Data Assocation for Semantic World Modeling from Partial Views

Lawson L.S. Wong, Leslie Pack Kaelbling, and Tomás Lozano-Pérez

Abstract Autonomous mobile-manipulation robots need to sense and interact with objects to accomplish high-level tasks such as preparing meals and searching for objects. To achieve such tasks, robots need semantic world models, defined as object-based representations of the world involving task-level attributes. In this work, we address the problem of estimating world models from semantic perception modules that provide noisy observations of attributes. Because attribute detections are sparse, ambiguous, and are aggregated across different viewpoints, it is unclear which attribute measurements are produced by the same object, so *data association* issues are prevalent. We present novel clustering-based approaches to this problem, which are more efficient and require less severe approximations compared to existing tracking-based approaches. These approaches are applied to data containing object type-and-pose detections from multiple viewpoints, and demonstrate comparable quality to the existing approach using a fraction of the computation time.

1 Introduction

Much of the everyday human physical environment is made up of coherent physical objects. Environmental dynamics are well described in terms of the effects of actions on those objects. Perceptual systems are able to report detections of objects with type, location, color, and other properties. Humans naturally designate both goals and prior information in terms of objects. Thus, it is appropriate for robots to construct 'mental models' of their environment that are structured around objects, their properties, and their relations to one another.

In this work, we define it a semantic world model to be a set of objects with associated attributes and relations. For concreteness, consider the following tasks and their potentially relevant objects and attributes:

Lawson L.S. Wong · Leslie Pack Kaelbling · Tomás Lozano-Pérez
CSAIL, MIT, Cambridge, MA 02139 e-mail: {lsw, lpk, tlp}@csail.mit.edu

- Cooking steaks on a pan: Objects — Steaks, pan, stove, etc.
  Attributes — *CookedTime*, *Thickness*, *SteakPositionRelativeToPan*
- Finding chairs for guests: Objects — Furniture, people
  Attributes — *IsChair*, *Sittable* (if $\neg$*IsChair*), *Movable*, *Location*, *SittingOn*
- Rearranging objects on a table: Objects — Items on table
  Attributes — *Shape*, *Type*, *RelativePositionAndOrientation*, *GraspPoints*

A common theme underlying these tasks, and many others, is that successful planning and execution hinges on good world-state estimation and monitoring. Dynamic attributes listed above also highlight why object-based representations are uniquely suitable for dynamic tasks: transition dynamics tends to operate on the level of objects. For example, it is much more natural to express and reason about a piece of steak that is being cooked, as opposed to points in a point cloud or cells in an occupancy grid that are 'cooked'. Although we focus on the static case in this paper, our ultimate goal is to provide a framework for estimating and monitoring large semantic world models involving objects and attributes that change over time as a result of physical processes as well as actions by the robot and other agents.

In this work, we address the problem of constructing world models from semantic perception modules that provide noisy observations of attributes. Due to noise, occlusion, and sensors' limited field of view, observations from multiple viewpoints will typically be necessary to produce a confident world model. Because attribute detections are sparse, noisy, and inherently ambiguous, where it is unclear which attribute measurements were produced by the same object across different views, *data association* issues become critical. This is the greatest challenge; if the measurement-object correspondences were known, the resulting object-attribute posterior distributions would be efficiently computable.

We begin by stating a formal model for a simplified 1-D version of the world-model estimation problem in Sect. 2, and then review a classic solution approach based on tracking in Sect. 3. The main contribution of this work is the development of several novel clustering-based data association approaches, described in Sects. 4 and 5. Application of the semantic world-modeling framework to object type-and-pose estimation is then briefly discussed in Sect. 6, followed in Sect. 7 by experimental results using data collected with a Kinect sensor on a mobile robot.

2 The 1-D Colored-Lights Domain

The approaches described in this paper apply to domains of arbitrary complexity. For clarity of explanation we begin by introducing a model of minimal complexity and then use it for an initial demonstration of the methods.

The world consists of an unknown number (K) of stationary lights. Each light is characterized by its color c_k and its location $l_k \in \mathbb{R}$. A finite universe of colors of size T is assumed. A robot moves along this 1-D world, occasionally gathering partial views of the world with known field of views $[a_v, b_v] \subset \mathbb{R}$. Within each view, M_v lights of various colors and locations are observed, denoted by $o_m^v \in [T] \triangleq \{1, \ldots, T\}$

and $x_m^v \in \mathbb{R}$ respectively. These (o_m^v, x_m^v) pairs may be noisy (in both color and location) or spurious (false positive – FP) measurements of the true lights. Also, a light may sometimes fail to be perceived (false negative – FN). Given these measurements, the goal is to determine the posterior distribution over configurations (number, colors, and locations) of lights in the explored region of the world.

We assume the following form of noise models. For color observations, for each color t, there is a known discrete distribution $\phi^t \in \Delta^T$ (estimable from perception apparatus statistics) specifying the probability of color observations:

$$\phi_i^t = \begin{cases} \mathbb{P}(\text{no observation for color } t \text{ light}) , & i = 0 \\ \mathbb{P}(\text{color } i \text{ observed for color } t \text{ light}) , & i \in [T] \end{cases} . \tag{1}$$

A similar distribution ϕ^0 specifies the probability of observing each color given that the observation was a false positive. False positives are assumed to occur in a proportion p_{FP} of object detections. For location observations, if the observation corresponds to an actual light, then the observed location is assumed to be Gaussian-distributed, centered on the actual location. The variance is *not* assumed known and will be estimated for each light from measurement data. For false positives, the location is assumed to be uniformly distributed over the field of view (Unif$[a_v, b_v]$).

Next, we present the core problem of this domain. Given sets of color-location detections from a sequence of views, $\{\{(o_m^v, x_m^v)\}_{m=1}^{M_v}\}_{v=1}^{V}$, we want to infer the posterior distribution on the configuration of lights $\{(c_k, l_k)\}_{k=1}^{K}$, where K is unknown as well. If we knew, for each light, which subset of the measurements were generated from that light, then we would get K decoupled estimation problems (assuming lights are independent from each other). With suitable priors, these single-light estimation problems admit efficient solutions; details can be found in the appendix.

