Analysis of Goal-directed Manipulation in Clutter using Scene Graph Belief Propagation

Karthik Desingh Anthony Opipari Odest Chadwicke Jenkins

***Abstract*— Goal-directed manipulation requires a robot to perform a sequence of manipulation actions in order to achieve a goal. A sequential pick-and-place actions can be planned using a symbolic task planner that requires current scene and goal scene represented symbolically. In addition to the symbolic representation, an object's pose in the robot's world is essential to perform manipulation. Axiomatic scene graph is a standard representation that consists of symbolic spatial relations between objects along with their poses. Perceiving an environment to estimate its axiomatic scene graph is a challenging problem especially in cluttered environments with objects stacking on top of and occluding each other. As a step towards perceiving cluttered scenes, we analyze the clutterness of a scene by measuring the visibility of objects. The more an object is visible to the robot, the higher the chances of perceiving its pose and manipulating it. As the visibility of the object is not directly computable, a measure of uncertainty on object pose estimation is necessary. We propose a belief propagation approach to perceive the scene as a collection of object pose beliefs. These pose beliefs can provide the uncertainty to allow or discard the planned actions from the symbolic planner. We propose a generative inference system that performs non-parametric belief propagation over scene graphs. The problem is formulated as a pairwise Markov Random Field (MRF) where each hidden node (continuous pose variable) is an observed object's pose and the edges (discrete relation variable) denote the relations between the objects. A robot experiment is provided to demonstrate the necessity of beliefs to perform goal-directed manipulation.**

I. Introduction

Autonomous robots must robustly perform goal-directed manipulation to achieve tasks specified by a human user. Goal-directed autonomy is a challenging field of research where robots must interpret a goal state and achieve it by performing sequence of manipulation actions. If initial and the goal states are represented symbolically, classical sequential planning algorithms [6], [20] can be used to plan a sequence of actions that achieves the goal state. In order to execute a planned action, the robot also requires object's pose in the robot's world. Axiomatic scene graph is a standard representation that encompasses the symbolic spatial relations between the objects along with their poses. However, perceiving an axiomatic scene graph is challenging especially in cluttered environments where the objects are occluded and only partially visible to the sensor. A classical planner assumes full perception of the scene while planning a sequence of actions. This assumption demands an accurate estimation of the scene from a perception system. This is an unrealistic assumption given the complexity of the typical human environments. An object's pose estimation depends on the observability/visibility of the object, which in turn affects the success of the manipulation. However, a perception system cannot measure the visibility of an object directly and hence should be able to provide a measure of uncertainty in the pose estimation. In this paper we propose a belief propagation based approach to estimate not only the pose of the objects but also a measure of uncertainty. This measure of uncertainty is used to allow or discard the actions planned by the symbolic planner.

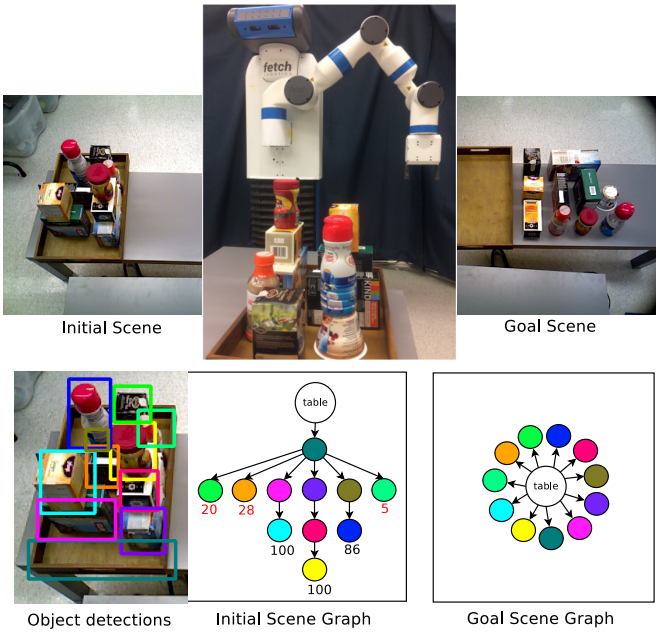


Fig. 1: Robot is observing a scene with objects in a tray. The goal of the object is to perform sequence of manipulation actions and achieve a goal scene. The bottom row shows object bounding boxes provided by a object detector, initial and goal scene graphs with object relations (child node *on* parent node) known a priori. The colors of the nodes in the scene graph corresponds to objects whose detections have the same color. The numbers below the leaf nodes on the initial scene graph corresponds to their percentage of visibility to the robot at this view point.

An example task is shown in Fig. 1 where a robot is given a tray full of objects and commanded to achieve a goal scene state where all the objects are on the table. If the scene graph is provided to the robot to simplify the goal-directed manipulation task, a planner can be used to generate a plan. As the planner doesn't have any information about the objects' pose in the world; all leaf nodes are equally likely

K. Desingh, A. Opipari, O.C. Jenkins are with the Department of Computer Science, University of Michigan, Ann Arbor, MI, USA, 48109-2121

to be part of the first pick-and-place action. But in the real world, only a some of these leaf node objects are visible. the objects whose visibilities are in red color are visible less than 30%. Hence, the perception system must output high uncertainty for these objects. In this fashion the perception system helps prune all the implausible actions.

