Constructing Semantic World Models from Partial Views

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

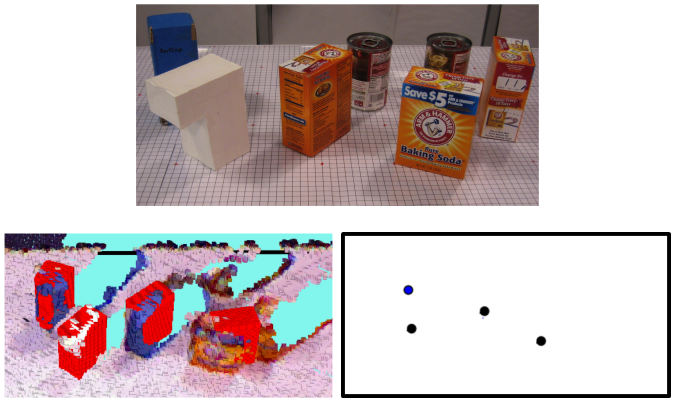


Fig. 1. Given a tabletop scene (top), we want to estimate the types and poses of objects in the scene using a black-box object detector. From a single Kinect RGB-D image, however, objects may be occluded or erroneously classified. The bottom left depicts a rendered image, with detections superimposed in red; three objects are missing due to occlusion, and two objects have been misidentified (second and fourth from left). The semantic attributes that result in our representation are very sparse (bottom right; dot location is measured 2-D pose, color represents type), but requires aggregation and association across many partial views in order to achieve estimates such as those in figure 2.

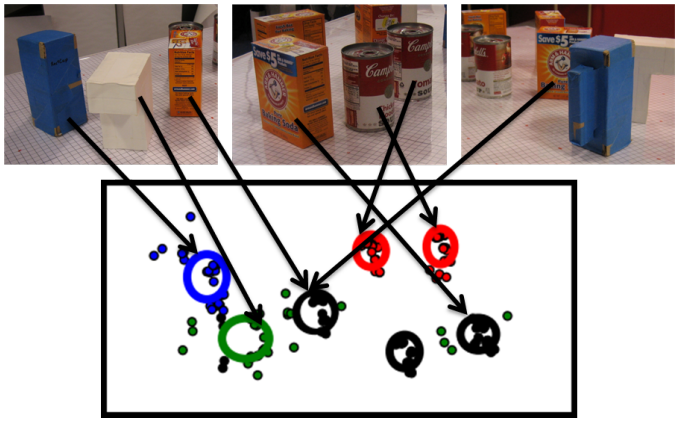


Fig. 2. A single viewpoint may be insufficient to identify all objects in a scene correctly (see figure 1). The natural solution is to observe the scene from different viewpoints, as depicted above. However, aggregating information from views of multiple objects introduces data association issues, especially 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.

***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. Behavior in these tasks is typically guided by goals supplied to task-level planners, which in turn assume a representation of the world in terms of objects. In this work, we explore the use of attribute-level perception to estimate high-level representations of the world. We run a black-box object detector in each range image, getting a set of detections of objects, labeled by their types and poses. We provide a formal description of a 1-D version of the problem, then develop three different solution approaches based on tracking, clustering, and a combination of the two. We evaluate the approaches empirically on data gathered by a robot moving around a table with objects on it, using a Kinect sensor to detect the objects from multiple viewpoints. We find that each of the methods performs better than using raw data, and that different methods perform best in different operational regimes.**

I. Introduction

Autonomous mobile-manipulation robots need to sense and interact with objects to accomplish high-level tasks such as preparing meals and searching for objects. Behavior in these tasks is typically guided by goals supplied to task-level planners, which in turn assume a representation of the world in terms of objects. Humans supplying goals will also refer to objects (fetch me a cup) or attributes (find the long plastic container), instead of using low-level geometric and visual features such as SIFT that are prevalent in recent robotic systems. Hence higher-level representations of the world at the level of semantic attributes and objects are necessary.

We address the problem of constructing high-level state representations of objects from multiple noisy observations (see figure 1). One strategy could be to perform the fusion of point clouds at the low level, and then do object segmentation and identification from a fused point cloud. This strategy however becomes fragile when the scene is large, views are cluttered, and possibly dynamic (objects may have moved) over long periods of time. Instead, we address the problem by performing information fusion from different views at a higher level of abstraction, where objects are the primitive entities, and the data are detections from an object detector.

In this work, we explore the use of attribute-level perception to estimate high-level representations of the world. We assume that we can run a black-box object detector in each image, getting detections of objects (types and poses). We assume that some low-level localization method is running sufficiently effectively that we can treat the robot's own pose estimates as being accurate, allowing us to put the object detections in a common coordinate frame. Then, using only the object type and pose measurements, we wish to construct a probabilistic estimate of the types and poses of the objects in the world, as illustrated in figure 2. Although we focus only on object type, the methods described in this work can be extended to incorporate other semantic attributes such as color or size.

In the following, we first state a formal model for a simplified 1-D version of the world model estimation problem in section II. Three different solution approaches based on tracking, clustering, and a combination of the two are then presented in sections III–V. Extensions to 3-D world model estimation is then briefly discussed in section VI, followed in section VII by experimental results using data collected with a Kinect sensor on a mobile robot.