The issue is that these associations are unknown. Therefore, we must reason over the *space* of possible data associations. For each observation, let z_m^v be the index of the light that the observation corresponds to (ranging in $[K]$ for a configuration with K lights), or 0 if the observation is a false positive. z_m^v is the latent association for measurement (o_m^v, x_m^v). Let $\mathbf{z^v}$ be the concatenated length-M_v vector of all z_m^v variables in view v, and let $\{\mathbf{z^v}\}$ be the collection of all correspondence vectors from the V views. We then aggregate estimates over all latent associations[1]:

$$\mathbb{P}\left(\{(c,l)\}_k \,\middle|\, \{\{(o,x)\}_m\}_v\right) = \sum_{\{\mathbf{z^v}\}} \mathbb{P}\left(\{(c,l)\} \,|\, \{\mathbf{z}\}, \{\{(o,x)\}\}\right) \mathbb{P}\left(\{\mathbf{z}\} \,|\, \{\{(o,x)\}\}\right) . \tag{2}$$

The first term is given by the decoupled estimation problems mentioned above, and results in a closed-form posterior distribution given in the appendix. The desired posterior distribution on the left is therefore, in exact form, a mixture over the closed-form posteriors. The problem is that the number of mixture components is exponential in M_v and V, one for each full association $\{\mathbf{z^v}\}$, so maintaining the full

[1] Indices have been dropped to reduce clutter, and will continue to be omitted in the future if clear from context. For this case, please refer to two paragraphs above for the full set of indices.

posterior distribution is intractable. Finding tractable approximations to this light configuration posterior distribution is the subject of Sects. 3–5.

3 A Tracking-Based Approach

If we consider the lights to be stationary targets and the views to be a temporal sequence, a tracking filter approach can be used. Tracking simultaneously solves the data association (measurement correspondence) and target parameter estimation (light colors and locations) problems. Of the wide variety of existing approaches for this classic problem [4], the multiple hypothesis tracking (MHT) filter [17] is most appropriate because it allows for an unknown number of targets. In fact, Elfring et al. [11] recently adopted this approach to the semantic world-modeling problem, and have provided extensive rationale for using MHTs over other tracking approaches.

We provide a gist of the MHT approach and discuss a problematic issue below; readers are referred to Elfring et al. [11] for details. The MHT algorithm maintains, at every timestep (view) v, a distribution over all possible associations of measurements to targets up to v. At each view, MHT therefore needs to propagate *each* previous hypothesis forward with *each* possible association in view v. One way to consider this is as a tree, where nodes of depth v are associations up to view v, and a distribution is maintained on the leaves. Each view introduces a new layer of nodes, where the branching factor is the number of valid associations in that view.

Estimating this branching factor highlights the intractability of the MHT. Assume we know which of the existing targets are within the current field of view based on the hypothesis on previous views (this can be found by gating). Denote the indices of these targets as the size-K_v set $\{k\}^v$. Another common assumption used in the tracking literature is that in a single view, each target can generate at most one non-spurious measurement. We will refer to this as the one-measurement-per-object (OMPO) assumption. We now define validity of correspondence vectors $\mathbf{z}^v$. First, by the OMPO assumption, no entry may be repeated in $\mathbf{z}^v$, apart from 0 for false positives. Second, an entry must either be 0, and target index in $\{k\}^v$, or be a new (non-existing) index; otherwise, it corresponds to an out-of-range target. A correspondence $\mathbf{z}^v$ is valid if and only if it satisfies both conditions.

The following quantities can be found directly from $\mathbf{z}^v$:

$$\begin{aligned} n_0 &\triangleq \text{Number of false positives (0 entries)} ; \\ n_\infty &\triangleq \text{Number of new targets (non-existing indices)} ; \\ \delta_k &\triangleq \mathbb{I}\{\text{Target } k \text{ is detected } (\exists m. z_m^v = k)\}, k \in \{k\}^v ; \\ n_1 &\triangleq \text{Number of matched targets} = M_v - n_0 - n_\infty = \sum_k \delta_k , \end{aligned} \tag{3}$$

where $\mathbb{I}\{\cdot\}$ denotes the indicator function and M_v is the number of measurements in view v. The number of valid associations is given by the following expression:

$$\sum_{n_0=0}^{M_v} \sum_{n_\infty=0}^{M_v-n_0} \binom{M_v}{n_0, n_1, n_\infty} \times \binom{K_v}{n_1} \times n_1! = \sum_{n_0=0}^{M_v} \sum_{n_\infty=0}^{M_v-n_0} \frac{M_v!}{n_0! n_\infty!} \times \frac{K_v!}{n_1!(K_v-n_1)!} \,. \quad (4)$$

Even with 4 measurements and 3 within-range targets, the branching factor is 304, so considering all hypotheses is clearly intractable. Many hypothesis-pruning strategies have been devised (e.g., [13, 6]), the simplest of which include keeping the best hypotheses or hypotheses with probability above a certain threshold. More complex strategies to combine similar tracks and reduce the branching factor have also been considered. In the experiments of Sect. 7 we simply keep hypotheses with probability above a threshold of 0.01. As we will demonstrate in the experiments, an MHT filter using this aggressive pruning strategy can potentially cause irreversible association errors and make overconfident conclusions.

4 A Clustering-Based Approach

If we consider all the measurements together and disregard their temporal relationship, we expect the measurements to form clusters in the product space of colors and locations ($[T] \times \mathbb{R}$), allowing us to derive estimates of the number of lights and their parameters. In probabilistic terms, the measurements are generated by a mixture model, where each mixture component is parameterized by the unknown parameters of a light. Since the number of lights in the world is unknown, we also do not want to limit the number of mixture components a priori.