We formulate the problem as a graphical model to perform belief propagation. The graphical model approach has been proven to perform well for articulated 3D human pose estimation [15] and visual hand tracking problems [17]. In such problems the inference leverages the probabilistic constraints posed on the articulations of body parts by the human skeletal structure. These structural constraints are modeled precisely with domain knowledge and do not change from scene to scene for human body. However, in the proposed framework we infer the pose of the objects using their scene specific relationship. For example, object A and B can have an *on* relationship in a current scene, however it is not a universal relationship to hold. But object A and B, not inter-penetrating each other is universal physical relationship to hold. Potentials defined on these relationships are generalized to perform for different scene graphs that changes scene to scene along with actions performed. Ideally an inference algorithm for goal-directed manipulation should produce axiomatic scene graph [18]. This paper however, is focused only on estimating the object poses. The problem is formulated as a Markov random field (MRF) where each hidden node represents an observed object's pose (continuous variable), each observed node has the information about the object from observation (discrete) and the edges of the graph denote the known relationships between object poses.

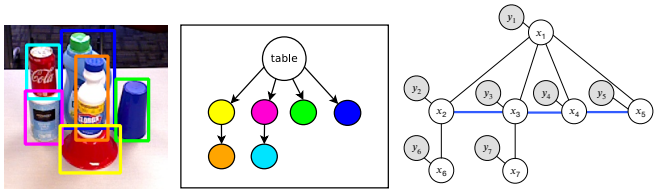
The main contribution of this paper is a generative scene understanding approach designed to support goal-directed manipulation. Our approach models inter-object relations in the scene and produces belief over object poses through nonparametric belief propagation (NBP)[16]. We detail the adaptation of NBP for scene estimation domain. In addition to this, we propose a measure to analyze the clutterness of a scene in terms of objects' visibility.

II. Related work

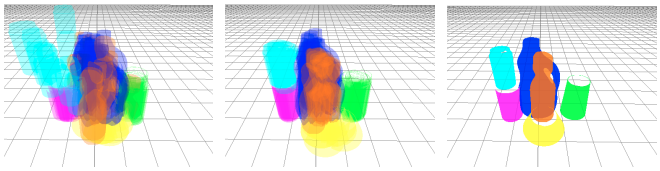
Our aim is to compute a scene estimate that allows a robot to use sequential planning algorithms, such as STRIPS [6] and SHRDLU [20]. A scene graph representation is one way to describe the world. However, the information on objects' poses should be perceived precisely to compliment the planning algorithms in order to let the robot make a physical change in the world. Chao et al. [1] perform taskable symbolic goal-directed manipulation by associating observed robot percepts with knowledge categories. This method uses background subtraction to adaptively build appearance models of objects and obtain percepts but is sensitive to lighting and object color. Narayanaswamy et al. [12] perform scene estimation and goal-directed robot manipulation in cluttered scenes for flexible assembly of structures. In contrast use the scene graph prior to run a perception system that estimates the object poses while maintaining the scene graph relations. Sui et. al [18] attempt to estimate a scene in a generative way; however, physical interactions are indirectly constrained by collision checks on object geometries. We avoid these expensive collision checks on object geometries and use potential functions that softly constraints inter-object collisions. Dogar et al. [5] use physics simulations and reasoning to accomplish object grasping and manipulation in clutter. However, their work does not include the object-object interactions in their clutter environments. The method proposed in this paper does not depend on a physics engine and also avoids explicit collision checking. Collet et al. [4] and Papazov et al. [13] apply a bottom up approach of using local features and 3D geometry to estimate the pose of objects for manipulation. In our work we product not only the object pose estimates but also belief over the object poses, which gives the confidence on the estimation. ten Pas and Platt [19] propose a model free approach localize graspable points in highly unstructured scenes of diverse unknown objects, which is directed towards pick and drop applications and doesn't apply to pick and place context. Model based robot manipulation as described in [3] can benefit from a scene graph based belief propagation proposed in this paper.

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) [10] is empirically proven to perform well for discrete variables. The problem becomes non-trivial when the variables take continuous values. Sudderth et.al (NBP) [16] and Particle Message Passing (PAMPAS) by Isard et.al [8] provide sampling approaches to perform belief propagation with hidden variables that take continuous values. Both of these approaches approximate a continuous valued function as a mixture of weighted Gaussians and use local Gibbs sampling to approximate the product of mixtures. This has been effectively used in applications such as human pose estimation [15] and hand tracking [17] by modeling the graph as a tree structured particle network. This approach has not been applied to scene understanding problems where a scene is composed of household objects with no strict constraints on their interactions. In this paper we propose a framework that can compute belief over object poses with relaxed constraints on their relations using a scene graph.

Model based generative methods [11] are increasingly being used to solve scene estimation problems where heuristics from discriminative approaches such as Convolutional Neural Networks (CNNs) [14], [7] are utilized to infer object poses. These approaches don't account for object-object interactions and rely significantly on the effectiveness of recognition. Our framework doesn't completely rely on the effectiveness of training the recognition system and can handle noisy priors as long as the priors have 100% recall.