II. The 1-D Colored Lights Domain

We first formalize the problem in a 1-D domain ($\mathbb{R}$). 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 on the real line. 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, which are known intervals $[a_v, b_v] \subset \mathbb{R}$. 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) measurements of the true lights. Also, a light may sometimes fail to be perceived (false negative). 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 we assume, for each color t, a known distribution $\phi^t \in \Delta^T$ that specifies how likely each color in $[T]$, or none at all, is observed:

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

A similar distribution ϕ^0 specifies the probability of observing each color given that the observation was a false positive.[1] False positives are assumed to occur in a p_{FP} proportion of object detections.[2] 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 of this distribution 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 range of the view (Unif$[a_v, b_v]$).

A. Data likelihood

We assume that views are independent given the hypothesized configuration $\{(c_k, l_k)\}_{k=1}^K$, hence the likelihood term is a product of V terms, one for each view. To reduce clutter, in the remainder we will focus on a single view.

Within each view, the correspondence between observations and lights (or false positives) is unknown, and it is useful to introduce latent variables to encode the correspondences. 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. Then:

$$\mathbb{P}\left(\{(o_m, x_m)\}_{m=1}^M \,\middle|\, \{(c_k, l_k)\}_{k=1}^K\right) \quad (2)$$
$$= \sum_{\{z_m\}} \mathbb{P}\left(\{(o_m, x_m)\} \,|\, \{z_m\}, \{(c, l)\}\right) \mathbb{P}\left(\{z_m\} \,|\, \{(c, l)\}\right)$$

By assuming that observations in a single view are independent given their correspondences $\{z_m^v\}$, and further that the color and location observations are independent:

$$\mathbb{P}\left(\{(o_m, x_m)\} \,|\, \{z_m\}, \{(c_k, l_k)\}\right) \quad (3)$$
$$= \prod_{m=1}^M \begin{cases} \phi_o^0 \cdot \text{Unif}\left[x\,;\, a_v, b_v\right], & z_m = 0 \\ \phi_o^{c_z} \cdot \mathcal{N}\left(x\,;\, l_z, \sigma_z^2\right), & z_m \in [K] \end{cases}$$

Here σ_z is unknown; a suitable prior for it will be given in section IV. Also, the above expression explicitly handles false positives only; false negatives (measurement is absent for a hypothesized light) will be handled in section V.

To combine the above equations, the final term in equation 2, $\mathbb{P}\left(\{z_m^v\} \,|\, \{(c_k, l_k)\}\right)$, needs to be resolved. This is the probability of a correspondence given only the configuration of lights (and other known parameters such as the view range $[a_v, b_v]$). Section III adapts a well-known multiple hypothesis tracking filter to this problem. Section IV gives an alternate clustering-based approach that is more tractable but arguably less realistic. Section V uses a more careful approach, borrowing ideas from the former approach in attempt to relax the less realistic assumptions of the latter.

B. Posterior and predictive distributions for a single light

Before examining approaches to solve for correspondences of measurements, we first consider the more straightforward problem of finding the posterior distribution on color and location for a single light, assuming we know exactly which observations correspond to the light. The developments in this section will be fundamental to all approaches discussed later.

Suppose we know that $\{(o, x)\}$ correspond to a light with unknown parameters (c, l). Since we assume independence between color and location, we can consider the two separately. We assume a known discrete prior distribution $\pi \in \Delta^{(T-1)}$ on colors, reflecting their relative prevalence. Using the color noise model (equation 1), the posterior distribution on c is:

$$\mathbb{P}\left(c \,|\, \{o\}\right) \propto \mathbb{P}\left(\{o\} \,|\, c\right) \mathbb{P}\left(c\right) \propto \left[\prod_o \phi_o^c\right] \cdot \pi_c \quad (4)$$

The posterior predictive distribution for the next color observation o', given that the observation is not a false positive, is obtained by summing over the latent color c:

$$\mathbb{P}\left(o' \,|\, \{o\}\right) = \sum_{c=1}^T \mathbb{P}\left(o'|c\right) \mathbb{P}\left(c|\{o\}\right) = \sum_{c=1}^T \phi_{o'}^c \mathbb{P}\left(c \,|\, \{o\}\right) \quad (5)$$

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

$$p_{\text{D}} \triangleq 1 - \mathbb{P}\left(o' = 0 \,|\, \{o\}\right) = 1 - \sum_{c=1}^T \phi_0^c \cdot \mathbb{P}\left(c \,|\, \{o\}\right) \quad (6)$$

[1] These distributions can be obtained from empirical perception apparatus statistics. Also, $\phi_0^0 = 0$ since it corresponds to an inconsistent measurement.

[2] Each view may have multiple detections and hence multiple false positives. The false positive rate is currently independent of camera pose and neighboring objects, an assumption that will be addressed in future work.

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 again that both the mean μ *and* precision τ of the Gaussian noise model is unknown. Modeling the variance as unknown allows us to attain a better representation of the inherent empirical uncertainty there is in the location estimate, and not naïvely assume that repeated measurements give a known fixed reduction in uncertainty each time. Since we are ultimately interested in the marginal distribution of the location estimate μ, the precision uncertainty will frequently need to be integrated out. Using a standard conjugate prior, the normal-gamma distribution $\text{NormalGamma}(\mu, \tau; \lambda, \nu, \alpha, \beta)$, will prove convenient. In this case, the marginal distribution on μ is a t-distribution with mean ν, precision $\frac{\alpha\lambda}{\beta(\lambda+1)}$, and 2α degrees of freedom.