A useful model for performing clustering with an unbounded number of clusters is the Dirichlet process mixture model (DPMM) [2, 14], a Bayesian non-parametric approach that can be viewed as an elegant extension to finite mixture models. The Dirichlet process (DP) acts as a prior on *distributions* over the cluster parameter space. A random distribution over cluster parameters G is first drawn from the DP, then a countably infinite number of cluster parameters are drawn from G, from which the measurement data is finally drawn according to our assumed observation models. Although the model can potentially be infinite, the number of clusters is finite in practice, as they will be bounded by the total number of measurements (typically significantly fewer if the data exhibits clustering behavior). The flexibility of the DPMM clustering model lies in its ability to 'discover' the appropriate number of clusters from the data.

We now derive the DPMM model specifics and inference procedure for the colored-lights domain. A few more assumptions need to be made and parameters defined first. Our model assumes that the cluster parameter distribution G is drawn from a DP prior $\mathrm{DP}(\alpha, H)$, where H is the base distribution and α is the concentration hyperparameter (controlling the similarity of G and H, and also indirectly the number of clusters). H acts as a 'template' for the DP, and is hence also a distribution over the space of cluster parameters. We set it to be the product distribution of π, the prior on colors, and a uniform distribution over the explored region. To

accommodate false positives, which occur with probability p_{FP}, we scale G from the DP prior by a factor of $(1-p_{\text{FP}})$ for true positives.

For ease of notation when deriving the inference procedure, we express the DP prior in an equivalent form based on the stick-breaking construction [18]. The idea is that the sizes of clusters are determined by a random process that first selects some proportion of the whole ('breaks the stick'), uses one part to define the size of a cluster, and then recursively subdivides the rest. Parameters are drawn from the base distribution H and associated with each cluster. More formally:

$$\begin{aligned} \theta &\sim \text{GEM}(\alpha)\ ; \\ (c_s, l_s) &\sim H \triangleq \pi \times \text{Unif}[A;B]\ , \end{aligned} \tag{5}$$

where GEM (Grifths-Engen-McCloskey) is the distribution over stick weights θ, and $\pi \in \Delta^{(T-1)}$ is a prior distribution on colors, reflecting their relative prevalence. By defining $G(c,l) \triangleq \sum_{s=1}^{\infty} \theta_s \times \mathbb{I}[(c,l) = (c_s, l_s)]$, G is a distribution over the cluster parameters and is distributed as $\text{DP}(\alpha, H)$. The rest of the generative process is:

$$\begin{aligned} \theta'_k &= \begin{cases} p_{\text{FP}}\ , & k = 0 \\ (1-p_{\text{FP}})\,\theta_k\ , & k \neq 0 \end{cases}\ ; && \text{Cluster proportions (with FPs)} \\ \mu_k, \tau_k &\sim \text{NormalGamma}(\nu, \lambda, \alpha, \beta)\ ; && \text{Cluster location distr. params.} \\ z_m^v &\sim \theta'\ , \qquad m \in [M_v]\ ,\ v \in [V]\ ; && \text{Cluster assignment (for each obs.)} \\ o_m^v &\sim \begin{cases} \phi^0\ , & z_m^v = 0 \\ \phi^{c_z}\ , & z_m^v \neq 0 \end{cases}\ ; && \text{Color observation} \\ x_m^v &\sim \begin{cases} \text{Unif}[a_v, b_v]\ , & z_m^v = 0 \\ \mathcal{N}\left(\mu_k, \tau_k^{-1}\right)\ , & z_m^v \neq 0 \end{cases}\ . && \text{Location observation} \end{aligned} \tag{6}$$

The most straightforward way to perform inference in a DPMM is by Gibbs sampling. In particular, we derive a collapsed Gibbs sampler for the cluster correspondence variables z and integrate out the other latent variables c, μ, τ, θ. In Gibbs sampling, we iteratively sample from the conditional distribution of each z_m^v, given all other correspondence variables (which we will denote by z^{-vm}). By Bayes' rule:

$$\begin{aligned} &\mathbb{P}\left(z_m^v = k \,\middle|\, z^{-vm}, \{\{(o,x)\}\}\right) \\ &\propto \mathbb{P}\left(o_m^v, x_m^v \,\middle|\, z_m^v = k, z^{-vm}, \{\{(o,x)\}\}^{-vm}\right)\ \mathbb{P}\left(z_m^v = k \,\middle|\, z^{-vm}, \{\{(o,x)\}\}^{-vm}\right) \\ &\propto \mathbb{P}\left(o_m^v, x_m^v \,\middle|\, \{\{(o,x)\}\}_{z=k}^{-vm}\right) \mathbb{P}\left(z_m^v = k \,\middle|\, z^{-vm}\right)\ . \end{aligned} \tag{7}$$

In the final line, the first term can be found from the posterior predictive distributions described in the appendix (Eqs. 14 and 17), noting that the observations being conditioned on *exclude* (o_m^v, x_m^v) and depend on the current correspondence variable samples (to determine which observations belong to cluster k).

The second term is given by the Chinese restaurant process (CRP), obtained by integrating out the DP prior on θ. Together with our prior on false positives:

$$\mathbb{P}\left(z_m^v = k \,\middle|\, z^{-vm}\right) = \begin{cases} (1-p_{\text{FP}})\,\frac{N_k^{-vm}}{\alpha+N-1}\,, & k \text{ exists} \\ (1-p_{\text{FP}})\,\frac{\alpha}{\alpha+N-1}\,, & k \text{ new} \\ p_{\text{FP}}\,, & k=0 \end{cases}, \tag{8}$$

where N_k^{-vm} is the number of observations currently assigned to cluster k (excluding (v,m)), and N is the total number of non-false-positive observations across all views.

By combining Eqs. 7 and 8, we have a method of sampling from the conditional distribution of individual correspondences z_m^v. Although the model supports an infinite number of clusters, the modified CRP expression (Eq. 8) shows that we only need to compute $k+2$ values for one sampling step, which is finite as clusters without data are removed. One sampling sweep over all correspondence variables $\{\{z\}\}$ constitutes one sample from the DPMM. Given the correspondence sample, finding the posterior configuration sample is simple. The number of lights is given by the number of non-empty clusters. Eq. 14 applied with all data belonging to one cluster provides the posterior distribution on the light's color. The hyperparameter updates in Eq. 16 similarly gives the posterior joint distribution on the light's location and precision of the observation noise model.

