

Factored Pose Estimation of Articulated Objects using Efficient Nonparametric Belief Propagation

Karthik Desingh¹, Shiyang Lu¹, Anthony Opiari¹, Odest Chadwicke Jenkins¹

Abstract—Robots working in human environments often encounter a wide range of articulated objects, such as tools, cabinets, and other jointed objects. Such articulated objects can take an infinite number of possible poses, as a point in a potentially high-dimensional continuous space. A robot must perceive this continuous pose to manipulate the object to a desired pose. This problem of perception and manipulation of articulated objects remains a challenge due to its high dimensionality and multi-modal uncertainty. In this paper, we propose a factored approach to estimate the poses of articulated objects using an efficient nonparametric belief propagation algorithm. We consider inputs as geometrical models with articulation constraints, and observed RGBD sensor data. The proposed framework produces object-part pose beliefs iteratively. The problem is formulated as a pairwise Markov Random Field (MRF) where each hidden node (continuous pose variable) is an observed object-part’s pose and the edges denote the articulation constraints between the parts. We propose articulated pose estimation by Pull Message Passing algorithm for Nonparametric Belief Propagation (PMPNBP) and evaluate its convergence properties over scenes with articulated objects.

I. INTRODUCTION

Robots working in human environments often encounter a wide range of articulated objects, such as tools, cabinets, and other kinematically jointed objects. For example, the cabinet with three drawers shown in Fig 1 functions as a storage container. An robot would need to perform open and close actions on the drawers to accomplish storage and retrieval tasks. Accomplishing such tasks involves repeated sense-plan-act phases under uncertainty in the robot’s observations and demands pose estimation that accommodates uncertainty to inform a planner with current state of the world. This uncertainty poses the challenge of dealing with sensor noise and inherent environmental occlusions. Ability to perceive articulated pose under partial observations due to self and environmental occlusions makes the inference problem multi-modal. Further, the inference becomes a high-dimensional problem when the number of object parts grow in number.

Pose estimation methods have been proposed that take a generative approach to the problem [22], [2], [26]. These methods aim to explain a scene as a collection of object/parts poses, using a particle filter formulation to iteratively maintain belief over possible states in the form of particles. Though these approaches hold the power of modeling the world generatively, they have a inherent drawback of being slow with the increase in the number of rigid bodies. In this

paper, we focus on overcoming this drawback by factoring the state as individual object parts constrained by their articulations to create an efficient inference framework for pose estimation.

Generative methods exploiting articulation constraints are widely used in human pose estimation problems [17], [21], [24] where human body parts have constrained articulation. We take a similar approach and factor the problem using a Markov Random Field (MRF) formulation where each hidden node in the probabilistic graphical model represents an observed object-part’s pose (continuous variable), each observed node has the information about the object-part from observation and edges of the graph denote the articulation constraints between the parts. Inference on the graph is performed using a message passing algorithm that share the information between the parts’ pose variables, to produce their pose beliefs which collectively gives the state of the articulated object.

Existing message passing approaches [10], [20] represent message as a mixture of Gaussian components and provide Gibbs sampling based techniques to approximate message product and update operations. Their message representation and product techniques limits the number of samples used in the inference and is not applicable to our application domain. In this paper we provide a more efficient “pull” Message Passing algorithm for Nonparametric Belief Propagation (PMPNBP). The key idea of pull message updating is to evaluate samples taken from the belief of the receiving node with respect to the densities informing the sending node. The mixture product approximation can then be performed individually per sample, and then later normalized into a distribution. This pull updating avoids the computational pitfalls of push updating of message distributions used in [10], [20].

Our system takes in 3D point cloud as the sensor data and object geometry models in the form of an URDF (Unified Robot Description Format) as input and outputs belief samples in continuous pose domain. We use these belief samples to compute a maximum likely estimate to let the robot act on the object. We evaluate the performance of the system by quantifying over an articulated object on compelling scenes. Contributions of this paper include: a) proposal of an efficient belief propagation algorithm to estimate articulated object poses, b) discussion and comparisons with traditional particle filter as baseline, c) a belief representation from perception to inform a task planner. A simple task is illustrated to show how the belief propagation informs a task planner to choose an information gain action and overcome uncertainty in the perceptual estimation.

¹Department of Electrical Engineering and Computer Science, Robotics Institute, University of Michigan, Ann Arbor {kdesingh, shiyang, topipari, ocj}@umich.edu

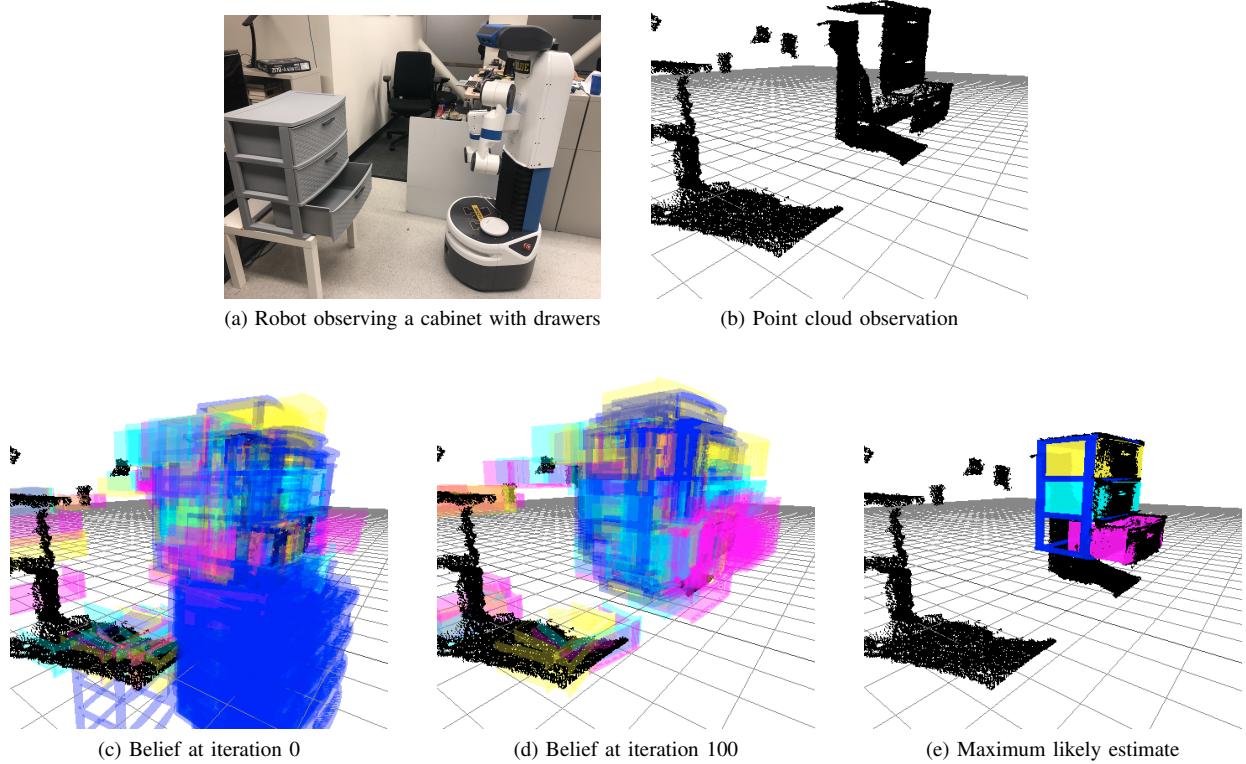


Fig. 1: Robot estimating the state of a cabinet with 3 prismatic articulated drawers from a 3D point cloud.