To model the distribution of μ_k to be close to that of l_k initially, which we assume to be uniform over the explored range of the world, we use hyperparameters that are non-informative. 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 $\frac{\beta}{\alpha}$. Hence we set the initial $\lambda = 0$ and let ν be arbitrary since it will not affect the posterior (the posterior mean will simply be the empirical mean).

For a normal-gamma prior on (μ, τ) with hyperparameters $\lambda, \nu, \alpha, \beta$, 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:

$$\begin{aligned} \lambda' &= \lambda + n; \quad \nu' = \frac{\lambda}{\lambda+n}\nu + \frac{n}{\lambda+n}\hat{\mu} \\ \alpha' &= \alpha + \frac{n}{2}; \quad \beta' = \beta + \frac{1}{2}\left(n\hat{s}^2 + \frac{\lambda n}{\lambda+n}\left(\hat{\mu}-\nu\right)^2\right) \end{aligned} \tag{7}$$

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:

$$\begin{aligned} &\mathbb{P}\left(x' \,|\, \{x\} \;;\, \lambda, \nu, \alpha, \beta\right) \\ &= \int_{(\mu,\tau)} \mathbb{P}\left(x \,|\, \mu, \tau\right) \mathbb{P}\left(\mu, \tau \,|\, \{x\} \;;\, \nu, \lambda, \alpha, \beta\right) \\ &= \frac{1}{\sqrt{2\pi}} \frac{{\beta'}^{\alpha'}}{{\beta^+}^{\alpha^+}} \frac{\sqrt{\lambda'}}{\sqrt{\lambda^+}} \frac{\Gamma(\alpha^+)}{\Gamma(\alpha')} \end{aligned} \tag{8}$$

where the hyperparameters with '′' superscripts are updated according to equation 7 using the empirical statistics of $\{x\}$ only (excluding x'), and the ones with '+' superscripts are likewise updated but including x'. The ratio in equation 8 assesses the fit of x' with the existing observations $\{x\}$ associated with the light.

III. A Tracking-based Solution

If we consider the lights as stationary targets and the views as 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. A wide variety of approaches exist for this classic problem ([4]). Our problem setting has two interesting features that will restrict the choice of potential methods. First, estimation and prediction of target dynamics is unnecessary since the lights do not move. Second, the number of lights is unknown, so accounting for new targets and performing track initiation is crucial. Many tracking methods are track-oriented, focusing on tractably tracking object dynamics, often at the expense of the second requirement (by assuming that the number and initial parameters of tracks are known in advance), and hence are not suitable. We therefore opt for a multiple hypothesis filter ([15]), a measurement-oriented approach that considers all possible correspondences of each measurement, including the possibility of a new target.

Without loss of generality, assume that the views are in chronological order. A multiple hypothesis tracking algorithm maintains, at every timestep (view) v, a distribution over all possible associations to measurements of views up to v. For each view, let $\mathbf{z}^v$ be the concatenation of the view's latent correspondence variables $\{z_m^v\}_{m=1}^{M_v}$. The distribution at v is:

$$\begin{aligned} &\mathbb{P}\left(\left\{\mathbf{z}^j\right\}_{j=1}^{v} \,\middle|\, \{\{(o,x)\}\}_{j=1}^{v}\right) \\ &= \mathbb{P}\left(\mathbf{z}^v \,\middle|\, \left\{\mathbf{z}^j\right\}_{j=1}^{v-1}, \{\{(o,x)\}\}_{j=1}^{v}\right) \mathbb{P}\left(\left\{\mathbf{z}^j\right\}_{j=1}^{v-1} \,\middle|\, \{\{(o,x)\}\}_{j=1}^{v-1}\right) \\ &\propto \mathbb{P}\left(\{(o^v, x^v)\} \,\middle|\, \mathbf{z}^v, \left\{\mathbf{z}^j\right\}_{j=1}^{v-1}, \{\{(o,x)\}\}_{j=1}^{v-1}\right) \\ &\cdot \mathbb{P}\left(\mathbf{z}^v \,\middle|\, \left\{\mathbf{z}^j\right\}_{j=1}^{v-1}, \{\{(o,x)\}\}_{j=1}^{v-1}\right) \mathbb{P}\left(\left\{\mathbf{z}^j\right\}_{j=1}^{v-1} \,\middle|\, \{\{(o,x)\}\}_{j=1}^{v-1}\right) \end{aligned} \tag{9}$$

The first term is the likelihood of the current view's observations, the second is the prior on the current view's correspondences given previously identified targets, and the final term is the filter's distribution from the previous views.

The likelihood term for view v follows mostly from the derivation in section II-B. The observations are independent given the view's correspondence vector $\mathbf{z}^v$, and the likelihood is a product of M_v of the following terms:

$$\begin{aligned} &\mathbb{P}\left(o_m^v, x_m^v \,\middle|\, z_m^v = k, \left\{\mathbf{z}^j\right\}_{j=1}^{v-1}, \{\{(o,x)\}\}_{j=1}^{v-1}\right) \\ &= \begin{cases} \frac{\phi_o^0}{b_v - a_v}, & k = 0 \\ \mathbb{P}\left(o_m^v \,\middle|\, \{\{o\}\}_{z=k}^{1:v-1}\right) \mathbb{P}\left(x_m^v \,\middle|\, \{\{x\}\}_{z=k}^{1:v-1}\right), & k \neq 0 \end{cases} \end{aligned} \tag{10}$$

where $\{\{(o,x)\}\}_{z=k}^{1:v-1}$ refers to the observations in the previous views that were assigned to target k according to $\left\{\mathbf{z}^j\right\}_{j=1}^{v-1}$. In the last line, the two terms can be found from the posterior predictive distribution (equations 5, 8 respectively). For new targets (where k does not index an existing target), the conditioning set of previous observations will be empty, but can be likewise handled by the predictive distributions. The false positive probability ($k = 0$) is a direct consequence of the observation model (equation 3).