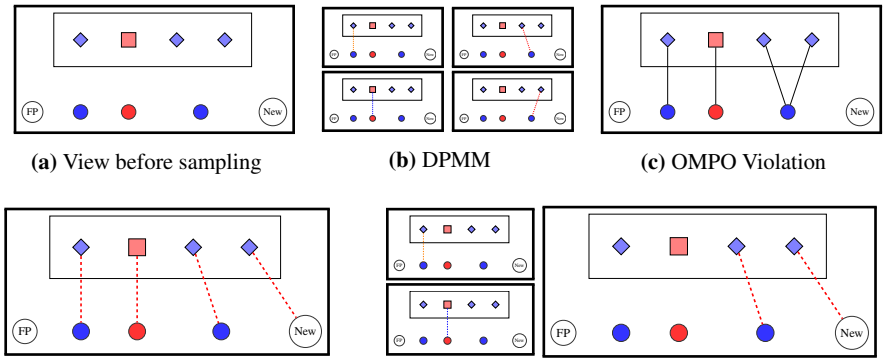
5 Incorporating View-Level Information and Constraints

The DPMM-based solution to the colored-lights problem is a straightforward application of the DPMM, but ignores two fundamental pieces of information:

- **False negatives (FN)**: The DPMM does not consider which clusters are visible when a measurement is made. It may therefore posit a cluster for a spurious measurement when its absence in other views would have suggested otherwise.
- **One-measurement-per-object (OMPO) assumption**: Consider the scenario depicted in Fig. 1(c), where two blue lights are placed close to each other and hence easily confusable. The DPMM ignores the OMPO assumption and may associate both to the same cluster, even if they were both observed in every view.

Both are consequences of the DPMM's conditional independence assumptions.

To see this, consider the concrete example depicted in Fig. 1, where we wish to sample cluster assignments for an entire view's $M_v = 4$ measurements. The DPMM Gibbs sampler samples the cluster assignment for each measurement *individually*, as shown in Fig. 1(b). This causes the two right-most measurements to be assigned to the same cluster, a violation of the OMPO assumption. The assumption states that *at most one* measurement in a single view can be assigned to each cluster; this view-level constraint cannot be incorporated on the level of individual measurements (DPMM). Likewise, a false negative only arises if *none* of the measurements in a view are assigned to a cluster within the field of view. To handle these constraints we must couple the measurements and sample their assignments *jointly*.



(a) View before sampling **(b)** DPMM **(c)** OMPO Violation

(d) DPMM-FullView **(e)** DPMM-Factored

Fig. 1: A concrete example for illustrating concepts in Sect. 5. **(a)** Each thick outer box depicts measurements in the same single view (inner box), and the clusters that each measurement can be assigned to (row below inner box). The view we consider has 4 measurements of lights' locations and colors. The existing clusters within the field of view are shown as colored circles (these were determined from other views). Measurements can also be assigned to the two 'clusters' to the left and right, for false positives and new clusters respectively. The task is to assign one of the 5 clusters in the bottom row to each measurement in the inner box. **(b)** The DPMM samples cluster assignments for each measurement independently. **(c)** This causes potential violations of the one-measurement-per-object (OMPO) assumption, where each cluster generates at most one observation within each view. **(d)** One solution is to consider all measurement assignments in the view jointly. However, as explained in Sect. 5.1, this is inefficient. **(e)** A more efficient approximation is derived in Sect. 5.2 by jointly considering *only* measurements that are OMPO-violating. Measurements that are unlikely to cause constraint violation, such as the two left ones in the example, are considered independently. This provides a trade-off between DPMM and DPMM-FullView.

5.1 DPMM-FullView

More formally, consider the view's joint correspondence vector $\mathbf{z}^v$. The induced conditional distribution $\mathbb{P}(\mathbf{z}^v \,|\, \mathbf{z}^{-v})$ that the DPMM Gibbs sampler uses is, by conditional independence, the product of M_v copies of Eq. 8:

$$\mathbb{P}_{\text{DPMM}}\left(\mathbf{z}^v \,\middle|\, \mathbf{z}^{-v}\right) = \frac{p_{\text{FP}}^{n_0}\,(1-p_{\text{FP}})^{(n_1+n_\infty)}\;\alpha^{n_\infty}\;\left[\prod_{\{m\}_1} N_{\mathbf{z}_m}^{-v}\right]}{\prod_{m=1}^{(n_1+n_\infty)}\alpha+N-m}\,, \tag{9}$$

where the definitions of n_0, n_1, n_∞ are given in Eq. 3, and $\{m\}_1$ is the set of indices that are matched to existing targets (i.e., $n_1 = |\{m\}_1|$). To incorporate absence information, suppose we knew which K_v of the existing K lights are within the field of view, i.e., $\{k\}^v$ from Sect. 3.[2] This, together with $\mathbf{z}^v$, allows us to determine the detection indicator variables $\{\delta_k\}$ (Eq. 3) and their probabilities:

[2] The correct Bayesian approach is to integrate over the posterior distribution of each light's location, which is intractable. This can be approximated by sampling the locations, then averaging the subsequent computations. In practice we found that using the posterior mean was sufficient.

$$\mathbb{P}(\{\delta_k\}) = \prod_{k \in \{k\}^v} [p_{\mathrm{D}}(k)]^{\delta_k} \; [1 - p_{\mathrm{D}}(k)]^{1-\delta_k} \;, \tag{10}$$

where p_{D} is the (target-specific) detection probability defined in Eq. 15. We combine the additional information with the DPMM conditional distribution in a conceptually simple fashion:

$$\mathbb{P}_{\mathrm{FullView}}\left(\mathbf{z}^v \,\middle|\, \mathbf{z}^{-v}, \{k\}^v\right) \propto \mathbb{P}_{\mathrm{DPMM}}\left(\mathbf{z}^v \,\middle|\, \mathbf{z}^{-v}\right) \; \mathbb{P}(\{\delta_k\}) \; \mathbb{I}[\mathbf{z}^v \text{ satisfies OMPO}] \;. \tag{11}$$

The final term evaluates to 1 if the joint correspondence satisfies the OMPO assumption, and 0 otherwise. Hence by construction the correspondence variables sampled from this conditional distribution will incorporate the FN and OMPO constraints.