II. RELATED WORK

Existing methods in the literature have set out to address the challenge of manipulating articulated objects by robots in complex human environments. Particular focus has been placed on addressing the task of estimating novel articulated objects' kinematic models by a robot through interactive perception. Hausman et al. [7] propose a particle filtering approach to estimate articulation models and plan actions to reduce the model uncertainty. In [12], Martin et al. suggest an online interactive perception technique for estimating kinematic models by incorporating low-level point tracking and mid-level rigid body tracking with high-level kinematic model estimation over time. Sturm et al. [19], [18] addressed the task of estimating articulation models in a probabilistic fashion by human demonstration of manipulation examples.

All of these approaches discover the articulated object's kinematic model by alternating between action and sensing and are important methods to enable a robot to reliably interact with novel articulated objects. In this paper we assume that such kinematic models once learned for an object can be reused to localize their articulated pose under real world ambiguous observations. The method proposed in this paper could compliment the existing body of work towards task completion in unstructured human environment.

Probabilistic graphical model representations such as Markov random field (MRF) are widely used in computer vision problems where the variables take discrete labels

such as foreground/background. Many algorithms have been proposed to compute the joint probability of the graphical model. Belief propagation algorithms are guaranteed to converge on tree-structured graphs. For graph structures with loops, Loopy Belief Propagation (LBP) [13] is empirically proven to perform well for discrete variables. The problem becomes non-trivial when the variables take continuous values. Sudderth et.al (NBP) [20] and Particle Message Passing (PAMPAS) by Isard et.al [10] provide sampling approaches to perform belief propagation with continuous variables. Both of these approaches approximate a continuous function as a mixture of weighted Gaussians and use local Gibbs sampling to approximate the product of mixtures. NBP has been effectively used in applications such as human pose estimation [17] and hand tracking [21] by modelling the graph as a tree structured particle network. Scene understanding problems where a scene is composed of household objects with articulations demands large number of samples in the representation to handle the high-dimensional multimodal state space. The algorithm proposed in this paper produces promising results to handle such demands. We reported comparisons with existing NBP algorithm [10] in [3] with 2D examples.

Model based generative methods [14], [23], [25] are increasingly being used to solve scene estimation problems where heuristics from discriminative approaches [16], [5] are used to infer object poses. These approaches do not account

for object-object interactions or articulations and rely significantly on the effectiveness of recognition. Our framework doesn't rely on any prior detections but can benefit from them while inherently handling noisy priors [20], [10], [3]. Chua et. al [1] proposed a scene grammar representation and belief propagation over factor graphs, whose objective similar to ours for generating scenes with multiple-objects satisfying the scene grammars. This approach is similar to ours however, we specifically deal with 3D observations along with continuous variables.

III. PROBLEM STATEMENT

We consider an articulated object O to be comprised of N object-parts and $N - 1$ points of articulation. Such an object description conforms with the Unified Robot Description Format (URDF) commonly used in the Robot Operating System (ROS) [15]. Such a URDF-compliant kinematic model can be represented using an undirected graph $G = (V, E)$ with nodes V for object-part links and edges E for points of articulation. If G is a Markov Random Field (MRF), it has two types of variables X and Y that are, respectively, hidden and observed variables. Let $Y = \{Y_s \mid Y_s \in V\}$, where $Y_s = P_s \subseteq P$, with P being the point cloud observed by the robot's 3D sensor. Each object-part has an observed node in the graph G . P_s serves as a region of interest if a trained object detector is used to find the object in the scene, but is optional in our current approach. Each observed node Y_s is connected to a hidden node X_s that represents the pose of the underlying object part. Let $X = \{X_s \mid X_s \in V\}$, where $X_s \in \mathbb{H}_D$ is a dual quaternion pose of an object-part. Dual quaternions [4], [11] are a quaternion equivalent to dual numbers representing a 6D pose $X_s = (x, y, z, q_w, q_x, q_y, q_z)$ as $X_s = q_r + \epsilon q_d$ where q_r is the real component and q_d is the dual component. Alternatively it is represented as $X_s = [q_r][q_d]$. Constructing a dual quaternion X_s is similar to rotation matrices, with a product of dual quaternions representing translation and orientation as $X_s = dq_{pos} * dq_{ori}$, where $*$ is a dual quaternion multiplication. $dq_{ori} = [q_w, q_x, q_y, q_z][0, 0, 0, 0]$ is the dual quaternion representation of pure rotation and $dq_{pos} = [1, 0, 0, 0][0, \frac{x}{2}, \frac{y}{2}, \frac{z}{2}]$ is the dual quaternion representation of pure translation. This dual quaternion representation is widely used for rigid body kinematics, where the $*$ operation due to its efficiency and elegance compared with matrix multiplication. In addition to the representing the hidden variable X_s , dual quaternions can capture the constraints in the edges E and represent articulation types such as prismatic, revolute, and fixed effectively. This will be discussed in detail in Section IV-D.2.

Pose estimation of the articulated object involves inferring the hidden variables X_s that maximizes the joint probability of the graph G considering only second order cliques, which is given as:

$$p(X, Y) = \frac{1}{Z} \prod_{(s,t) \in E} \psi_{s,t}(X_s, X_t) \prod_{s \in V} \phi_s(X_s, Y_s) \quad (1)$$

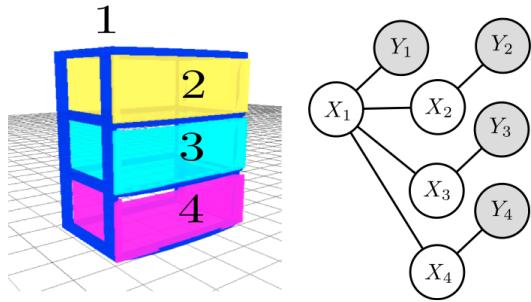


Fig. 2: Cabinet with 3 drawers connect to its frame is converted to a probabilistic graphical model with hidden nodes X_s representing the pose of the object-parts and observed nodes Y_s connected to each of the hidden nodes.

where $\psi_{s,t}(X_s, X_t)$ is the pairwise potential between nodes X_s and X_t , $\phi_s(X_s, Y_s)$ is the unary potential between the hidden node X_s and observed node Y_s , and Z is a normalizing factor. The problem is to infer belief over the possible articulation poses assigned to hidden variables X that are continuous such that the joint probability is maximized. This inference is generally performed by passing messages between hidden variables X until convergence of their belief distributions over several iterations. After converging over iterations, a maximum likelihood estimate of the marginal belief gives the pose estimate X_s^{est} of a object-part corresponding to the node in the graph G . The collection of all such object-part pose estimates form the entire object's pose estimate.