The prior on correspondences is due to Reid [15]. It assumes that we know which of the existing targets are within view based on the hypothesis on previous views, and can be found

by methods such as gating. Let the set $\{k\}^v$ denote the size-K_v set of target indices that we hypothesize are in view 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-light (OMPL) assumption. Based on these assumptions, we now define validity of correspondence vectors $\mathbf{z}^v$. First, by the OMPL 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. Invalid correspondences have probability 0.

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

$$\begin{aligned} n_0 &\triangleq \text{Number of false positives (0 entries)} \\ n_\infty &\triangleq \text{Number of new targets (non-existing indices)} \\ \delta_k &\triangleq \mathbb{I}\left\{\text{Target } k \text{ is detected } (\exists m.\, z_m^v = k)\right\}, k \in \{k\}^v \\ n_1 &\triangleq \text{Number of matched targets} = M_k - n_0 - n_\infty = \sum_k \delta_k \end{aligned} \tag{11}$$

Then we can split $\mathbb{P}(\mathbf{z}^v)$ by conditioning on these quantities:

$$\mathbb{P}(\mathbf{z}^v) = \mathbb{P}(\mathbf{z}^v \mid n_0, n_1, n_\infty, \{\delta_k\})\, \mathbb{P}(n_0, n_1, n_\infty, \{\delta_k\}) \tag{12}$$

By the assumed model characteristics, the second term is:

$$\begin{aligned} \mathbb{P}(n_0, n_1, n_\infty, \{\delta_k\}) = &\prod_{k \in \{k\}^v} p_{\text{D}}^{\delta_k}(k)\,(1 - p_{\text{D}}(k))^{1-\delta_k} \\ &\cdot \text{Bin}(n_0\,;\, M_v, p_{\text{FP}}) \cdot \text{Bin}(n_\infty\,;\, M_v, p_\infty) \end{aligned} \tag{13}$$

where p_∞ is the probability of a new target,and p_{D} is the (target-specific) detection probability defined in equation 6. Determining the correspondence given the quantities involves assigning $\mathbf{z}_m^v$ indices to the three groups of entries and matching $\{k\}^v$ to the indices in the corresponding group. A common assumption used in tracking is that all assignments and matches of indices are equally likely, so the first term in equation 12 is simply the reciprocal of the number of valid correspondence vectors given n_0, n_∞, $\{\delta_k\}$, given by:

$$\binom{M_k}{n_0, n_1, n_\infty} \cdot n_1! = \frac{M_k!}{n_0! n_1! n_\infty!} \cdot n_1! = \frac{M_k!}{n_0! n_\infty!} \tag{14}$$

By combining equations 10–14, along with the filter's distribution over association hypothesis for previous views (before v), we have derived all the expressions needed to use equation 9 to update the filter's distribution to include $\mathbf{z}^v$.

The main drawback of the multiple hypothesis filter is clearly the exponential growth in the hypothesis space. Viewing the set of hypotheses as a tree, at each step the branching factor is the number of valid correspondences:

$$\sum_{n_0=0}^{M_v} \sum_{n_\infty=0}^{M_v - n_0} \frac{M_k!}{n_0! n_\infty!} \cdot \frac{K_v!}{n_1!(K_v - n_1)!} \tag{15}$$

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 ([12, 7]), 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 section VII we simply keep hypotheses with probability above a threshold of 0.01.

IV. A Clustering-based Solution

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 ($[T] \times \mathbb{R}$), and estimates of the number of lights and their parameters can be derived from these clusters. 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 *a priori* limit the number of mixture components.

A recently popular model for performing clustering with an unbounded number of clusters is the Dirichlet process mixture model (DPMM) ([2, 13]), 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 $\text{DP}(\alpha, H)$, where H is the base distribution and α is the concentration hyperparameter (controlling the likeness 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 ([16]):

$$\begin{aligned} \theta &\sim \text{GEM}(\alpha) \\ (c_k, l_k) &\sim H \triangleq \pi \cdot \text{Unif}[A; B] \end{aligned} \tag{16}$$

where GEM is the distribution over stick weights θ. By defining $G(c, l) \triangleq \sum_{k=1}^{\infty} \theta_k \cdot \mathbb{I}\left[(c, l) = (c_k, l_k)\right]$, G is a distribution over the cluster parameters and is distributed as $\text{DP}(\alpha, H)$.

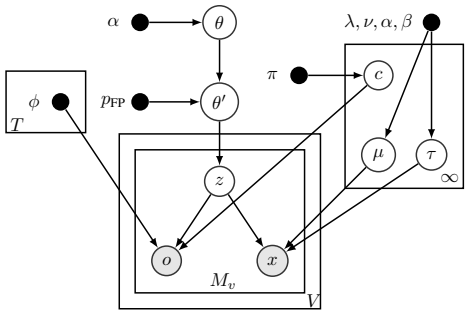


Fig. 3. Graphical model for DPMM-based solution; see section IV for details.