Although $\mathbb{P}_{\mathrm{FullView}}$ combines all the desired information, the inherent difficulty is hidden in the '$\propto$' sign. The distribution first needs to be normalized before we can sample from it, which is inefficient now because the support of the distribution is the set of correspondence vectors satisfying the OMPO assumption. The OMPO constraint fully couples the measurements' cluster assignments, and all assignments must be considered jointly, as depicted in Fig. 1(d). We have essentially reverted to the high branching factor of the MHT! (The exponential blowup of the hypothesis tree is still avoided by sampling.) In the Fig. 1 example, $\mathbb{P}_{\mathrm{FullView}}$ must be evaluated for 304 different values of $\mathbf{z}^v$, compared to the $4 \times 5 = 20$ required for the DPMM.

5.2 DPMM-Factored

A closer look at the nature of the OMPO violation suggests a potential approximation to $\mathbb{P}_{\mathrm{FullView}}$. In Fig. 1(c), the violation is caused by *only* the two right-most measurements; the two measurements on the left are not easily confusable with the others and hence are easy to handle from a data association perspective. This suggests coupling *only* those measurements that cause OMPO violations. More generally, suppose we can partition each view's set of measurements into 'violating' subsets, where all OMPO violations are contained within a single subset (with high probability). That is, a good partition has the property that any two measurements belonging to different subsets will have low probability of being assigned to the same cluster (and hence causing an OMPO violation). Let $\mathscr{P}$ denote such a partition, and let $\left\{\mathbf{z}_p^v\right\}_{p \in \mathscr{P}}$ denote the restrictions of $\mathbf{z}^v$ to each subset $p \in \mathscr{P}$. Then:

$$\mathbb{I}[\mathbf{z}^v \text{ satisfies OMPO}] \approx \prod_{p \in \mathscr{P}} \mathbb{I}\left[\mathbf{z}_p^v \text{ satisfies OMPO}\right] \;. \tag{12}$$

Returning to Fig. 1(c), the most refined partition contains three subsets, where the sole non-singleton contains the two right-most OMPO-violating measurements.

The other two terms in $\mathbb{P}_{\mathrm{FullView}}$ (Eq. 13) are product distributions that factor nicely according to $\mathscr{P}$. We therefore arrive at the following *factored* approximation:

$$\mathbb{P}_{\text{Factored}}\left(\mathbf{z}^{v} \,\middle|\, \mathbf{z}^{-v}, \{k\}^{v}\right) \propto \prod_{p \in \mathscr{P}} \mathbb{P}_{\text{DPMM}}\left(\mathbf{z}_{p}^{v} \,\middle|\, \mathbf{z}^{-vp}\right) \mathbb{P}\left(\{\boldsymbol{\delta}_{k}\}|_{p}\right) \mathbb{I}\left[\mathbf{z}_{p}^{v} \text{ OMPO}\right] . \quad (13)$$

This form makes clear that each factor can be normalized and sampled independently. With a good partition, this breaks up the large joint computation in DPMM-FullView into several smaller ones within each subset of $\mathscr{P}$. Using the partition described above for the concrete example in Fig. 1 gives us the sampling process depicted in Fig. 1(e), where only the OMPO-violating measurement pair is considered jointly. This results in computing $5+5+22=32$ values, which is slightly greater than DPMM (20) but significantly less than DPMM-FullView (304).

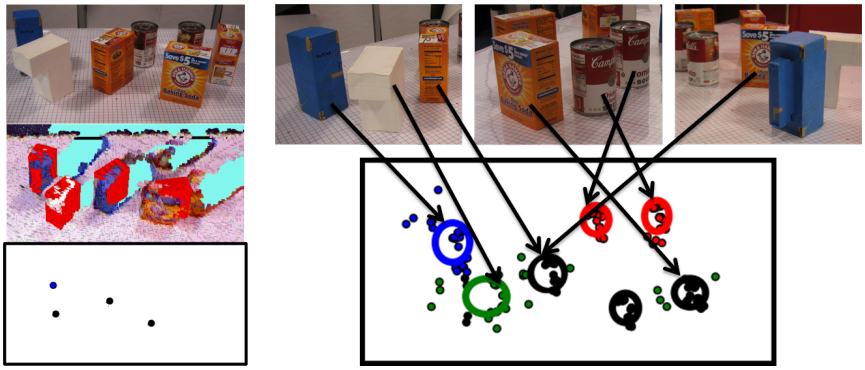
One issues remains: Where does the partition come from? This is crucial for all factored approximation: the aggressiveness of partitioning determines the trade-off between approximation error and efficiency. On one extreme, the DPMM model is similar to a fully-factored model (but does not take into account false negatives); on the other extreme, DPMM-FullView is equivalent to a one-set partition. The example in Fig. 1(c) once again provides an answer: 'violating' subsets can be found by examining clusters in the DPMM samples. Specifically, if measurements tend to be assigned to the same cluster across samples, then clearly they are strong violators and should be considered jointly. We therefore group measurements together if the proportion of samples in which they are assigned to the same cluster exceeds some threshold value. This proportion allows one to select an appropriate trade-off level.

6 Application to Object Type-and-Pose Estimation

As mentioned in Sect. 2, the colored-lights domain is representative of the semantic world-model estimation problem by considering lights as objects and locations and colors as attributes. Extension to additional attributes and higher-dimensional locations (3-D locations, 4-D or 6-D poses) is straightforward since the correspondence priors described in Sects. 3–5 do not depend on the observations. If attributes are independent, we simply take the product of their observation models when determining their posterior or predictive distributions, e.g., in Gibbs sampling (Eq. 7). Dependent attributes will need to be jointly considered as a single unit. For example, for pose estimates with non-diagonal error covariances, the normal-gamma prior needs to be replaced with a normal-Wishart prior.

