Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence

# A Model Attention and Selection Framework for Estimation of Many Variables, with Applications to Estimating Object States in Large Spatial Environments

**Lawson L. S. Wong**
Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology
32 Vassar St., 32-G418; Cambridge, MA 02139, USA • +1 (617) 735-5393 • lsw@csail.mit.edu

## Introduction

Robots performing service tasks such as cooking and cleaning in human-centric environments require knowledge of certain environmental states in order to complete tasks successfully. For example, storage locations of specific ingredients and utensils are needed for cooking; dirtiness of particular regions of space may be required for efficient cleaning. Typically these task-critical states cannot be directly observed, and must be estimated by using (noisy) perception and prior domain knowledge. Bayesian filtering solves such estimation problems for a wide variety of state characteristics: given a particular set of variables (uncertain states) to be estimated, Bayesian filtering techniques most likely already exist in that particular regime. While much effort has gone into developing various estimators, less attention has been placed on *why* the particular estimation problem arises.

In this work, I argue that state estimation should no longer be treated as a black box. 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 whose ultimate mission is to complete tasks, only variables that are relevant to successful completion should be estimated. Returning to cooking and cleaning, 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, as argued earlier, 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 the remainder, I demonstrate this through a proof-of-concept case study.



## A Tale of Two Estimators

Previously, I have developed two different estimators for the *world modeling* problem, the estimation of objects' states within the world. Abstractly, on the level of object attributes, a system exists that takes black-box attribute detections, such as object type and pose, and estimates the objects that are present (including their number, which is unknown) and their attribute values (Wong, Kaelbling, and Lozano-Pérez 2013). Although this gives an elegant 'semantic' view of objects as clusters in joint-attribute space, it ignores crucial information related to the geometric realization of objects, such as their physical extent in space. In particular, low-level observations on whether specific 'voxels' of space are occupied/free cannot be easily incorporated on the object-attribute level. Such observations are traditionally tracked using occupancy grids (Moravec and Elfes 1985), and we developed a second estimator that attempts to fuse object-attribute estimates with geometric occupancy grids (Wong, Kaelbling, and Lozano-Pérez 2014). More details on the estimators are provided in the complementary material.[1]

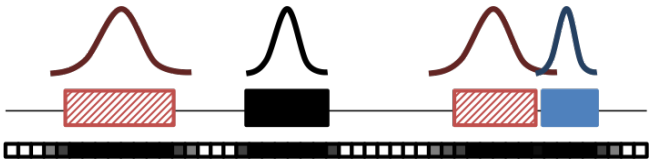
The latter estimator can be viewed as a *refinement* of the former, because it fuses extra observations with the former model. The drawback of doing so is computational complexity: because the method reasons over grids of space, its representation scales with the volume of space covered, which, under 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 the 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* 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).

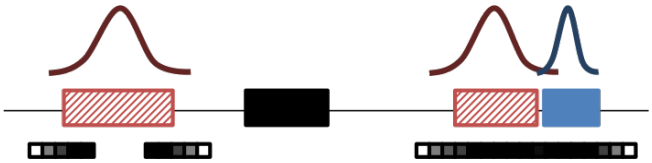
[1] http://people.csail.mit.edu/lsw/papers/aaai2014-models.pdf

(a) Task: Locate red (striped) objects in 1-D with 4 objects

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



(c) Naïve estimator: Estimate all variables



(d) After refinement: Includes task-correlated variables

Figure 1: 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.

## Case Study: 1-D Colored Intervals Domain

As a proof-of-concept, consider the domain and task depicted in Figure 1(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 1(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 1(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 a probabilistic condition detailed in the complementary material. This leads to the addition of two sets of variables. 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 attribute-level variables 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. Incorporating these new variables in the refined estimator sufficiently reduces the variance for successful task completion.

For details, please refer to the complementary material:

**http://people.csail.mit.edu/lsw/papers/aaai2014-models.pdf**

## Future Directions

Apart from determining which variables to include in refinement, *when* to trigger this process is also important. Currently this is determined by *ad-hoc* thresholds (for the level of mismatch); ultimately they should be automatically learned from task performance. Possible techniques for handling this issue include execution monitoring (Pettersson 2005), Bayesian optimization (Snoek, Larochelle, and Adams 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, Salakhutdinov, and Tenenbaum 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).

## References


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

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

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

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

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

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

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

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