The graphical model of the generative procedure is depicted in figure 3. The remainder of the process is as follows:

$$\theta'_k = \begin{cases} p_{\text{FP}}, & k = 0 \\ (1 - p_{\text{FP}})\,\theta_k, & k \neq 0 \end{cases} \tag{17}$$

$$z_m^v \sim \theta', \qquad m \in [M_v], v \in [V]$$

$$\mu_k, \tau_k \sim \text{NormalGamma}(\nu, \lambda, \alpha, \beta)$$

$$o_m^v \sim \begin{cases} \phi^0, & z_m^v = 0 \\ \phi^{c_z}, & z_m^v \neq 0 \end{cases}; \quad x_m^v \sim \begin{cases} \text{Unif}[a_v, b_v], & z_m^v = 0 \\ \mathcal{N}\left(\mu_k, \tau_k^{-1}\right), & z_m^v \neq 0 \end{cases}$$

The most straightforward way to perform inference in a DPMM is by Gibbs sampling. In particular, we will 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:

$$\begin{aligned} &\mathbb{P}\left(z_m^v = k \,\middle|\, z^{-vm}, \{\{(o,x)\}\}\right) \\ &\propto \mathbb{P}\left(o_m^v, x_m^v \,\middle|\, z_m^v = k, z^{-vm}, \{\{(o,x)\}\}^{-vm}\right) \\ &\cdot \mathbb{P}\left(z_m^v = k \,\middle|\, z^{-vm}, \{\{(o,x)\}\}^{-vm}\right) \\ &\propto \mathbb{P}\left(o_m^v, x_m^v \,\middle|\, \{\{(o,x)\}\}_{z=k}^{-vm}\right) \mathbb{P}\left(z_m^v = k \,\middle|\, z^{-vm}\right) \end{aligned} \tag{18}$$

In the final line, the first term can be found from the posterior predictive distributions (equations 5, 8), 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:

$$\mathbb{P}\left(z_m^v = k \,\middle|\, z^{-vm}\right) = \begin{cases} (1 - p_{\text{FP}})\,\frac{N_k^{-vm}}{\alpha + N - 1}, & k \text{ exists} \\ (1 - p_{\text{FP}})\,\frac{\alpha}{\alpha + N - 1}, & k \text{ new} \\ p_{\text{FP}}, & k = 0 \end{cases} \tag{19}$$

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

By combining equations 18, 19, 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 (19) 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. Equation 4 applied with all data belonging to one cluster provides the posterior distribution on the light's color. The hyperparameter updates in equation 7 similarly gives the posterior joint distribution on the light's location and precision of the observation noise model.

V. A Local View Correspondence (VC) Problem

The DPMM-based solution to the colored lights problem is relatively straightforward, but it makes a few unrealistic assumptions. In this section, we attempt to correct three issues:

1) The known view range limits a_v, b_v provide hard constraints on the correspondences $\{z^v\}$, in that they can only correspond to clusters likely within range. This information is not used; the only term preventing an observation from being assigned to a far-away cluster is the Gaussian location observation model.
2) The possibility of false negatives, are likewise absent in the DPMM. This may lead to clusters being posited for spurious measurements when its absence in repeated measurements would have suggested otherwise.
3) The DPMM ignores the OMPL assumption described in section III. If we consider a scenario where two red lights are placed very close to each other, the DPMM may associate both to the same cluster, even if two measurements are observed in every view of the lights.

Fig. 4. The DPMM ignores the OMPL assumption and may merge clusters.

In the tracking approach, the prior on correspondences described in equations 12–14 handles all three issues. The first two problems require the correspondences $\{z^v\}$ to depend on more information, namely the view range and all cluster parameters respectively. The final problem is due to the strong independence assumptions made by the DPMM on $\{z^v\}$. The OMPL assumption creates strong exclusion dependencies within a single view, and is impossible to enforce in a DPMM. The solution then is to couple the correspondences within a single view and consider the joint correspondence $\mathbf{z}^v$. To address the first two problems, we allow $\mathbf{z}^v$ to depend on the view parameters and cluster locations, as shown in figure 5.

First suppose we knew which K_v of the existing K lights lie within the view, i.e., $\{k\}^v$ from section III. By combining the CRP model of assigning cluster weights in equation 19 and the correspondence prior used in tracking, we attempt at a reasonable definition of a conditional distribution on $\mathbf{z}^v$:

$$\mathbb{P}\left(\mathbf{z}^v \,\middle|\, \mathbf{z}^{-v}, \{\{(o,x)\}\}^{-v}, \{k\}^v\right) \tag{20}$$

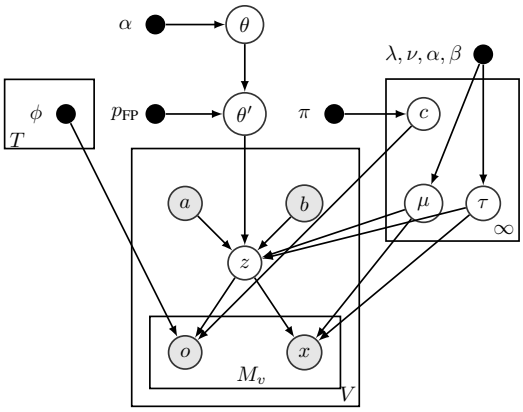


Fig. 5. Graphical model for the view correspondence extension to the DPMM; see section V for details. Also compare with the DPMM in figure 3.