As a proof of concept, we apply the discussed approaches to object type-and-pose estimation on tabletop scenes, illustrated in Fig. 2. This is similar to the colored-lights domain, where 'type' is equivalent to 'color', and 'pose' is a 3-D version of 'location'.[3] 3-D point cloud data was obtained from a Kinect sensor mounted on a mobile robot. A ROS perception service attempts to detect instances of the known shape models in a given point cloud. This is done by locating horizontal planes in the point cloud, finding clusters of points resting on the surface, and then doing

[3] For simplicity, we assume that the error covariance is axis-aligned and use an independent normal-gamma prior for each dimension, but it is straightforward to extend to general covariances.



(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 Kinect 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 circles centered around location estimate; color represents type, circle size reflects uncertainty). The estimate above identifies all types correctly with minimal error in pose.

stochastic gradient descent over the space of poses of the models to find one that best matches the cluster.[4] Example matches for a scene are illustrated in Fig. 2(a).

As shown, multiple instances of the same object type are present (increasing association difficulty), objects may be partially or fully occluded from a single viewpoint (cyan patches are occluded regions), object types can be confused (the white L-shaped block on the left), and pose estimates are noisy (the orange box in the center). Aggregation of object detections across different viewpoints and solving the subsequent data association issues, as depicted in Fig. 2(b), was therefore essential.

For our scenarios, objects of 4 distinct types were placed on a table. A robot moved around the table in a circular fashion, obtaining 20-30 views in the process. We constructed 12 scenes of varying object and occlusion density to test our approaches; results for 4 representative scenarios are described in the next section.

7 Results

Qualitative results for 4 representative scenarios are shown in Fig. 3. Images from above are for comparison convenience only; the camera's viewing height is much

[4] We thank Jared Glover and Sanja Popovic for the perception system implementation and support.

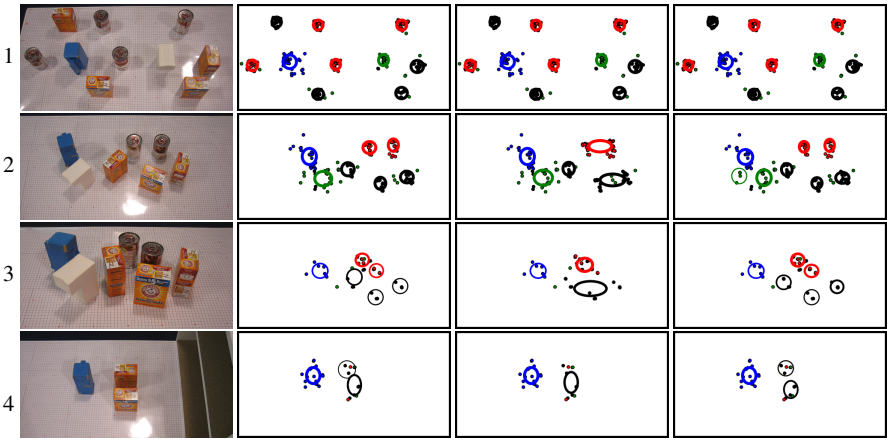


Fig. 3: Qualitative results for three world-model estimation approaches in four scenarios. The bird's-eye view of the scenes is for comparison convenience only; the actual viewing height is much closer to the table. The most likely hypothesis is shown for MHTF, and the maximum a posteriori sample is shown for the clustering-based approaches. Each small colored dot is a semantic (object type-and-pose) detection. Each target / cluster is depicted by an ellipse, centered at the posterior mean location. Ellipse axis lengths are proportional to the standard deviation in their respective dimensions. Ellipses are color-coded by the most likely posterior object type: red = red soup can, black = orange baking soda box, green = white L-shaped block, blue = blue rectangular cup. Line thickness is proportional to cluster size. See text in Sect. 7 for qualitative comparisons.

closer to the table height, as shown in Fig. 2(a), so in each view only a subset of objects is observable. We compare three approaches: multiple hypothesis tracking (**MHTF** from Sect. 3), generic DPMM clustering (**DPMM** from Sect. 4), and the factored approximation to DPMM-FullView (**DPMM-Factored** from Sect. 5.2). In Fig. 3, the most likely hypothesis is shown for MHTF, and the maximum a posteriori (MAP) sample (out of 100) is shown for the clustering-based approaches.

All approaches work well for scenario 1, where objects are spaced far apart. As objects of similar type are placed near each other, **DPMM** tends to combine clusters since it ignores the OMPO assumption. This is most apparent in scenario 3, where two soup cans (red) and three soda boxes (black) are combined into large clusters. By reconsidering the OMPO assumption, **DPMM-Factored** performs significantly better and is on par qualitatively with the **MHTF**, except for an extra cluster (bottom left, green) in scenario 2. In this case, the measurements corresponding to the white L-shaped object are dispersed, causing the shown extra-cluster error to be likely. Examining more samples reveals that a significant proportion (31%) do not have the extra cluster; they just happen not to be MAP samples. This means that the estimator has signficant uncertainty as to whether or not the extra object exists. Although in this case the **DPMM-Factored** MAP sample is wrong, it highlights a feature of our approach. Consider a task, e.g., grasping, that requires an accurate estimate of this object's neighborhood. Given the high uncertainty in the samples, the robot should decide to gather more observations of the region instead of operating based

Table 1: Average accuracy metrics and computation wall times for the four Fig. 3 scenarios. "**DPMM-Factored**" is denoted "Factored" below in the final row due to space constraints. Wall times are computed on a single core of an 2.66 GHz Intel Core i7 processor, using implementations in Python. Times are not provided for **Raw** since no processing on the measurements is required.