(a) Observed detections (b) Scene graph prior (c) Markov random field



(d) Initial belief samples (e) Final belief samples (f) Final estimate

Fig. 2: Graphical Model (c) is constructed using the scene graph prior along the observed 2D object detections derived from the RGBD sensor data. In the constructed graphical model the edges denote the type of relation between the objects: black for support relations, cyan: non-colliding relation. The colors in (b) correspond to the object bounding box colors in (a). Inference on this graph is initialized with the object bounding boxes and the corresponding point clouds. Initial belief samples are shown in (d) and the inference iteratively propagates these beliefs to (e). The final estimate is shown in (f) using a post processing step using Iterative Closest Point (ICP) (discussed later in the paper).

Recent works such as [21] propose systems that can work on wild image data and refine the object detections along with their relations. However these methods do not consider the continuous pose in their estimation and work in pixel domain. Chua et. al [2] propose a scene grammar representation and belief propagation over factor graphs, whose objective is in generating scenes with multiple-objects satisfying the scene grammars. We specifically deal with RGBD observations and infer with continuous variables pose variables whereas Chua et. al [2] applies to RGB observations and discrete variables.

III. Methodology

A. Problem Statement

A robot observes a scene using an RGBD sensor, which gives an RGB image I and a point cloud P. An object detector takes I as an input and produces object bounding boxes $B = \{b_1, b_2, ..., b_k\}$ and corresponding confidence scores $C = \{c_1, c_2, ..., c_k\}$, with k being the number of detections. We make an assumption that the detector has a 100% recall.Using these detections, an undirected graph $G = (V, E)$ is constructed with nodes V and edges E. For each unique object label from the object detection result, there exists a corresponding observed node in the graph G. Let $Y = \{y_s \mid y_s \in V\}$ denote the set of observed variables, where $y_s = (l_s, B_s, C_s)$, with detections $B_s \subseteq B$ and confidences $C_s \subseteq C$ from the object detector corresponding to object label $l_s \in L$. Each observed node is connected to a hidden node that represents the pose of the underlying object. Let $X = \{x_s \mid x_s \in V\}$ denote a set of hidden variables, where $x_s \in \mathbb{R}^d$ with d being the dimensions of the pose of the object. This graph G consists of N hidden nodes if there are N objects with labels $L = \{l_1, l_2, ..., l_N\}$ present in the scene. However, $k \geq N$ as there could be multiple detections with the same label. G represents a scene with the observed and hidden nodes. Scene estimation involves finding this graph structure along with the inference on the hidden nodes. In this paper we assume that the graph structure is known apriori. This known graph structure in the form of scene graph provides the edges E showing the relations between the hidden nodes. The edges in our problem represent the relation between the objects in the scene. More precisely, we have two types of edges: one showing that an object is supporting/supported by another object (dark black in Fig 2), the other one indicating that an object is not in contact with another object (cyan in Fig 2). The joint probability of this graph considering only second order cliques is given as:

$$p(x, y) = \frac{1}{Z} \prod_{(s,t)\in E} \psi_{s,t}(x_s, x_t) \prod_{s\in V} \phi_s(x_s, y_s) \quad (1)$$

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. Pairwise potential can be modeled using the type of edges to perform the inference over this graph. We use Nonparametric Belief Propagation (NBP) [16] to pass messages over continuous variables and perform inference on a loopy graph structure such as ours (see Fig 2).

After converging over iterations, the maximum likelihood estimate of this marginal belief gives the pose estimate x_s^{est} of the object corresponding to the node in the graph G. The collection of all these pose estimates form the scene estimate.

B. Nonparametric Belief Propagation

Loopy belief propagation in the context of continuous variables is shown in Algorithm 1. Computing message updates in continuous domain is nontrivial. A message update in a continuous domain at an iteration n from a node $t \rightarrow s$ is:

$$m_{ts}^n(x_s) \leftarrow \int_{x_t\in\mathbb{R}^d} \Big(\psi_{st}(x_s, x_t)\phi_t(x_t, y_t) \prod_{u\in\rho(t)\setminus s} m_{ut}^{n-1}(x_t) \Big) dx_t \quad (2)$$

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

$$bel_s^n(x_s) \propto \phi_s(x_s, y_s) \prod_{t\in\rho(s)} m_{ts}^n(x_s) \quad (3)$$

We approximate each message $m_{ts}^n(x_s)$ as a mixture of weighted Gaussian components given as:

$$m_{ts}^n(x_s) = \sum_{i=1}^{M} w_s(i)\mathcal{N}(x_s; \mu_s(i), \Lambda_s(i)) \quad (4)$$

where M is the number of Gaussian components, $w_s(i)$ is the weight associated with the i^{th} component, $\mu_s(i)$ and $\Lambda_s(i)$ are the mean and covariance of the i^{th} component respectively. Fixing $\Lambda_s(i)$ to a constant Λ simplifies the approximation without affecting the performance of the system [16].

NBP provides a Gibbs sampling based method to compute an approximate of the value $\prod_{u\in\rho(t)\setminus s} m_{ut}^{n-1}(x_t)$ that results in the same form as Eq 4. Assuming that $\phi_t(x_t, y_t)$ is pointwise computable, the product $\phi_t(x_t, y_t)\prod_{u\in\rho(t)\setminus s} m_{ut}^{n-1}(x_t)$ is computed as part of the sampling procedure. The pairwise term $\psi_{st}(x_s, x_t)$ should be approximated as marginal influence function $\zeta(x_t)$ to make the right side of Eq 2 independent of x_s. The marginal influence function is given by:

$$\zeta(x_t) = \int_{x_s\in\mathbb{R}^d} \psi_{st}(x_s, x_t)dx_s \tag{5}$$

If the marginal influence function is also pointwise computable then the entire product $\zeta(x_t)\phi_t(x_t, y_t)\prod_{u\in\rho(t)\setminus s} m_{ut}^{n-1}(x_t)$ can be computed as part of the sampling procedure proposed in NBP. Refer to the papers describing NBP [16] and PAMPAS [8] for further details on how changes to the nature of the potentials affect the message update computation (as in Eq 2). The marginal influence function provides the influence of x_s for sampling x_t. However, the function can be ignored if the pairwise potential function is based on the distance between the variables, which is true in our case.