Recall the definition of validity of correspondence vectors, and the definition of $n_0, n_1, n_\infty, \{\delta_k\}$ from equation 11. We account for false negatives of within-view clusters (targets) in the same way as in tracking (equation 13):

$$\mathbb{P}\left(\{\delta_k\}\right) = \prod_{k \in \{k\}^v} p_{\text{D}}^{\delta_k}(k) \left(1 - p_{\text{D}}(k)\right)^{1-\delta_k} \tag{21}$$

For the probability of $\mathbf{z}^v$ we use the DPMM instead of the tracking prior. We will use the CRP values given in equation 19 for each of the M_v indices. By exchangeability of the CRP, the probability will therefore be the same regardless of the order of the indices. This is convenient because the correspondence prior assumption remains valid, that all correspondences with the same n_0, n_1, n_∞ values (i.e., only involving a permutation of entries) are equally likely. The probability is:

$$\begin{aligned} &\left[\prod_{\{m\}_1} \frac{\left(1 - p_{\text{FP}}\right) N_{\mathbf{z}_m}^{-v}}{\alpha + N - n_1 - n_\infty + m - 1}\right] \\ &\cdot \left[\prod_{m=1}^{n_\infty} \frac{\left(1 - p_{\text{FP}}\right) \alpha}{\alpha + N - n_\infty + m - 1}\right] \cdot \left[\prod_{m=1}^{n_0} p_{\text{FP}}\right] \\ =&\frac{p_{\text{FP}}^{n_0} \left(1 - p_{\text{FP}}\right)^{(n_1 + n_\infty)} \alpha^{n_\infty} \prod_{\{m\}_1} N_{\mathbf{z}_m}^{-v}}{\prod_{m=1}^{(n_1+n_\infty)} \alpha + N - m} \end{aligned} \tag{22}$$

where $\{m\}_1$ is the set of indices that are matched to existing targets (i.e., $n_1 = |\{m\}_1|$). Note however that the expression above gives non-zero probability to invalid correspondence vectors as well (such as those those that do not satisfy the OMPL assumption), which we must disallow. Hence to achieve a distribution over valid correspondences, we define the conditional distribution 20 to be proportional to the product of equations 21, 22 for valid correspondences, and 0 otherwise.

To remove the assumption that we know $\{k\}^v$, we need to integrate it out using the posterior distribution on cluster locations after observing $\{\{x\}\}^{-v}$:

$$\begin{aligned} &\mathbb{P}\left(\mathbf{z}^v \,\middle|\, \mathbf{z}^{-v}, \{\{(o,x)\}\}^{-v}\right) \\ &= \int_{\{l_k\}} \mathbb{P}\left(\mathbf{z}^v \,\middle|\, \mathbf{z}^{-v}, \{\{o\}\}^{-v}, \{k\}^v\right) \mathbb{P}\left(\{l_k\} \,\middle|\, \{\{x\}\}^{-v}\right) \end{aligned} \tag{23}$$

The integral is deceptively simple but intractable even though we know that the locations have a t-distribution posterior. However, since it is straightforward to sample from t-distributions, we can compute the first term in the integral for every set of location samples $\left\{\hat{l}_k\right\}$, and average the result to produce a Monte Carlo estimate of the integral. The approximate conditional distribution for correspondences can then be used in conjunction with a likelihood term similar to equation 9 to give a conditional distribution similar to equation that in 18 for Gibbs sampling. In practice, because we limit estimation only to lights that are likely (above some threshold) to be in the view, and assuming that there are neither many measurements nor lights within a view (~ 5), brute-force enumeration remains tractable. More sophisticated techniques to sample correspondences exist ([9]) but were not considered.

VI. Application to World Model Estimation

Returning to world model estimation, the solutions above can be directly applied to object type and pose measurements by mapping them to the concepts of 'colors' and 'locations' respectively. Since we are interested in 3-D location estimates, and ultimately 4-D or 6-D poses, the approaches must be extended to handle higher-dimensional measurements. In all three methods, the observations only affect the probabilities through the observation model (equation 3); the correspondence priors do not depend on the observations. Hence we only need to extend the observation (location) model, of which a natural multivariate extension exists–a normal-Wishart prior.[3] As for attributes besides object type, if desired, it is again straightforward to treat them as independent and let the extended observation model be a product of the individual observation distributions, or to construct factored joint distributions (conditioned on the state) for dependent attributes.

VII. Results

We tested all three world model estimation approaches using a mobile robot with a Kinect sensor. The sensor yields three-dimensional point clouds; 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 stochastic gradient descent over the space of poses of the models to find the pose that best matches the cluster. Example matches for a scene are illustrated in figure 1.

Objects of 4 distinct types were placed on a table, as shown in the 6 scenarios of the left column of figure 6. Note that the bird's-eye view shown is for comparison convenience only; the camera's viewing height is much closer to the table height, as shown in figure 1, so in each view only a subset of objects is observable. As illustrated in the figure, objects may be partially or fully occluded, object types can be confused (the white L-shaped block on the left), and pose estimates are noisy (the orange box in the center). In all cases, one or two object types

[3] For simplicity in the current implementation, we will 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.

had multiple instances on the table to increase association difficulty. The robot moved around the table in a circular fashion, obtaining 20–30 views in the process.