IV. NONPARAMETRIC BELIEF PROPAGATION

A. Overview

A message is denoted as $m_{t \rightarrow s}$ directed from node t to node s if there is an edge between the nodes in the graph G . The message represents the distribution of what node t thinks node s should take in terms of the hidden variable X_s . Typically, if X_s is in the continuous domain, then $m_{t \rightarrow s}(X_s)$ is represented as a Gaussian mixture to approximate the real distribution:

$$m_{t \rightarrow s}(X_s) = \sum_{i=1}^M w_{ts}^{(i)} \mathcal{N}(X_s; \mu_{ts}^{(i)}, \Lambda_{ts}^{(i)}) \quad (2)$$

where $\sum_{i=1}^M w_{ts}^{(i)} = 1$, M is the number of Gaussian components, $w_{ts}^{(i)}$ is the weight associated with the i^{th} component, $\mu_{ts}^{(i)}$ and $\Lambda_{ts}^{(i)}$ are the mean and covariance of the i^{th} component, respectively. We use the terms components, particles and samples interchangeably in this paper. Hence, a message can be expressed as M triplets:

$$m_{t \rightarrow s} = \{(w_{ts}^{(i)}, \mu_{ts}^{(i)}, \Lambda_{ts}^{(i)}) : 1 \leq i \leq M\} \quad (3)$$

Assuming the graph has tree or loopy structure, computing these message updates is nontrivial computationally. A message update in a continuous domain at an iteration n from a

Algorithm - Message update

Given input messages $m_{u \rightarrow t}^{n-1}(X_t) = \{(\mu_{ut}^{(i)}, w_{ut}^{(i)})\}_{i=1}^M$ for each $u \in \rho(t) \setminus s$, and methods to compute functions $\psi_{ts}(X_t, X_s)$ and $\phi_t(X_t, Y_t)$ point-wise, the algorithm computes $m_{t \rightarrow s}^n(X_s) = \{(\mu_{ts}^{(i)}, w_{ts}^{(i)})\}_{i=1}^M$

1. Draw M independent samples $\{\mu_{ts}^{(i)}\}_{i=1}^M$ from $bel_s^{n-1}(X_s)$.
 - (a) If $n = 1$ the $bel_s^0(X_s)$ is a uniform distribution or informed by a prior distribution.
 - (b) If $n > 1$ the $bel_s^{n-1}(X_s)$ is a belief computed at $(n-1)^{th}$ iteration using importance sampling.
2. For each $\{\mu_{ts}^{(i)}\}_{i=1}^M$, compute $w_{ts}^{(i)}$
 - a Sample $\hat{X}_t^{(i)} \sim \psi_{ts}(X_t, X_s = \mu_{ts}^{(i)})$
 - b Unary weight $w_{unary}^{(i)}$ is computed using $\phi_t(X_t = \hat{X}_t^{(i)}, Y_t)$.
 - c Neighboring weight $w_{neigh}^{(i)}$ is computed using $m_{u \rightarrow t}^{n-1}$.
 - (i) For each $u \in \rho(t) \setminus s$ compute $W_u^{(i)} = \sum_{j=1}^M w_{ut}^{(j)} w_u^{(ij)}$ where $w_u^{(ij)} = \psi_{ts}(X_s = \mu_{ts}^{(i)}, X_t = \mu_{ut}^{(j)})$.
 - (ii) Each neighboring weight is computed by $w_{neigh}^{(i)} = \prod_{u \in \rho(t) \setminus s} W_u^{(i)}$
 - d The final weights are computed as $w_{ts}^{(i)} = w_{neigh}^{(i)} \times w_{unary}^{(i)}$.
3. The weights $\{w_{ts}^{(i)}\}_{i=1}^M$ are associated with the samples $\{\mu_{ts}^{(i)}\}_{i=1}^M$ to represent $m_{t \rightarrow s}^n(X_s)$.

Algorithm - Belief update

Given incoming messages $m_{t \rightarrow s}^n(X_t) = \{(w_{ts}^{(i)}, \mu_{ts}^{(i)})\}_{i=1}^M$ for each $t \in \rho(s)$, and methods to compute functions $\phi_s(x_s, y_s)$ point-wise, the algorithm computes $bel_s^n(X_s) \propto \phi_s(x_s, y_s) \prod_{t \in \rho(s)} m_{t \rightarrow s}^n(X_t) = \{(w_s^{(i)}, \mu_s^{(i)})\}_{i=1}^T$

1. For each $t \in \rho(s)$
 - a Update weights $w_{ts}^{(i)} = w_{ts}^{(i)} \times \phi(X_s = \mu_{ts}^{(i)}, Y_s)$.
 - b Normalize the weights such that $\sum_{i=1}^M w_{ts}^{(i)} = 1$.
2. Combine all the incoming messages to form a single set of samples and their weights $\{(w_s^{(i)}, \mu_s^{(i)})\}_{i=1}^T$, where T is the sum of all the incoming number of samples.
3. Normalize the weights such that $\sum_{i=1}^T w_s^{(i)} = 1$.
4. Perform a resampling step followed by diffusion with Gaussian noise, to sample new set $\{\mu_s^{(i)}\}_{i=1}^T$ that represent the marginal belief of X_s .

node $t \rightarrow s$ is given by

$$m_{t \rightarrow s}^n(X_s) \leftarrow \int_{X_t \in \mathbb{H}_D} \left(\psi_{st}(X_s, X_t) \phi_t(X_t, Y_t) \prod_{u \in \rho(t) \setminus s} m_{u \rightarrow t}^{n-1}(X_t) \right) dX_t \quad (4)$$

where $\rho(t)$ is a set of neighbor nodes of t . The marginal belief over each hidden node at iteration n is given by

$$\begin{aligned} bel_s^n(X_s) &\propto \phi_s(X_s, Y_s) \prod_{t \in \rho(s)} m_{t \rightarrow s}^n(X_s) \\ bel_s^n &= \{(w_s^{(i)}, \mu_s^{(i)}, \Lambda_s^{(i)}) : 1 \leq i \leq T\} \end{aligned} \quad (5)$$

where T is the number of components used to represent the belief.

