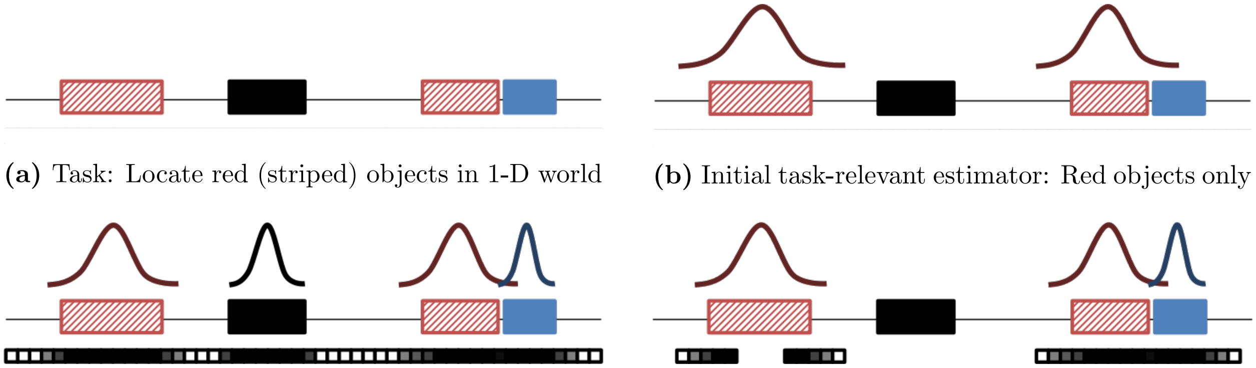
- **Attention: Task relevance.** Autonomous robots need to 'focus' on *relevant variables* for given tasks. Without constraints, the model-free 'model' is always most faithful to observed data. For many tasks, however, only a small subset of variables benefit the task with additional accuracy. Improved accuracy for other variables is irrelevant and is a waste of resources. Monolithic estimation of all variables is not practical; *only a small subset matters.*

- **Mismatch: Fault detection.** If deviations between expected and observed values exceed thresholds, *as informed by the task*, the current model must be refined.

- **Refinement: Model class expansion.** When a relevant variable's model is found inadequate, explore a larger model class *for a small subset of related variables only.*

- **Learning: Estimating parameters.** Expanded model classes will have additional parameters to be learned. Here is where the inference techniques developed for various estimators apply. Non-parametric 'models' (empirical estimates) can be considered the final refinement, where empirical estimates are used directly, as in model-free approaches.

### 5.2.1 A proof-of-concept case study: 1-D colored intervals domain



**(a)** Task: Locate red (striped) objects in 1-D world

**(b)** Initial task-relevant estimator: Red objects only

**(c)** Naïve estimator: Estimate all variables

**(d)** After refinement: Includes task-correlated vars.

**Figure 4:** Locating unknown red (striped) objects in a 1-D domain (line). Curves above objects represent Gaussian distributions on the object's centroid. Shaded boxes below the line show a discretized occupancy grid, where darker shades indicate greater probabilities of being occupied. Different estimators keep track of different sets of variables; those not shown are ignored.

Consider the domain and task depicted in Figure 4(a). The task is to locate (to some specified uncertainty tolerance) red (striped) objects on the real line, given a list of 'images' as input, each containing a small set of noisy attribute (location, length, and color) detections and a larger set of occupancy observations. The naïve solution is to run all estimators on all the observations, as depicted in Figure 4(c). Since the task is to locate only red objects, this approach, while sound, is inefficient, especially if the domain is significantly larger and contains few red objects.

Instead, consider the estimator in Figure 4(b). Only objects whose color attribute is red with high probability are given *attention*; the rest is discarded/ignored. This is conceivably the minimal estimator for the task. However, these observations are very noisy (e.g., the output of an entire object detection pipeline) and lead to large variance in the posterior attribute distribution, above the required tolerance. The performance of this estimator is therefore *mismatched* for the task, and therefore estimator *refinement* is necessary.

The refinement process involves adding new variables to the estimator and estimating their values based on a buffer of lazily-stored recent observation values. Variables are ranked and added (up to a threshold) based on the expected cost reduction obtained by incorporating each of them. More formally, let $X$ be the variables already estimated (e.g., red object locations), and $Y$ be a candidate variable to add. For each candidate, consider the benefit in cost $f(\cdot)$ over the distribution $p_X$, in expectation over the possible outcomes of $Y$ (given what is already known about $X$):

$$\mathbf{R}(\mathbf{p}_{\mathbf{X}|\mathbf{Y}}) \triangleq \mathbb{E}_{\mathbf{y}\sim\mathbf{p}_{\mathbf{Y}}}\left[\mathbf{f}(\mathbf{p}_{\mathbf{X}}) - \mathbf{f}(\mathbf{p}_{\mathbf{X}|\mathbf{Y}=\mathbf{y}})\right] \text{ , where } p_Y = \int p_{Y|X=x}\, p_X(x)\,\mathrm{d}x$$

In our case, the cost function is the variance on red objects' location distributions. The criterion leads to the addition of two sets of variables (Figure 4(d)). The first set, for the left red object, is a subset of occupancy grid cells; their primary purpose is to distinguish the boundary of the object more finely. The second set, for the right red object, is more interesting: not only does it include associated occupancy grid cells, it also includes the location attribute of the nearby blue object. This latter variable is helpful because of the domain constraint that objects cannot overlap each other, which correlates the states of the two objects (Wong et al., 2012). Incorporating these new variables in the refined estimator sufficiently reduces the variance for successful task completion.