C. 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 observation y_t. The observation $y_t \in (l_t, B_t, C_t)$ provides the 2D bounding boxes B_t corresponding to the x_t. Let b_t be a bounding box sampled from the B_t using the C_t as corresponding weights. Let p_t be a subset of original point cloud corresponding to a bounding box b_t. The hypothesized object pose x_t is used to position the given geometric object model for object l_t 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 geometric model available a priori. The likelihood is calculated as

$$\phi_t(x_t, y_t) = e^{\lambda_r D(p_t, p_t^*)} \tag{6}$$

where λ_r is the scaling factor, $D(p_t, p_t^*)$ is the sum of 3D Euclidean distance between each point in p_t and p_t^* associated by their pixel location in the observed bounding box b_t on I.

2) Pairwise potential: Pairwise potential gives information about how compatible two object poses are given the support relation in the form of edges. We consider three different support cases: 1) object s is not directly in physical contact with object t, 2) object s is supporting object t and 3) object t is supporting object s. This support structure is provided as input to the system as shown in Fig 2. A binary value $\{0, 1\}$ is assigned based on the compatibility of object poses and their support relations. Object geometries are available to the system for modeling this potential. We consider the number of the dimensions of a pose to be 3 so that $x_s = \{x_s^x, x_s^y, x_s^z\}$ and $x_t = \{x_t^x, x_t^y, x_t^z\}$. The dimensions of the object models are denoted as $d_s = \{d_s^x, d_s^y, d_s^z\}$ and $d_t = \{d_t^x, d_t^y, d_t^z\}$ for the objects associated with the nodes s and t respectively. The potentials between the nodes s and t in the graph G can be computed using simple rules as shown in the case below.

Case 1: Object s is not in physical contact with t

$$\psi_{ts}(x_s, x_t) = \begin{cases} 1, & \text{if } \Delta x > \frac{(d_s^x + d_t^x)}{2} \\ & \text{or } \Delta y > \frac{(d_s^y + d_t^y)}{2} \\ & \text{or } \Delta z > \frac{(d_s^z + d_t^z)}{2} \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where $\Delta x = |x_s^x - x_t^x|$ and d_s^x denotes the size of the geometry associated with the node s in x direction. The size of the geometry is used to avoid collisions and hence an approximation such as $d_s^x = d_s^y = \max(d_s^x, d_s^y)$ can be used.

Case 2: Object s supports object t

$$\psi_{ts}(x_s, x_t) = \begin{cases} 1, & \text{if } \Delta x < \frac{1}{2}(d_s^x) \\ & \text{and } \Delta y < \frac{1}{2}(d_s^y) \\ & \text{and } |\Delta z - \frac{1}{2}(d_s^z + d_t^z)| < \Delta d \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

where Δd denotes such a threshold in the z direction that the objects are touching each other.

Case 3: An object associated with t is supporting an object associated with s

$$\psi_{ts}(x_s, x_t) = \begin{cases} 1, & \text{if } \Delta x < \frac{1}{2}(d_t^x) \\ & \text{and } \Delta y < \frac{1}{2}(d_t^y) \\ & \text{and } |\Delta z - \frac{1}{2}(d_s^z + d_t^z)| < \Delta d \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

However, in the sampling based algorithm it is computationally optimal to sample for cases 2 and 3. In case 2, given x_t, the x_s^x and x_s^y components of x_s can be sampled using $\mathcal{N}(x_t^x, \frac{1}{2}(d_t^x))$ and $\mathcal{N}(x_t^y, \frac{1}{2}(d_t^y))$ respectively. The x_s^z is computed as $x_t^z + \frac{d_s^z + d_t^z}{2} + \mathcal{N}(0, \sigma)$ with a small noise. Similar to case 2, in case 3, the x_s^x and x_s^y components of x_s can be sampled using $\mathcal{N}(x_t^x, \frac{1}{2}(d_t^x + d_s^x))$ and $\mathcal{N}(x_t^y, \frac{1}{2}(d_t^y + d_s^y))$. The x_s^z is computed as $x_t^z - \frac{d_s^z + d_t^z}{2} + \mathcal{N}(0, \sigma)$ with a small noise.

The Algorithm 1 denotes the high level overview of a non-parametric belief propagation, Algorithm 2 gives the details of how the Message update is performed using Gibbs sampling approach. Algorithm 3 gives the details of how the belief is computed which results in the maximum likelihood estimate.

Algorithm 1: Overall Belief Propagation

Given node potentials $\phi(x_s, y_s)\ \forall s \in V$, pairwise potentials $\psi(x_s, x_t)\ \forall (s,t) \in E$ and initial messages for every edge $m^0_{st}\ \forall (s,t) \in E$, the algorithm iteratively updates all messages and computes the belief till the graph G till converges.