B. "Push" Message Update

NBP [20] provides a Gibbs sampling approach to compute an approximation of the product $\prod_{u \in \rho(t) \setminus s} m_{u \rightarrow t}^{n-1}(X_t)$. Assuming that $\phi_t(X_t, Y_t)$ is pointwise computable, a "pre-message" [9] is defined as

$$M_{t \rightarrow s}^{n-1}(X_t) = \phi_t(X_t, Y_t) \prod_{u \in \rho(t) \setminus s} m_{u \rightarrow t}^{n-1}(X_t) \quad (6)$$

which can be computed in the Gibbs sampling procedure. This reduces Equation 4 to

$$m_{t \rightarrow s}^n(X_s) \leftarrow \int_{X_t \in \mathbb{R}^b} \left(\psi_{st}(X_s, X_t) M_{t \rightarrow s}^{n-1}(X_t) \right) dX_t \quad (7)$$

NBP[20] sample $\hat{X}_t^{(i)}$ from the "pre-message" followed by a pairwise sampling where $\psi_{st}(X_s, X_t)$ is acting as $\psi_{st}(X_s | X_t = \hat{X}_t^{(i)})$ to get a sample $\hat{X}_s^{(i)}$.

The Gibbs sampling procedure in itself is an iterative procedure and hence makes the computation of the "pre-message" (as the Foundation function described for PAMPAS) expensive as M increases.

C. "Pull" Message Update

Given the overview of Nonparametric Belief Propagation above in Section IV-A, we now describe our "pull" message passing algorithm. We represent message as a set of pairs instead of triplets in Equation 3 which is

$$m_{t \rightarrow s} = \{(w_{ts}^{(i)}, \mu_{ts}^{(i)}) : 1 \leq i \leq M\} \quad (8)$$

Similarly, the marginal belief is summarized as a sample set

$$bel_s^n(X_s) = \{\mu_s^{(i)} : 1 \leq i \leq T\} \quad (9)$$

where T is the number of samples representing the marginal belief. We assume that there is a marginal belief over X_s as $bel_s^{n-1}(X_s)$ from the previous iteration. To compute the $m_{t \rightarrow s}^n(X_s)$, at iteration n , we initially sample $\{\mu_{ts}^{(i)}\}_{i=1}^M$ from the belief $bel_s^{n-1}(X_s)$. Pass these samples over to the neighboring nodes $\rho(t) \setminus s$ and compute the weights $\{w_{ts}^{(i)}\}_{i=1}^M$. This step is described in Algorithm - Message update. The computation of $bel_s^n(X_s)$ is described in Algorithm - Belief update. The key difference between the "push" approach of the earlier methods (NBP [20] and PAMPAS [10]) and our "pull" approach is the message $m_{t \rightarrow s}$ generation. In the "push" approach, the incoming messages to t determines the outgoing message $t \rightarrow s$. Whereas, in the "pull" approach, samples representing s are drawn from its belief bel_s from previous iteration and weighted by the incoming messages

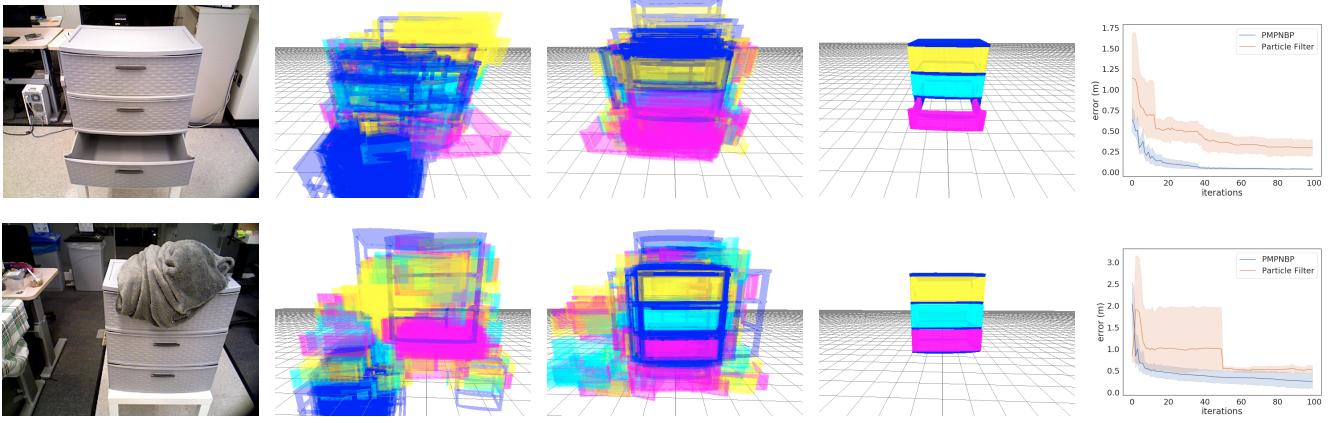


Fig. 3: Convergence of pose estimation on two different scenes: the first column shows the RGB image of each scene, second to fourth columns show the convergence results of PMPNBP. The second column shows randomly initialized belief particles, the third column shows the belief particles after 100 iterations, and the fourth column shows the maximum likely estimates of each part. The fifth column shows the estimation error (0.95 confidence interval) using PMPNBP with respect to the baseline particle filter method across 10 runs (400 particles and 100 iterations each). It can be seen that the baseline suffers from local minimas while PMPNBP is able to recover from them effectively.

to t . This weighting strategy is computationally efficient. Additionally, the product of incoming messages to compute bel_s is approximated by a resampling step as described in Algorithm - Belief update.

D. Potential functions

1) *Unary potential*: Unary potential $\phi_t(X_t, Y_t)$ is used to model the likelihood by measuring how a pose X_t explains the point cloud observation P_t . The hypothesized object pose X_t is used to position the given geometric object model and generate a synthetic point cloud P_t^* that can be matched with the observation P_t . The synthetic point cloud is constructed using the object-part's geometric model available *a priori*. The likelihood is calculated as

$$\phi_t(X_t, Y_t) = e^{\lambda_r d(P_t, P_t^*)} \quad (10)$$

where λ_r is the scaling factor, $d(P_t, P_t^*)$ is the sum of 3D Euclidean distance between the observed point $p \in P_t$ and rendered point $p^* \in P_t^*$ at each pixel location in the region of interest.