### 5.2.2 Relevant approaches and future directions

So far, I have considered one possible criterion for determining which variables to include in the estimator refinement step. The decision-theoretic criterion is related to the expected error reduction approach to active learning (Settles, 2012), where new training data is selected based on its expected benefit to reducing error in the learned hypothesis. Other approaches to active learning may serve as an inspiration for alternative ways to perform estimator refinement. Approaches from feature/variable selection (Guyon, 2003) may also provide insight.

Apart from determining which variables to include in refinement, *when* to trigger this process is also important. Various metrics for Bayesian 'surprise' (e.g., Ranganathan and Dellaert, 2009; Maier and Steinbach, 2011; White et al., 2014) can be used to measure how unlikely measurements and belief states are, compared to model/filter predictions. Methods from the mature literature of fault detection/identification/diagnosis can also be applied to determine when a mismatch event has occurred (Kurien and Nayak, 2000; Verma et al., 2004; Pettersson, 2005; Dearden, 2010). Most of these approaches still require *ad-hoc* thresholds (for the level of mismatch); ultimately they should be automatically learned from task performance. Possible techniques for this include Bayesian optimization (Snoek et al., 2012) and metareasoning (Cox and Raja, 2011).

The presented framework should in principle work for any hierarchy of estimators and models. Possible candidates for testing this include using grammars that generate increasingly-complex models (Grosse et al., 2012), and a recent approach that uses a hierarchical decomposition of variables to produce a partition of variables with varying fineness (Steinhardt and Liang, 2014).

### 5.2.3 Closing thought

Ultimately, I view the proposed model attention and selection framework as part of an answer to the following observation by Anderson and Oates (2007):

> Natural intelligent systems tend to be robust;
> artificial intelligent systems tend to be brittle.

As mortal designers, the models we supply to systems will eventually be wrong, and hence systems become brittle. Robust AI systems need to be able to discover model mismatch autonomously and apply flexible estimators that learn to deal with uncertainty in the world.

# 6 Proposed Timeline

- Fall 2014: Work on model selection problem (Section 5.2). Formulate problem rigorously and demonstrate utility of coarse-to-fine, task-relevant estimation scheme. Appropriate heuristics may need to be developed for world modeling on mobile-manipulation robots.

- Fall 2014 / Winter 2015: Devise 'kitchen' demonstration for the PR2 robot that showcases object-based world models, and possibly task-relevant estimation.

- Winter 2015: Work on tracking object changes over time (Section 5.1), in particular using the DDPMM as the theoretical foundation. Devise approximations to ensure inference is scalable.

- Winter 2015 / Spring 2015: Increase complexity of demonstrations, possibly including the task of object search (Wong et al., 2013a). Identify missing pieces that still require work.

- Spring 2015: Devise plan for completion. Write thesis, defend, and revise!

# References

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.

R. Anati, D. Scaramuzza, K.G. Derpanis, and K. Daniilidis. Robot localization using soft object detection. In *IEEE International Conference on Robotics and Automation*, 2012.

M.L. Anderson and T. Oates. A review of recent research in metareasoning and metalearning. *AI Magazine*, 28(1):7–16, 2007.

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

N. Atanasov, B. Sankaran, J. Le Ny, T. Koletschka, G. Pappas, and K. Daniilidis. Hypothesis testing framework for active object detection. In *International Conference on Robotics and Automation*, 2013.

Y. Bar-Shalom and T.E. Fortmann. *Tracking and Data Association.* Academic Press, 1988.

D.M. Blei and J.D. Lafferty. Dynamic topic models. In *International Conference on Machine Learning*, 2006.

T. Campbell, M. Liu, B. Kulis, J.P. How, and L. Carin. Dynamic clustering via asymptotics of the dependent Dirichlet process mixture. In *Advances in Neural Information Processing Systems*, 2013.

D. Chakrabarti, R. Kumar, and A. Tomkins. Evolutionary clustering. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

I.J. Cox and S.L. Hingorani. An efficient implementation of Reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(2):138–150, 1996.

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

M.T. Cox and A. Raja, editors. *Metareasoning: Thinking About Thinking.* MIT Press, 2011.

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.

R. Dearden. Bayesian fault diagnosis: Common approaches and challenges. In *International Workshop on Cognitive Information Processing*, 2010.