Metric →	Error (cm) in location estimate				% most likely type is correct				Num. missed objects (FNs)				Num. spurious clusters (FPs)				Computation wall time (s)			
Scenario	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Raw	2.5	2.7	1.9	2.1	98	93	67	56	2.0	3.7	5.4	2.0	0.8	1.3	0.3	0.7	N/A			
MHTF	2.1	2.8	**1.5**	2.6	**100**	**100**	**97**	**100**	**0.0**	**0.0**	**0.9**	0.6	0.6	**0.7**	**0.1**	0.6	492	263	833	3.0
DPMM	**1.8**	2.7	3.2	2.6	95	95	91	92	2.8	4.9	4.1	0.7	1.3	2.1	0.5	**0.1**	89	23	4	2.6
Factored	2.0	**2.6**	**1.5**	**2.0**	95	95	94	94	0.5	0.4	1.8	**0.2**	**0.2**	1.6	0.3	**0.1**	146	83	13	4.3

on the incorrect MAP sample. In contrast, the **MHTF** is over 90% certain of its estimate because most other possibilities have been pruned. Although **MHTF** would have been less certain as well if all hypotheses were retained during filtering, the necessary aggressive pruning tends to make **MHTF** overconfident in its estimates.

Scenario 4 highlights another related difference between the tracking filter and batch approaches. Here two closely-arranged orange boxes are placed near a shelf, such that from most views at most one of the two boxes can be seen. Only in the final views of the sequence can both be seen (imagine a perspective from the bottom-left corner of the image). Due to the proximity of the boxes, and the fact that at most one was visible in the early views, **MHTF** eventually pruned all the then-unlikely hypotheses positing that measurements came from two objects. When finally both are seen together, although a hypothesis with two orange boxes resurfaces, it is too late: the remaining association hypotheses already associate all previous measurements of the boxes to the same target, in turn giving an inaccurate location estimate. In contrast, **DPMM-Factored** re-examines previous associations (in the next sampling iteration) after the two boxes are seen together, and can correct such errors. One way to consider this difference is that **DPMM-Factored** is a batch algorithm, whereas **MHTF** is simply a forward filter and does not have this capability.

Quantitative metrics are given in Table 1, averaged over the association hypotheses for **MHTF** and over 100 samples (after discarding burn-in) for **DPMM** and **DPMM-Factored**. To evaluate predicted targets and clusters against our manually-collected ground truth, for each ground truth object, the closest cluster within a 5 cm radius is considered to be the estimate of the object. If no such cluster exists, then the object is considered missed; all predicted clusters not assigned to objects at the end of the process are considered spurious. We also compare against a baseline approach, **Raw**, that does not perform any data association. It uses the object types and poses perceived in each view directly as a separate prediction of the objects present within the visible field of view. The metrics in the table are evaluated for each view's prediction, and the **Raw** table rows show the average value over all views. The location and type metrics are only computed for clusters assigned to detected objects, i.e., the clusters whose number is being averaged in the third metric.

The need for aggregating measurements across views is exemplified by **Raw**'s tendency to miss objects or confuse their types within single views. **DPMM** overcomes the latter issue by clustering across views, but still misses many objects be-

cause it ignores the OMPO assumption and agglomerates nearby similar objects. **DPMM-Factored** approximately respects this constraint and performs significantly better, missing few objects while maintaining accuracy in the posterior type-and-pose estimates. Although quantitatively it is slightly behind **MHTF**, this extra improvement comes at a several-factor computational expense, and potentially introduces filtering-related overconfidence issues mentioned earlier.

8 Related Work

Cox and Leonard [7] first considered data association for world modeling, using an MHT approach as well, but for low-level sonar features. The motion correspondence problem, which is similar to ours, has likewise been studied by many (e.g., [8, 9]), but typically again using low-level geometric and visual features only. Perhaps most similar to our problem is the recent work of Elfring et al. [11], which considers attribute-based anchoring and semantic world modeling with an MHT approach.

To our knowledge, our application of clustering to semantic world modeling is novel. More generally, sampling-based approaches have been applied to data association ([9, 15]), and may be applicable to approximate our DPMM-FullView model.

The important role of objects in spatial representations and semantic mapping was explored by Ranganathan and Dellaert [16], although their focus was on place modeling and recognition. Anati et al. [1] have also used the notion of objects for robot localization, but did not explicitly estimate their poses.

Object recognition and pose estimation has received widespread attention from the computer vision and robotics communities. Hager and Wegbreit [12] provide a good review as well as a unique approach. For pose estimation from multiple viewpoints, active perception has also been popular recently (e.g., [10, 3]). Our work differs in that we place no assumptions on the choice of camera poses, and we focus on data association issues. Moreover, we emphasize that object type-and-pose estimation was only chosen as a concrete and familiar proof of concept application, and our framework is applicable to many other semantic attributes and tasks.

9 Discussion

We have presented several clustering-based data association approaches for estimating semantic world models. We use Dirichlet process mixture models (DPMM) as our underlying framework. However, DPMMs perform poorly in their generic form because they ignore a crucial view-level constraint. Two improvements were therefore developed by incorporating the constraint exactly and approximately respectively. In preliminary experiments based on tabletop object type-and-pose estimation, the latter approach (**DPMM-Factored**) achieved performance comparable to an existing tracking-based approach using a fraction of the computation time.

As discussed in the introduction, semantic world models are useful in many object-centric tasks, involving a diverse set of attributes. We are currently exploring applications involving attributes beyond object type and pose. To be truly applicable, world models must also cope with objects moving over extended periods of time. Extending our framework to handle temporal dynamics while maintaining tractability over long horizons is the subject of future work.

Appendix: Posterior and predictive distributions for a single light

In this appendix, we verify the claim from Sect. 2 that finding the posterior and predictive distributions on color and location for a single light is straightforward, given that we know which observations were generated by that light. Let $\{(o,x)\}$ denote the set of light color-location detections that correspond to a light with unknown parameters (c,l). Color and location measurements are assumed to be independent given (c,l) and will be considered separately. We assume a known discrete prior distribution $\pi \in \Delta^{(T-1)}$ on colors, reflecting their relative prevalence. Using the color noise model (Eq. 1), the posterior and predictive distributions on c are:

$$\mathbb{P}(c\,|\,\{o\}) \propto \left[\prod_o \phi_o^c\right] \times \pi_c\,; \quad \mathbb{P}\left(o'\,|\,\{o\}\right) = \sum_{c=1}^{T} \mathbb{P}\left(o'|c\right)\,\mathbb{P}(c|\{o\}) = \sum_{c=1}^{T} \phi_{o'}^c\,\mathbb{P}(c\,|\,\{o\})\,. \tag{14}$$

We can use this to find the light's probability of detection:

$$p_{\mathrm{D}} \triangleq 1 - \mathbb{P}\left(o'=0\,\middle|\,\{o\}\right) = 1 - \sum_{c=1}^{T} \phi_0^c\,\mathbb{P}(c\,|\,\{o\})\,. \tag{15}$$

Unlike the constant false positive rate p_{FP}, the detection (and false negative) rate is dependent on the light's color posterior.