1 For $n \in [1$:MAX BELIEF ITERATIONS]

(a) **Message update:**
Update messages from iteration $(n-1) \rightarrow n$ using Eq 2 to 5
$m^n_{ts}(x_s) \leftarrow$
$\int_{x_t \in \mathbb{R}^d} \left(\zeta(x_t)\phi_t(x_t, y_t) \prod_{u \in \rho(t) \setminus s} m^{n-1}_{ut}(x_t) \right) dx_t$

(b) **Belief update:**
(Optional) Update belief at every iteration if necessary. However, it doesn't affect the message update at 1(a) in next iteration.
$bel^n_s(x_s) \propto \phi_s(x_s) \prod_{t \in \rho(s)} m^n_{ts}(x_s)$

Algorithm 2: Message update

Given input messages $m^{n-1}_{ut}(x_t) = \{\mu^{(i)}_{ut}, \Lambda^{(i)}_{ut}, w^{(i)}_{ut}\}^M_{i=1}$ for each $u \in \rho(t) \setminus s$, and methods to compute functions $\psi_{ts}(x_t, x_s)$ and $\phi_t(x_t, y_t)$ pointwise, the algorithm computes $m^n_{ts}(x_s) = \{\mu^{(i)}_{ts}, \Lambda^{(i)}_{ts}, w^{(i)}_{ts}\}^M_{i=1}$

1. Draw M independent samples $\{\hat{x}^{(i)}_t\}^M_{i=1}$ from the product $\zeta(x_t)\phi_t(x_t, y_t) \prod_{u \in \rho(t) \setminus s} m^n_{ut}(x_t)$ using Gibbs sampler in NBP.
   (a) In our specific case $\zeta(x_t)$ is 1.0.
2 For each $\{\hat{x}^{(i)}_t\}^M_{i=1}$, sample $\hat{x}^{(i)}_s \sim \psi_{st}(x_s, x_t = \hat{x}^{(i)}_t)$
   (a) Using Eq:(7, 8, 9), rejection sampling is performed to sample $\hat{x}^{(i)}_s$.
3 $m^n_{ts}(x_s) = \{\mu^{(i)}_{ts}, \Lambda^{(i)}_{ts}, w^{(i)}_{ts}\}^M_{i=1}$ is constructed
   (a) $\mu^{(i)}_{ts}$ is the sampled component $\hat{x}^{(i)}_s$
   (b) kernel density estimators can be used to select the appropriate kernel size $\Lambda^{(i)}_{ts}$. We use "rule of thumb" estimator [9].
   (c) $w^{(i)}_{ts}$ is initialized to $1/M$; however, if the pairwise potential is a density function then importance weights in the selection of $\mu^{(i)}_{ts}$ can be used.

Algorithm 3: Belief update

Given incoming messages $m^n_{ts}(x_t) = \{\mu^{(i)}_{ts}, \Lambda^{(i)}_{ts}, w^{(i)}_{ts}\}^M_{i=1}$ for each $t \in \rho(s)$, and methods to compute functions $\phi_s(x_s, y_s)$ pointwise, the algorithm computes $bel^n_s(x_s) \propto \phi_s(x_s, y_s) \prod_{t \in \rho(s)} m^n_{ts}(x_s)$

1 Draw T independent samples $\{\hat{x}^{(i)}_s\}^M_{i=1}$ from the product $\phi_s(x_s, y_s) \prod_{t \in \rho(s)} m^n_{ts}(x_s)$ using Gibbs sampler in NBP.
   (a) If $\phi_s(x_s, y_s)$ is modeled as a mixture of Gaussians, then it can be part of the product.
   (b) If $\phi_s(x_s, y_s)$ can be computed pointwise, then the variant proposed in NBP [16] can be used to compute the product using Gibbs sampling.
2 $bel^n_s(x_s) \sim \{\mu^{(i)}_s, \Lambda^{(i)}_s, w^{(i)}_s\}^T_{i=1}$ is used to compute the scene estimate.
3 (Optional) To perform robot grasping and manipulation, a single estimate is required. We draw K samples from the belief product in step (2) to initialize the ICP algorithm. From the ICP fits, get the weighted mean and covariance. These are the estimates for manipulation experiments.

IV. Experiments

In this section we firstly discuss the results of the proposed belief propagation system qualitatively. Secondly, we introduce the measure of clutterness of scene in terms of visibility of objects. Lastly, we discuss a goal-directed manipulation experiment performed with the proposed belief propagation framework.