2) *Pairwise potential and sampling*: Pairwise potential $\psi_{t,s}(X_t|X_s)$ gives information about how compatible two object poses are given their joint articulation constraints captured by the edge between them. As mentioned in the Section 2, these constraints are captured using dual quaternions. Most often, the joint articulation constraints have minimum and maximum range in either prismatic or revolute types. We capture this information from URDF to get $R_{t|s} = [dq_{t|s}^a, dq_{t|s}^b]$ giving the limits of articulations. For a given X_s and $R_{t|s}$, we find the distance between X_t and the limits as $A = d(X_t, dq_{t|s}^a)$ and $B = d(X_t, dq_{t|s}^b)$, as well as the distance between the limits $C = d(dq_{t|s}^a, dq_{t|s}^b)$. Using a joint limit kernel parameterized by $(\sigma_{pos}, \sigma_{ori})$, we evaluate the pairwise potential as:

$$\psi_{t,s}(X_t|X_s) = e^{-\frac{(A_{pos} + B_{pos} - C_{pos})^2}{2(\sigma_{pos})^2} - \frac{(A_{ori} + B_{ori} - C_{ori})^2}{2(\sigma_{ori})^2}} \quad (11)$$

The pairwise sampling uses the same limits $R_{t|s}$ to sample for X_t given a X_s . We uniformly sample a dual quaternion \bar{X}_t that is between $[dq_{t|s}^a, dq_{t|s}^b]$ and transform it back to the X_s 's current frame of reference by $X_t = X_s * \bar{X}_t$.

V. EXPERIMENTS AND RESULTS

A. Experimental setup

We use Fetch robot, a mobile manipulation platform for our data collection and manipulation experiments. RGBD data is collected using an ASUS Xtion RGBD sensor mounted on the robot along with the intrinsic and camera to robot base transform. We use CUDA-OpenGL interoperation to render synthetic scenes on large set of poses in a single render buffer on a GPU. We render scenes as depth images, then project them back to 3D point clouds via camera intrinsic parameters.

B. Articulated Objects Models

We used a cabinet with three drawers as our articulated object in the experiment. CAD model of the object is obtained from the Internet and annotation of their articulations are performed on Blender to generate URDF models. Obtaining geometrical models and articulation models can either be crowd-sourced [6] or learned using human or robot interactions [12].

C. Baseline

We implemented Monte Carlo localization (particle filter) method that has object specific state representation. For example, the Cabinet with 3 drawers have state representation of $(x, y, z, \phi, \psi, \chi, t_a, t_b, t_c)$ where the first 6 elements describe the 6D pose of the object in the world and t_a, t_b, t_c represent the prismatic articulation. The measurement model in the implementation uses the unary potential described in the Section IV-D.1. Instead of rendering a point cloud of each object-part, the entire object in the hypothesized pose is rendered for measuring the likelihood. As the observations

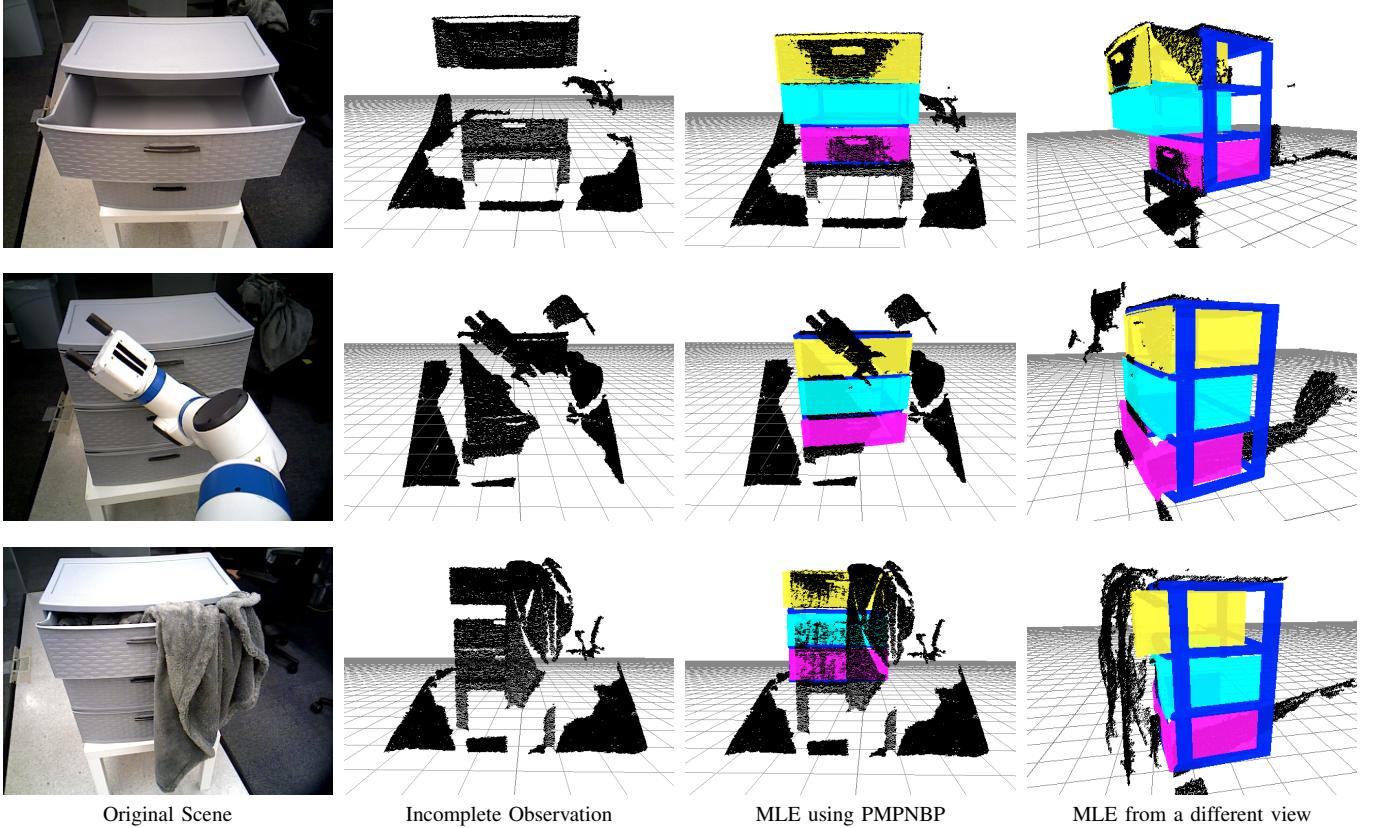


Fig. 4: Partial and incomplete observations due to self and environmental occlusions are handled by PMPNBP in estimating plausible pose with accuracy

are static, the action model in the standard particle filter is replaced with a Gaussian diffusion over the object poses.