F. Dellaert, S.M. Seitz, C.E. Thorpe, and S. Thrun. EM, MCMC, and chain flipping for structure from motion with unknown correspondence. *Machine Learning*, 50(1–2):45–71, 2003.

R. Eidenberger and J. Scharinger. Active perception and scene modeling by planning with probabilistic 6D object poses. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010.

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

J. Elfring, S. van den Dries, M.J.G. van de Molengraft, and M. Steinbuch. Semantic world modeling using probabilistic multiple hypothesis anchoring. *Robotics and Autonomous Systems*, 61(2):95–105, 2013.

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

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.

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

T. Grundmann, M. Fiegert, and W. Burgard. Probabilistic rule set joint state update as approximation to the full joint state estimation applied to multi object scene analysis. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010.

Isabelle Guyon. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

G.D. Hager and B. Wegbreit. Scene parsing using a prior world model. *International Journal of Robotics Research*, 30(12):1477–1507, 2011.

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

B. Kuipers. The spatial semantic hierarchy. *Artificial Intelligence*, 119(1–2), 2000.

B. Kulis and M.I. Jordan. Revisiting k-means: New algorithms via Bayesian nonparametrics. In *International Conference on Machine Learning*, 2012.

J. Kurien and P.P. Nayak. Back to the future for consistency-based trajectory tracking. In *National Conference on Artificial Intelligence*, 2000.

T. Kurien. Issues in the design of practical multitarget tracking algorithms. In Y. Bar-Shalom, editor, *Multitarget-Multisensor Tracking: Advanced Applications*, pages 43–84. Artech House, 1990.

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.

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

Z. Liu and G. von Wichert. Extracting semantic indoor maps from occupancy grids. *Robotics and Autonomous Systems*, 62(5):663–674, 2014.

S.N. MacEachern. Dependent nonparametric processes. In *ASA Section on Bayesian Statistics*, 1999.

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

R.P.S. Mahler. *Statistical Multisource-Multitarget Information Fusion.* Artech House, 2007.

W. Maier and E. Steinbach. Surprise-driven acquisition of visual object representations for cognitive mobile robots. In *IEEE International Conference on Robotics and Automation*, 2011.

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.

J. Mason and B. Marthi. An object-based semantic world model for long-term change detection and semantic querying. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012.

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

R.M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

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.

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

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.

H. Pasula, S. Russell, M. Ostland, and Y. Ritov. Tracking many objects with many sensors. In *International Joint Conference on Artificial Intelligence*, 1999.

O. Pettersson. Execution monitoring in robotics: A survey. *Robotics and Autonomous Systems*, 53 (2):73–88, 2005.

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

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

A. Ranganathan and F. Dellaert. Bayesian surprise and landmark detection. In *IEEE International Conference on Robotics and Automation*, 2009.

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

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.

J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

B. Settles. *Active Learning.* Morgan & Claypool, 2012.

J. Snoek, H. Larochelle, and R.P. Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, 2012.

J. Steinhardt and P. Liang. Filtering with abstract particles. In *International Conference on Machine Learning*, 2014.

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

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

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.

J. Velez, G. Hemann, A.S. Huang, I. Posner, and N. Roy. Modelling observation correlations for active exploration and robust object detection. *Journal of Artificial Intelligence Research*, 44: 423–453, 2012.

V. Verma, G. Gordon, R. Simmons, and S. Thrun. Real-time fault diagnosis. *IEEE Robotics & Automation Magazine*, 11(2):56–66, 2004.

T. Whelan, M. Kaess, M.F. Fallon, H. Johannsson, J.J. Leonard, and J.B. McDonald. Kintinuous: Spatially extended KinectFusion. In *Robotics: Science and Systems Workshop on RGB-D: Advanced Reasoning with Depth Cameras*, 2012.

T. Whelan, M. Kaess, H. Johannsson, M.F. Fallon, J.J. Leonard, and J.B. McDonald. Real-time large scale dense RGB-D SLAM with volumetric fusion. *International Journal of Robotics Research*, 2014. To appear.

A. White, J. Modayil, and R.S. Sutton. Surprise and curiosity for big data robotics. In *AAAI Conference on Artificial Intelligence Workshop on Sequential Decision-Making with Big Data*, 2014.

L.L.S. Wong, L.P. Kaelbling, and T. Lozano-Pérez. Collision-free state estimation. In *IEEE International Conference on Robotics and Automation*, 2012.

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*, 2013a.

L.L.S. Wong, L.P. Kaelbling, and T. Lozano-Pérez. Data association for semantic world modeling from partial views. In *International Symposium of Robotics Research*, 2013b.

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*, 2014a.

L.L.S. Wong, L.P. Kaelbling, and T. Lozano-Pérez. Data association for semantic world modeling from partial views. *International Journal of Robotics Research*, 2014b. In submission.

K.M. Wurm, D. Hennes, D. Holz, R.B. Rusu, C. Stachniss, K. Konolige, and W. Burgard. Hierarchies of octrees for efficient 3D mapping. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011.

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.