A. Belief Propagation Experiments

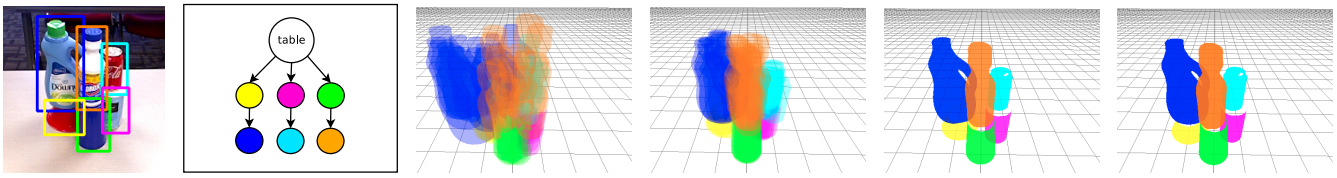
We initially tested the proposed framework on 6 scenes with 6 objects in different configurations for each scene. The scenes included objects in stacked, partially occluded, and completely occluded configurations. In Fig 3, we show the qualitative results of the pose estimations for a mildly cluttered scene. The belief samples after 15 iterations shown in Fig 3d cannot be directly used for manipulation as they represent the samples drawn from a distribution of poses. However, when these samples go through a post processing step with ICP they transform into reliable a single pose estimates for each object that fits to the point cloud. ICP fits for each belief sample are used to compute the weighted mean, which acts as the single final estimate for manipulation. For every sample output from the ICP fit, distance is computed to all the other samples in output of ICP. The distances are normalized to sum to one. Weights used in the weighted mean is the $1-$distance computed and assigned to the sample. The weighted variance measures the uncertainty in perception for manipulation. If this variance is higher than a threshold $(0.25cm^2)$ on all dimensions, then no manipulation action is performed with this final pose estimate. For this mildly cluttered scenes with 6 objects, the average error in the pose of the final scene estimate is 0.91cm.

B. Analysis of Clutterness

We analyze the visibility of the objects in the scene and measure its clutterness. This analysis is done using the ground truth poses of the objects. If a scene consists of N objects, the entire scene is rendered to a depth map S using OpenGL rendering. Each of the N objects are rendered at their ground truth poses in isolation to generate N depth maps $O_{1:N}$. Rendered image has non-zero depth value at pixels where objects are present. Let J_i be the set of all the pixels in the rendered depth map with non-zero depth value for object O_i. Then the visibility V^o_i of object O_i and scene V^s is given by the below equation.

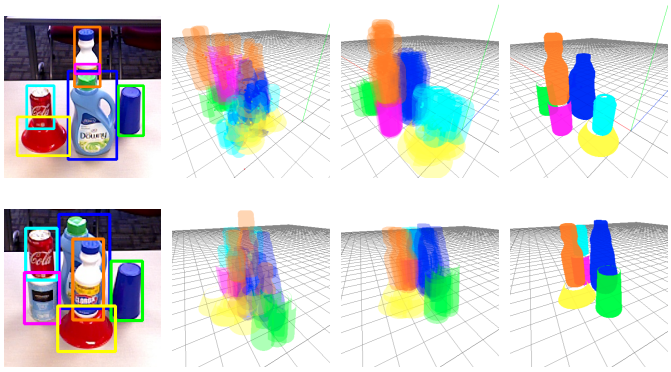
$$V^o_i = \frac{v^o_i}{K^o_i}$$
$$V^s = \frac{\sum_{i=1}^N v^o_i}{\sum_{i=1}^N K^o_i} \tag{10}$$

where v^o_i is the total number of pixels with same depth values in the isolated depth map of object O_i and full scene S; K^o_i is the total number of pixels in J_i. See Fig. 5 to get a sense of the scene visibility and object visibility.



(a) RGBD Observation with detections (b) Scene graph prior (c) Initial belief samples (d) Final belief samples (e) Final estimate (f) Ground truth

Fig. 3: Qualitative results showing the estimation of the proposed framework: This experiment uses 15 Gaussian components for each message and is run for 15 iterations to get the final belief samples. Final estimation is achieved through post processing the final belief samples with ICP registration, followed by the weighted average on the converged poses.



(a) Observation Initialized belief Belief after 5 iterations Ground truth from different view

Fig. 4: The first row shows a complete occlusion case: pink object (color of bounding box in first column or object colors in 2-4 columns) is behind the blue object and supporting the orange object. The bounding box for the pink object drives the initial sampling. In the later iterations, samples are drawn using support relations. The second row demonstrates a partial occlusion case: The blue object (downy detergent) is behind the orange object in the front. The bounding box initializes the belief propagation confusingly. However, the pairwise potential resolves these cases by sampling away from the objects without support relations.

C. Goal-directed Manipulation Experiment

To evaluate the proposed framework's success in robotic manipulation tasks we use the weighted mean pose and the weighted variance that results from the post processing step to perform a scene reorganization task. The reorganization task is provided a goal scene with desired place locations for the objects. Specifically, the robot's task is to unload a tray full of objects on to the table at the desired place locations. To accomplish this task, a sequence of pick-and-place actions are executed by the robot using the estimated poses of each object. After all estimated object poses with high confidence (with a threshold variance of $0.25cm^2$) have been acted on, we reapply the scene estimation to produce updated pose estimates. We iterate until the entire task is accomplished.

In Fig 6, we report the results of an object manipulation experiment on a heavily cluttered scene. This scene contains 10 objects stacked on a tray with a mixture of visible and non-visible objects with partial and complete occlusions. This experiment ended up with 4 sequence of perception and manipulations stages. Sequence 1: begins with robot perceiving the scene, which provides accurate pose estimates for the coffee (yellow) and ginger tea (cyan) along with high confidence in terms of variance less than $0.25cm^2$. These two estimates are used by the robot manipulate these objects from the tray onto the table. It should be noted from the Fig. 5(a) that these two objects were 100% visible. However the perception system was not confident about the coffee mate which had 86% visibility. The perception system didn't consider any other leaf node for the manipulation action as their belief samples didn't converge due to low visibility and had high uncertainty. Sequence 2: robot perceives the scene again, which provided accurate poses with high confidence for not only the leaf nodes -lemon tea (red) and granula bar (pink)) but also the parent node for the lemon tea which is oats container (purple). It should be noted from Fig. 5(b) that only the fully visible (100%) leaf nodes are manipulated along with 62% visible non-leaf node. This is because the planner's sequence of actions contains the entire manipulation of the scene graph at any point of the time. Sequence 3: the robot perceives the scene again. At this stage all the objects on the tray are the leaf nodes except for the protein box which is visible to the sensor for the first time with visibility of 5%. Only three objects gets accurate pose estimates along with high confidence for manipulation. Sequence 4: we perceive the scene that contains only two objects whose visibility is 100%. The poses are estimated with high confidence for manipulation.