D. Convergence Results

In the Figure. 3, we show the convergence of the proposed method visually for two scenes containing different point cloud observations. We collected point cloud observations of the objects in arbitrary poses and performed inference using both the proposed PMPNBP and the baseline Monte Carlo localization. Entire point cloud observed by the sensor is used as the observation for all the object-parts. The first column shows the scene (RGB not used in the inference). Second column shows the uniformly initialized poses of the object-parts on the entire point cloud. Third column shows the propagated belief particles for each object-part after 100 iterations. Fourth column shows the Maximum Likely Estimate (MLE) of each object-part using the belief particles from the third column.

For the results shown in Figure. 3, we ran our inference for 100 iterations with 400 particles representing the messages. 10 different runs are used to generate the convergence plot that shows the mean and variance in error across the runs. We adopt the average distance metric (ADD) proposed in [8], [25] for the evaluation. The point cloud model of the object-part is transformed to its ground truth dual quaternion (dq) and to the estimated pose's dual quaternion (\bar{dq}). Error is calculated as the pointwise distance of these transformation

pairs normalized by the number of points in the model point cloud.

$$ADD = \frac{1}{m} \sum_{p \in \mathcal{M}} \|\bar{dq} * p * \bar{dq}_c - dq * p * dq_c\| \quad (12)$$

where (\bar{dq}_c) and (dq_c) are the conjugates of the dual quaternions [4], [11], m is the number of 3D points in the model set \mathcal{M} .

E. Partial and incomplete observations

Articulated models suffer from self-occlusions and often environmental occlusions. By exploiting the articulation constraints of an object in the pose estimation, our inference method is able to estimate a physically plausible estimate that can explain the partial or incomplete observations. In Figure. 4 we show three compelling cases that indicates the strength of our inference method. In the first case, the drawer 1 heavily occludes the bottom drawers resulting in limited observations on drawer 2 and 3. PMPNBP is able to estimate a plausible pose given the constraints. In the second case, the cabinet is occluded by the robot's arm, while in the third case, a blanket from the drawer 1 occludes half of the object. PMPNBP is able to recover from these occlusions and produce a plausible estimate along with belief of possible poses.

The factored approach proposed in this paper scales to objects with higher number of links and joints with com-

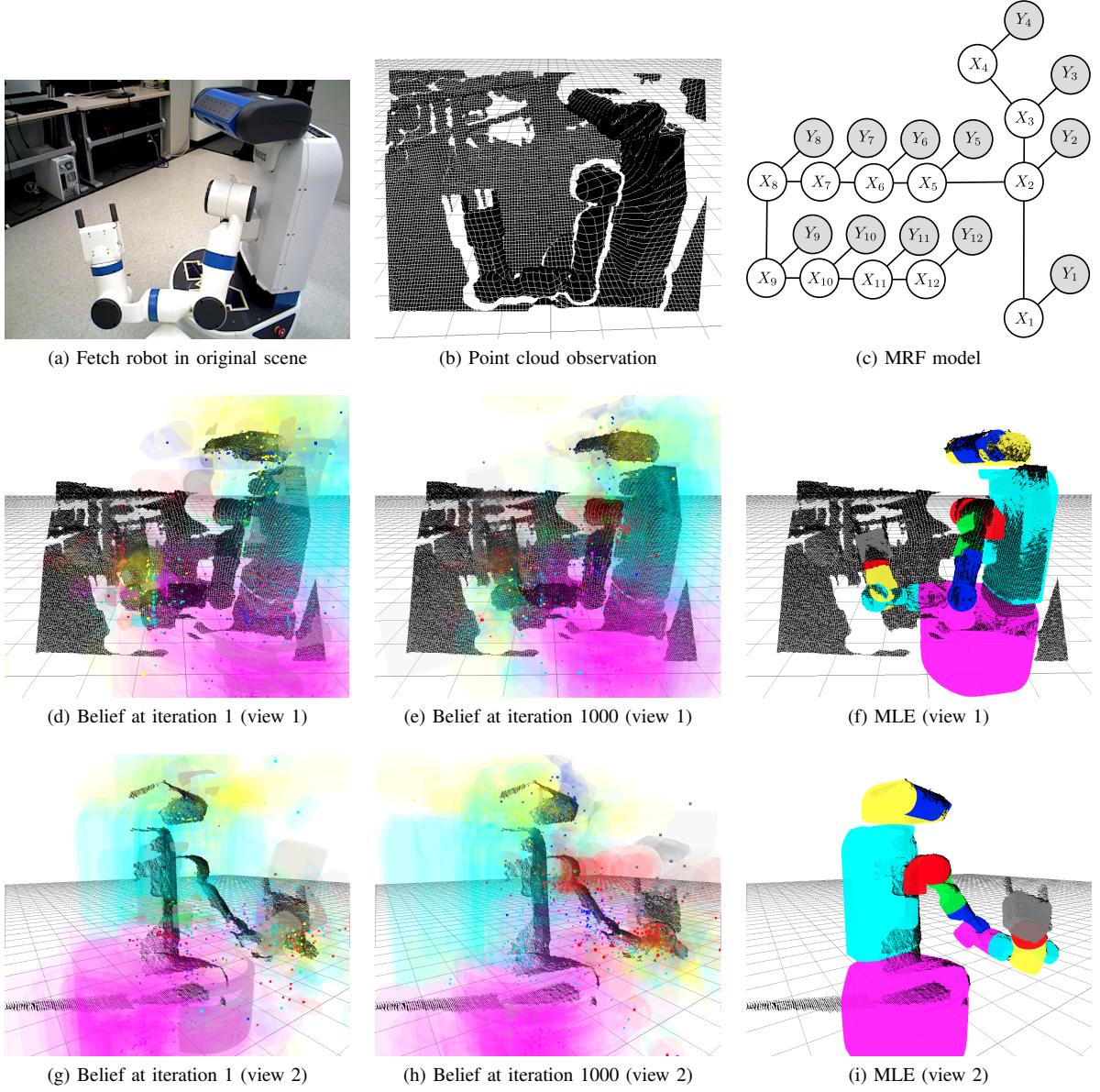


Fig. 5: Factored pose estimation using PMPNBP extends to articulated objects such as Fetch robot (a) which has 12 nodes and 11 edges in the probabilistic graphical model (c). For a scene (a), which has partial 3D point cloud observation (b), the PMPNBP message passing algorithm, propagates the belief samples from iteration 1 (d and g) to iteration 1000 (e and h), that leads to MLE (f and i).

binations of articulations. This is evaluated by estimating the pose of a Fetch robot that has 12 nodes and 11 edges in its graphical model. The graphical model is constructed using the URDF model of the robot. This is shown in Figure. 5(c) where the robot is observed using the a depth camera. Figure. 5(a & b) show the original scene and its point cloud observation with partial sensor data on the base, torso and the head of the robot. PMPNBP is able to estimate the pose of the robot by iteratively passing messages for 1000 iterations. Figures. 5(d-f) and Figures. 5(g-i) show the belief samples of the robot links at iteration 1 and 1000 followed by the most likely estimation (MLE) from two different view points for better visualization.