Some qualitative results are shown in figure 6, showing the best hypothesis for tracking (**MHTF**) and the final sample for clustering (**DPMM**, **DPMM-VC**). All approaches work similarly well for the first two scenarios, where objects are spaced relatively far apart. As objects of similar type are placed near each other, **DPMM** tends to combine clusters since it ignores the OMPL assumption (which the other two methods satisfy). This is most apparent in the fifth scenario, where four nearby soup cans (red) are combined into one large cluster. Although this cluster has many points, the variance is large, from which we see the utility of our normal-gamma prior compared to a fixed observation variance. By monitoring the variance, a higher-level process could prompt the robot to take more views to try to obtain a more accurate estimate. In the last scenario, there is significant occlusion early in the sequence, which throws off MHTF, causing it to make incorrect associations which result in poor pose estimates.

Quantitative metrics are given in table I, averaged over the association hypotheses for **MHTF** and over 60 samples (after discarding burn-in) for **DPMM** and **DPMM-VC**. 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. **Raw** is a baseline approach 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 range. 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 first two metrics are only computed for clusters assigned to detected objects, i.e., the clusters whose number is being averaged in the third metric.

Ultimately for robot tasks, we are interested in the estimates of object types and poses, and we see from the first two metrics that all three data association approaches work better than the baseline in most scenarios. The differences in location estimate between the three approaches is not signficant except in the final scenario. For type estimates, **MHTF** has slightly better performance overall. As for detection characteristics considered by the final three metrics, we see that the baseline does significantly worse in the number of missed objects, which affects the number of correct clusters as well. Here we see that considering multiple views is beneficial, and further considering the correspondence problem in views helps even more. The clustering approaches tend to have more spurious clusters because we chose hyperparameters that encourage positing new clusters and faster exploration of the association space (high concentration parameters), but this can be corrected at the expense of convergence speed.

The final scenario highlights the risks of using a tracking filter. 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 consistently in the early views at most one was visible, **MHTF** eventually pruned all the then-unlikely hypotheses positing that the measurements came from two objects. When finally both can be 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-VC** is allowed to re-examine previous associations (in the next sampling iteration) after the two boxes are seen together, and hence does not suffer from this problem. One way to consider this difference is that **DPMM-VC** can essentially perform smoothing in the association space, whereas **MHTF** is simply a forward filter and does not have this capability.

TABLE I
AVERAGE ACCURACY METRICS FOR FIGURE 6 SCENARIOS

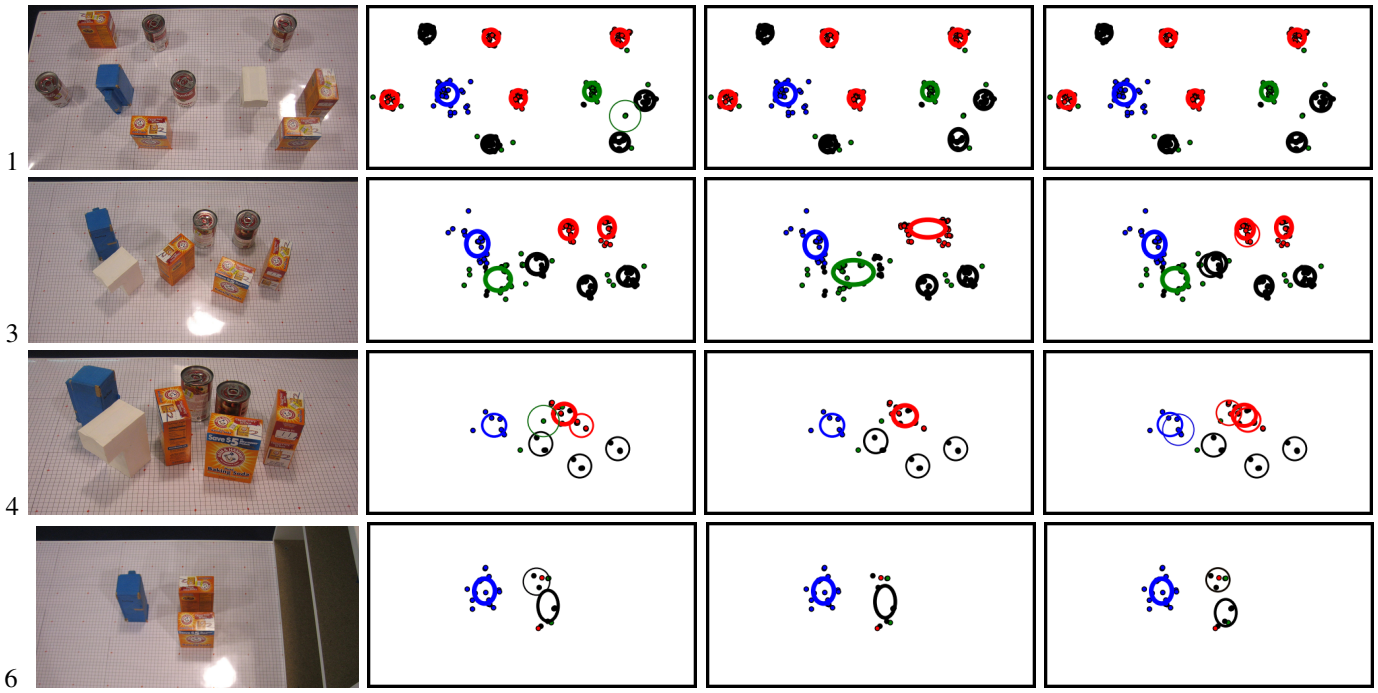
Metric	Method	1	2	3	4	5	6
Error in location estimate (cm)	Raw	2.54	3.20	2.69	1.90	2.24	2.07
	MHTF	2.04	2.17	2.78	1.89	**1.32**	2.64
	DPMM	**1.94**	**1.98**	2.64	2.17	1.51	2.83
	DPMM-VC	1.95	2.04	**2.63**	**1.82**	1.34	**2.02**
% most likely type is correct	Raw	98	93	93	67	85	56
	MHTF	**100**	**100**	**100**	**88**	**100**	**100**
	DPMM	95	95	95	**88**	92	92
	DPMM-VC	95	95	95	84	95	94
Num. clusters assigned to objects	Raw	8.0	4.6	3.3	1.6	5.3	1.0
	MHTF	**10.0**	**7.0**	**7.0**	**6.0**	**10.0**	2.4
	DPMM	9.2	6.6	5.5	4.6	7.2	2.3
	DPMM-VC	9.5	6.7	6.7	**6.0**	9.5	**2.8**
Num. spurious clusters	Raw	**0.8**	1.5	1.3	**0.3**	**0.1**	0.7
	MHTF	1.0	**0.3**	**0.4**	0.7	0.5	0.6
	DPMM	1.2	1.3	1.8	0.5	2.1	**0.1**
	DPMM-VC	2.4	1.3	2.2	3.1	2.5	**0.1**
Num. missed objects	Raw	2.0	2.4	3.7	5.4	4.7	2.0
	MHTF	**0.0**	**0.0**	**0.0**	**1.0**	**0.0**	0.6
	DPMM	0.8	0.4	1.5	2.4	2.8	0.7
	DPMM-VC	0.5	0.3	0.3	**1.0**	0.5	**0.2**