In addition to the manipulation results, we would like to emphasize the advantage of using scene graph in the inference. The first row of Fig 4 shows a case, where the location of a completely occluded object supporting a visible object is correctly estimated. Similarly second row of Fig 4 demonstrates how the scene graph enables a correct estimation for the location of an object partially occluded by another one.

V. Conclusion

We analyze the approach towards goal-directed manipulation especially in cluttered scenes. We observe that the visibility can affect perception system. To overcome the planner using the pose estimates directly from the perception system, a confidence measure is calculated. We proposed a nonparametric scene graph belief propagation to estimate a scene as a collection of object poses and their interactions.

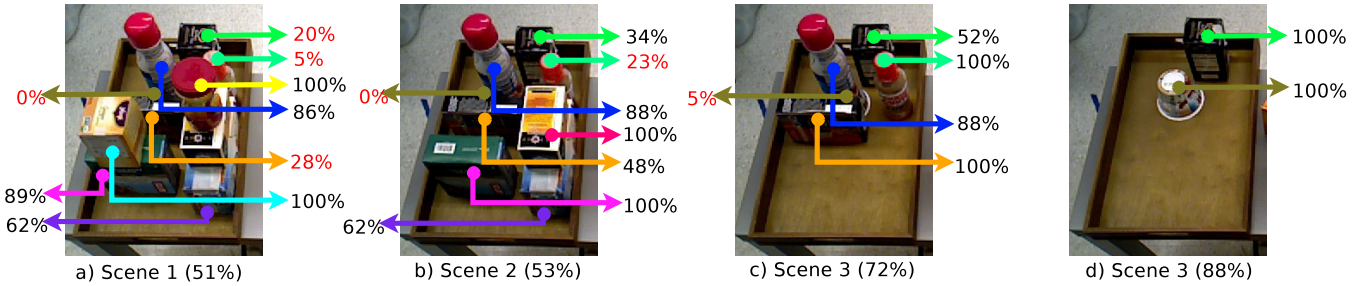


Fig. 5: Visibility of the nodes in the scene graph as viewed by the robot. Arrows towards left side of the scene corresponds to non-leaf nodes and right side corresponds to the leaf nodes. These scenes are same as in Fig. 6 with same color coding as in scene graphs. The visibility is computed using the measure described in Eq. 10. The red color % correspond to values less than 30%. The scene complexity is provided in the caption of every scene.

This problem is formulated as a graph inference problem on a Markov Random Field. We show the benefit of the belief propagation approach in manipulation by presenting the qualitative results of pose estimations used for robot manipulation. We also show how to measure the clutterness of the scene in terms of object visibilities. In the future work we would like to infer the graph structure or the scene graph which is currently provided as the input to the system.

REFERENCES

[1] C. Chao, M. Cakmak, and A. L. Thomaz. Towards grounding concepts for transfer in goal learning from demonstration. In *Proceedings of the Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics (ICDL-EpiRob)*. IEEE, 2011.

[2] J. Chua and P. F. Felzenszwalb. Scene grammars, factor graphs, and belief propagation. *arXiv preprint arXiv:1606.01307*, 2016.

[3] M. Ciocarlie, K. Hsiao, E. G. Jones, S. Chitta, R. B. Rusu, and I. A. Şucan. Towards reliable grasping and manipulation in household environments. In *Experimental Robotics*, pages 241–252. Springer Berlin Heidelberg, 2014.

[4] A. Collet, M. Martinez, and S. S. Srinivasa. The moped framework: Object recognition and pose estimation for manipulation. *Int. J. Rob. Res.*, 30(10):1284–1306, Sept. 2011.

[5] M. Dogar, K. Hsiao, M. Ciocarlie, and S. Srinivasa. Physics-based grasp planning through clutter. 2012.

[6] R. E. Fikes and N. J. Nilsson. Strips: A new approach to the application of theorem proving to problem solving. *Artificial intelligence*, 2(3):189–208, 1972.

[7] 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.

[8] M. Isard. Pampas: Real-valued graphical models for computer vision. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2003.

[9] H. Läuter. Silverman, bw: Density estimation for statistics and data analysis. chapman & hall, london–new york 1986, 175 pp.,£ 12., 1988.

[10] 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. Morgan Kaufmann Publishers Inc., 1999.

[11] V. Narayanan and M. Likhachev. Discriminatively-guided deliberative perception for pose estimation of multiple 3d object instances. In *Robotics: Science and Systems*, 2016.

[12] S. Narayanaswamy, A. Barbu, and J. M. Siskind. A visual language model for estimating object pose and structure in a generative visual domain. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 4854–4860. IEEE, 2011.