F. Benefits of maintaining belief towards planning actions

We show how the belief propagation approach aids in planning with a simple task illustration. Assume that the robot is performing a larger task of storing elements into the drawer 3. In a subtask, the goal is to open the drawer 3. With this setting (see Figure. 6) the robot is perceiving the current scene by estimating the pose of the cabinet, along with covariance on the belief for each part. We set a maximum threshold of 0.25cm on the standard deviation of (x, y, z) dimensions to decide if the estimation is certain or not. In this case, the standard deviation from the belief falls within this threshold and the robot is certain that the drawer 1 is open and drawer 3 is closed. Hence, the robot performs opening drawer 3 action. For the same task but with

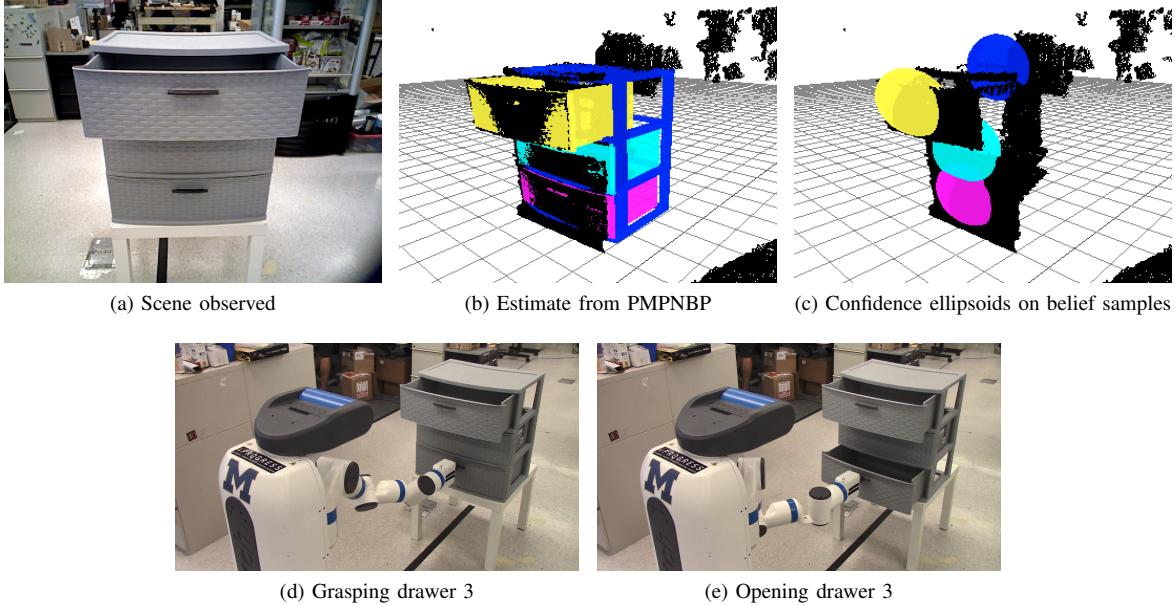


Fig. 6: Robot Experiments 1: The task for the robot is to open the drawer 3 (bottom) while having the drawer 1 open. The robot estimates the state of the object with certainty as shown in (b) with drawer 1 open and drawer 3 closed. In addition to the estimate, covariance can be calculated (shown as ellipsoids in (c) with 75% confidence interval). This could be used to decide that the estimation is certain with a threshold on the standard deviation on each of the dimensions of the pose. In this case the standard deviation of (x, y, z) falls below the threshold 0.25cm, and allows the robot to perform the opening action. (d-e) shows the robot performing opening action on drawer 3 using the estimate.

a different observation (see Figure. 7), the robot estimates the pose of the cabinet, along with its covariance. However, in this case, the robot is not certain about the estimation as the standard deviation is bigger than the threshold. This enables the robot to take an intermediate action (of lowering its torso) that provides a new observation of the cabinet. With this new observation, the robot perceives that the drawer 3 is closed with more certainty and performs an open action. This is an illustration of how the belief can be used in planning actions. More rigorous experiments with the choice of thresholds for different objects and tasks will be detailed in the future work.

VI. CONCLUSION

We proposed Pull Message Passing algorithm for Non-parametric Belief Propagation (PMPNBP), an efficient algorithm to estimate the poses of articulated objects. This problem was formulated as a graph inference problem for a Markov Random Field (MRF). We showed that the PMPNBP outperforms the baseline Monte Carlo localization method quantitatively. Qualitative results are provided to show the pose estimation accuracy of PMPNBP under a variety of occlusions. We also showed the scalability of the algorithm to articulated objects with higher number of nodes and edges in their probabilistic graphical models. In addition, we illustrated how belief propagation can benefit robot manipulation tasks. The notion of uncertainty in the inference is inevitable in robotic perception. Our proposed PMPNBP algorithm is able to accurately estimate the pose of articulated objects and maintain belief over possible poses that can benefit a robot in performing a task.

REFERENCES

- [1] J. Chua and P. F. Felzenszwalb. Scene grammars, factor graphs, and belief propagation. *arXiv preprint arXiv:1606.01307*, 2016.
- [2] K. Desingh, O. C. Jenkins, L. Reveret, and Z. Sui. Physically plausible scene estimation for manipulation in clutter. In *IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, pages 1073–1080, 2016.
- [3] K. Desingh, A. Oripipi, and O. C. Jenkins. Pull message passing for nonparametric belief propagation. *CoRR*, abs/1807.10487, 2018.
- [4] I. Gilitschenski, G. Kurz, S. J. Julier, and U. D. Hanebeck. A new probability distribution for simultaneous representation of uncertain position and orientation. In *Information Fusion (FUSION), 2014 17th International Conference on*, pages 1–7. IEEE, 2014.
- [5] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
- [6] S. R. Gouravajhala, J. Yim, K. Desingh, Y. Huang, O. C. Jenkins, and W. S. Lasecki. Eureca: Enhanced understanding of real environments via crowd assistance. 2018.
- [7] K. Hausman, S. Niekum, S. Osentoski, and G. S. Sukhatme. Active articulation model estimation through interactive perception. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3305–3312, May 2015.
- [8] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Asian conference on computer vision*, pages 548–562. Springer, 2012.
- [9] A. Ihler and D. McAllester. Particle belief propagation. In *Artificial Intelligence and Statistics*, pages 256–263, 2009.
- [10] M. Isard. PAMPAS: Real-valued graphical models for computer vision. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 613–620, 2003.
- [11] B. Kenwright. A beginners guide to dual-quaternions: what they are, how they work, and how to use them for 3d character hierarchies. 2012.
- [12] R. M. Martin and O. Brock. Online interactive perception of articulated objects with multi-level recursive estimation based on task-specific priors. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, Chicago, IL, USA, September 14-18, 2014*, pages 2494–2501, 2014.