For location measurements, we emphasize that both the mean μ *and* precision $\tau = \frac{1}{\sigma^2}$ of the Gaussian noise model is unknown. Modeling the variance as unknown allows us to attain a better representation of the location estimate's empirical uncertainty, and not naïvely assume that repeated measurements give a known fixed reduction in uncertainty each time. We use a standard conjugate prior, the distribution NormalGamma$(\mu,\tau;\lambda,\nu,\alpha,\beta)$. The typical interpretation of normal-gamma hyperparameters is that the mean is estimated from λ observations with mean ν, and the precision from 2α observations with mean ν and variance $\frac{\beta}{\alpha}$. It is well known (e.g., [5]) that after observing n observations with sample mean $\hat{\mu}$ and sample variance $\hat{s}^2$, the posterior is a normal-gamma distribution with parameters:

$$\lambda' = \lambda + n\,; \quad \nu' = \frac{\lambda}{\lambda+n}\nu + \frac{n}{\lambda+n}\hat{\mu}\,; \quad \alpha' = \alpha + \frac{n}{2}\,; \quad \beta' = \beta + \frac{1}{2}\left(n\hat{s}^2 + \frac{\lambda n}{\lambda+n}\left(\hat{\mu}-\nu\right)^2\right)\,. \tag{16}$$

Often we are only interested in the mean; the marginal distribution on μ is a t-distribution with mean ν, precision $\frac{\alpha\lambda}{\beta(\lambda+1)}$, and 2α degrees of freedom.

The upshot of using a conjugate prior for location measurements is that the marginal likelihood of location observations has a closed-form expression. The posterior predictive distribution for the next location observation x' is obtained by integrating out the latent parameters μ, τ, and has the following expression:

$$\mathbb{P}\left(x' \,|\, \{x\}\,; \lambda, \nu, \alpha, \beta\right) = \int_{(\mu,\tau)} \mathbb{P}(x|\,\mu,\tau)\,\mathbb{P}(\mu,\tau|\,\{x\}\,; \nu, \lambda, \alpha, \beta) = \frac{1}{\sqrt{2\pi}} \frac{\beta'^{\alpha'}}{\beta^{+\alpha^+}} \frac{\sqrt{\lambda'}}{\sqrt{\lambda^+}} \frac{\Gamma(\alpha^+)}{\Gamma(\alpha')}\,, \tag{17}$$

where the hyperparameters with '$'$' superscripts are updated according to Eq. 16 using the empirical statistics of $\{x\}$ only (excluding x'), and the ones with '$+$' superscripts are likewise updated but including x'. The ratio in Eq. 17 assesses the fit of x' with the existing observations $\{x\}$ associated with the light.

References

1. Anati, R., Scaramuzza, D., Derpanis, K., Daniilidis, K.: Robot localization using soft object detection. In: ICRA (2012)
2. Antoniak, C.: Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. The Annals of Statistics pp. 1152–1174 (1974)
3. Atanasov, N., Sankaran, B., Ny, J.L., Koletschka, T., Pappas, G., Daniilidis, K.: Hypothesis testing framework for active object detection. In: ICRA (2013)
4. Bar-Shalom, Y., Fortmann, T.: Tracking and Data Association. Academic Press (1988)
5. Bernardo, J., Smith, A.: Bayesian Theory. John Wiley (1994)
6. Cox, I., Hingorani, S.: An efficient implementation of Reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking. IEEE Trans. PAMI **18**(2), 138–150 (1996)
7. Cox, I., Leonard, J.: Modeling a dynamic environment using a Bayesian multiple hypothesis approach. AIJ **66**(2), 311–344 (1994)
8. Cox, I.J.: A review of statistical data association techniques for motion correspondence. IJCV **10**(1), 53–66 (1993)
9. Dellaert, F., Seitz, S., Thorpe, C., Thrun, S.: EM, MCMC, and chain flipping for structure from motion with unknown correspondence. Machine Learning **50**(1-2), 45–71 (2003)
10. Eidenberger, R., Scharinger, J.: Active perception and scene modeling by planning with probabilistic 6D object poses. In: IROS (2010)
11. Elfring, J., van den Dries, S., van de Molengraft, M., Steinbuch, M.: Semantic world modeling using probabilistic multiple hypothesis anchoring. RAS **61**(2), 95–105 (2013)
12. Hager, G., Wegbreit, B.: Scene parsing using a prior world model. IJRR **30**(12), 1477–1507 (2011)
13. Kurien, T.: Issues in the design of practical multitarget tracking algorithms. In: Y. Bar-Shalom (ed.) Multitarget-Multisensor Tracking: Advanced Applications, pp. 43–84. Artech House (1990)
14. Neal, R.: Markov chain sampling methods for Dirichlet process mixture models. Journal of Computational and Graphical Statistics **9**(2), 249–265 (2000)
15. Oh, S., Russell, S., Sastry, S.: Markov chain Monte Carlo data association for multi-target tracking. IEEE Trans. on Automatic Control **54**(3), 481–497 (2009)
16. Ranganathan, A., Dellaert, F.: Semantic modeling of places using objects. In: RSS (2007)
17. Reid, D.: An algorithm for tracking multiple targets. IEEE Trans. on Automatic Control **24**, 843–854 (1979)
18. Sethuraman, J.: A constructive definition of Dirichlet priors. Statistical Sinica **4**, 639–650 (1994)