[13] C. Papazov, S. Haddadin, S. Parusel, K. Krieger, and D. Burschka. Rigid 3d geometry matching for grasping of known objects in cluttered scenes. *The International Journal of Robotics Research*, page 0278364911436019, 2012.

[14] 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.

[15] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard. Tracking loose-limbed people. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2004.

[16] E. B. Sudderth, A. T. Ihler, W. T. Freeman, and A. S. Willsky. Nonparametric belief propagation. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2003.

[17] E. B. Sudderth, M. I. Mandel, W. T. Freeman, and A. S. Willsky. Visual hand tracking using nonparametric belief propagation. In *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on*, pages 189–189. IEEE, 2004.

[18] 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.

[19] A. ten Pas and R. Platt. Localizing handle-like grasp affordances in 3d point clouds. In *International Symposium on Experimental Robotics (ISER)*, 2014.

[20] T. Winograd. Understanding natural language. *Cognitive psychology*, 3(1):1–191, 1972.

[21] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei. Scene graph generation by iterative message passing. *arXiv preprint arXiv:1701.02426*, 2017.

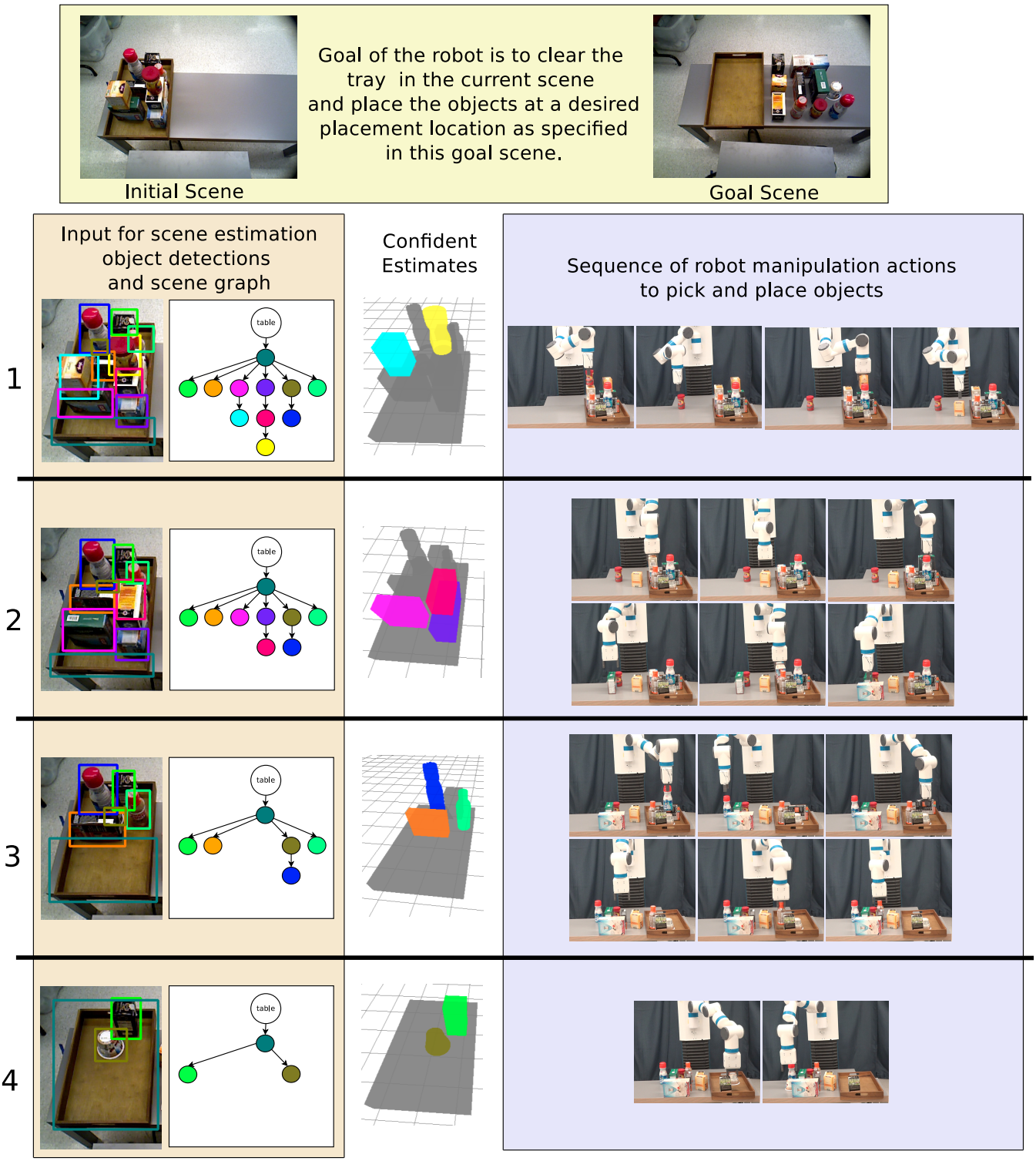


Fig. 6: Goal-directed manipulation experiment: The robot is given a task of reconfiguring the scene and achieve a predefined goal configuration. In this experiment, the robot takes 4 sequences to finish the task. Each sequence starts with the proposed system perceiving the scene and providing object pose estimates and measure of uncertainty. The robot performs the pick-and-place actions on the objects whose estimates have high confidence from the perception system. In the last run the robot fails to pickup an object as it was not reachable, though the estimate was precise with respect to the ground truth.