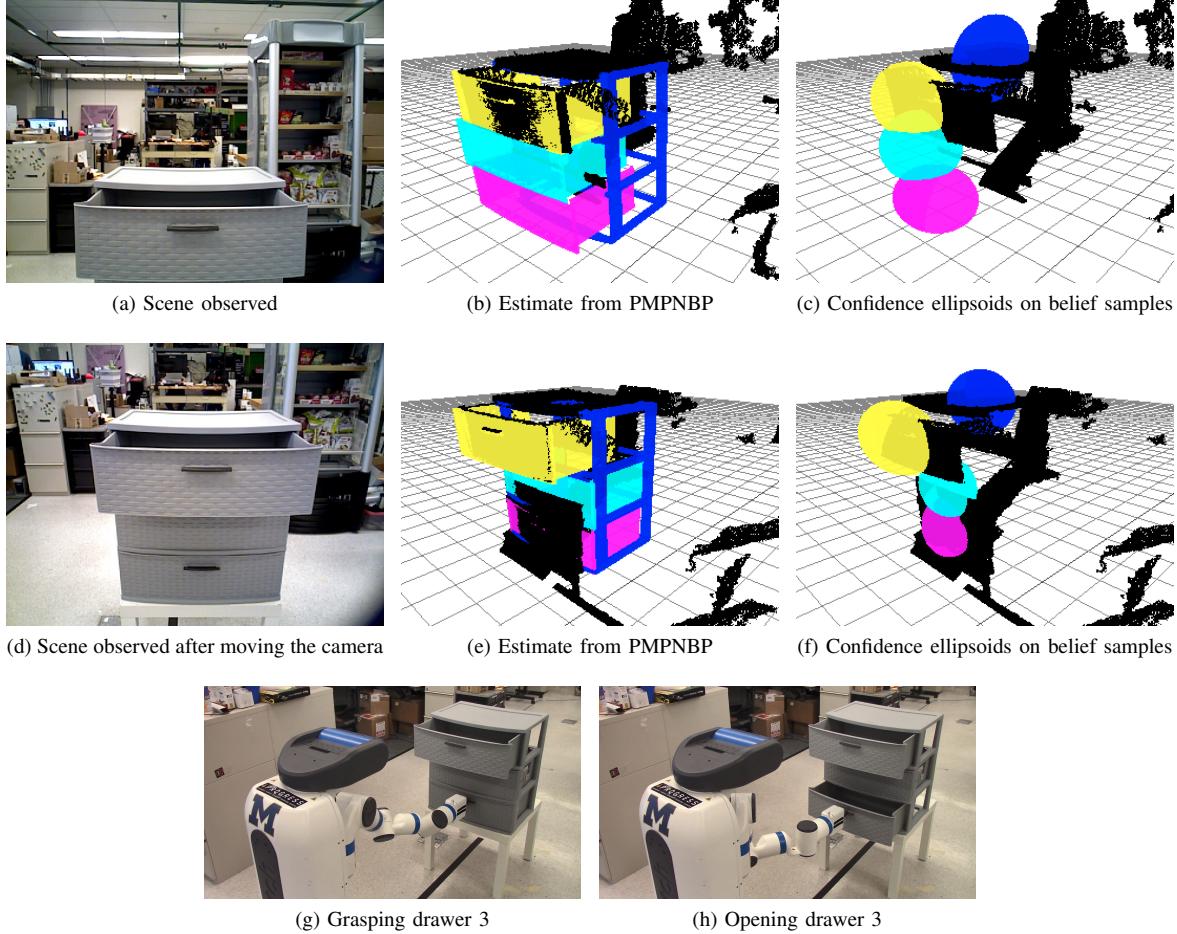


Fig. 7: Robot Experiments 2: The task remains the same as Figure. 6. However the robot is not directly observing the drawer 3 (a). The inference using PMPNPB gives an estimate (b). However, the standard deviation in the position (as seen in (c) with ellipsoids of larger radii), is higher than the threshold 0.25cm. To reduce the uncertainty in the estimation, the robot takes an intermediate action (changing the viewpoint) to gain more information about the scene. This results in new scene (d). Running inference using PMPNPB on this scene gives an estimate (e) with covariance ellipsoids as shown in (f). This satisfies the threshold on the standard deviations, enabling the robot to perform the (g-h) grasping and opening of drawer 3.

- [13] K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 467–475, 1999.
- [14] V. Narayanan and M. Likhachev. Discriminatively-guided deliberative perception for pose estimation of multiple 3d object instances. In *Robotics: Science and Systems*, 2016.
- [15] M. Quigley, K. Conley, B. P. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng. Ros: an open-source robot operating system. In *ICRA Workshop on Open Source Software*, 2009.
- [16] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015.
- [17] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard. Tracking loose-limbed people. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 421–428, 2004.
- [18] J. Sturm. *Approaches to Probabilistic Model Learning for Mobile Manipulation Robots*. Springer Tracts in Advanced Robotics (STAR). Springer, 2013.
- [19] J. Sturm, C. Stachniss, and W. Burgard. A probabilistic framework for learning kinematic models of articulated objects. *J. Artif. Intell. Res.*, 41:477–526, 2011.
- [20] E. B. Sudderth, A. T. Ihler, W. T. Freeman, and A. S. Willsky. Non-parametric belief propagation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, page 605, 2003.
- [21] E. B. Sudderth, M. I. Mandel, W. T. Freeman, and A. S. Willsky. Visual hand tracking using nonparametric belief propagation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04)*, pages 189–189, 2004.
- [22] Z. Sui, L. Xiang, O. C. Jenkins, and K. Desingh. Goal-directed robot manipulation through axiomatic scene estimation. *The International Journal of Robotics Research*, 36(1):86–104, 2017.
- [23] Z. Sui, Z. Zhou, Z. Zeng, and O. C. Jenkins. Sum: Sequential scene understanding and manipulation. In *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*. IEEE, 2017.
- [24] M. Vondrak, L. Sigal, and O. C. Jenkins. Dynamical simulation priors for human motion tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):52–65, 2013.
- [25] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017.
- [26] Z. Zeng, Z. Zhou, Z. Sui, and O. C. Jenkins. Semantic robot programming for goal-directed manipulation in cluttered scenes. In *IEEE/RSJ International Conference on Robotics and Automation (ICRA)*, 2018.