VIII. RELATED WORK

Cox and Leonard ([8]) first considered data association for world modeling, using a multiple hypothesis approach as well, but for low-level sonar features. The motion correspondence problem, which is similar to ours, has likewise been studied by many ([6, 9]), but typically again using low-level geometric and visual features only. For additional work in tracking and clustering, please refer to the respective sections (III, IV).

The important role of objects in semantic mapping was explored by Ranganathan and Dellaert ([14]), 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; instead, they used "soft" heatmaps of local image features as their representation.

Active perception has also been applied to object pose estimation in complex and potentially cluttered scenes (e.g., [10, 3]). This approach determines the next best view (camera pose) where previously occluded objects may be visible,



(a) Scene from above (b) MHTF (most likely hypothesis) (c) DPMM (final sample) (d) DPMM-VC (final sample)

Fig. 6. Qualitative results for the three approaches in six scenarios (four shown). The bird's-eye view of the scenes is for comparison convenience only; the actual viewing height is much closer to the table. The clusters are color-coded by the most likely posterior object type: red = red soup can, black = orange soda box, green = white L-shaped block, blue = blue rectangular cup. Thickness in lines is proportional to cluster size. See text in section VII for details.

typically by formulating the problem as a POMDP. Our work differs in that we place no assumptions on how camera poses were chosen, and we have emphasized data association issues.

Perhaps most similar to our problem and approach is the recent work of Elfring et al. ([11]), which considers attribute-based anchoring and world modeling, likewise with a multiple hypothesis approach. However, our application of DPMM clustering to the world modeling problem, as well as the view correspondence in section V, appears to be novel.

IX. Discussion

Through our exploration of three different approaches to the world model estimation problem, we have found that both a multiple hypothesis tracking filter and a Dirichlet process mixture model with view correspondence constraints perform very well, with complementary strengths in different scenarios. A generic Dirichlet process model is less robust and prone to over-association. From a practical standpoint, however, all three approaches perform similarly well when objects are either spaced sufficiently far apart or are not easily confusable. If that is the case, **DPMM** offers significant computational advantages, since in each view the computational time is linear in the number of observations, instead of the combinatorial expression in equation 15. Although the relative speeds depends on the difficulty of the scenario and the heuristics employed, in our implementation we observed that **DPMM** is always one to two orders of magnitude faster. Characterizing the regimes where each approach dominates more thoroughly and designing a scalable hybrid system that takes advantage of their differences is the subject of future work.

References

[1] R. Anati, D. Scaramuzza, K.G. Derpanis, and K. Daniilidis. Robot localization using soft object detection. In *ICRA*, 2012.

[2] C.E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, pages 1152–1174, 1974.

[3] N. Atanasov, B. Sankaran, J. Le Ny, T. Koletschka, G. Pappas, and K. Daniilidis. Hypothesis testing framework for active object detection. In *ICRA*, 2013.

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

[5] J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. John Wiley, 1994.

[6] I.J. Cox. A review of statistical data association techniques for motion correspondence. *IJCV*, 10(1):53–66, 1993.

[7] 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 Trans. PAMI*, 18(2):138–150, 1996.

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

[9] 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.

[10] R. Eidenberger and J. Scharinger. Active perception and scene modeling by planning with probabilistic 6D object poses. In *IROS*, 2010.

[11] 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.

[12] 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.

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

[14] A. Ranganathan and F. Dellaert. Semantic modeling of places using objects. In *RSS*, 2007.

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

[16] J. Sethuraman. A constructive denition of Dirichlet priors. *Statistical Sinica*, 4:639–650, 1994.