What Is Where: Inferring Containment Relations from Videos

Wei Liang\textsuperscript{1,2}, Yibiao Zhao\textsuperscript{2}, Yixin Zhu\textsuperscript{2}, Song-Chun Zhu\textsuperscript{2}
liangwei@bit.edu.cn, ybzhao@ucla.edu, yixin.zhu@ucla.edu, sczhu@stat.ucla.edu
\textsuperscript{1}School of Computer Science, Beijing Institute of Technology (BIT), China
\textsuperscript{2}Center for Vision, Cognition, Learning, & Autonomy, University of California, Los Angeles, USA

Abstract

In this paper, we present a probabilistic approach to explicitly infer containment relations between objects in 3D scenes. Given an input RGB-D video, our algorithm quantizes the perceptual space of a 3D scene by reasoning about containment relations over time. At each frame, we represent the containment relations in space by a containment graph, where each vertex represents an object and each edge represents a containment relation. We assume that human actions are the only cause that leads to containment relation changes over time, and classify human actions into four types of events: move-in, move-out, no-change and paranormal-change. Here, paranormal-change refers to the events that are physically infeasible, and thus are ruled out through reasoning. A dynamic programming algorithm is adopted to finding both the optimal sequence of containment relations across the video, and the containment relation changes between adjacent frames. We evaluate the proposed method on our dataset with \textit{1326} video clips taken in \textit{9} indoor scenes, including some challenging cases, such as heavy occlusions and diverse changes of containment relations. The experimental results demonstrate good performance on the dataset.

1 Introduction and Motivations

For many AI tasks, such as scene understanding in visual perception, task planning in robot autonomy, and symbol grounding in natural language understanding, a key problem is to infer “what is where over time”. A person may say “the pizza is in a pizza box, and the pizza box is in a fridge”. In such a description, the object locations are described in a qualitative and hierarchical way, in which 	extit{containers} play an important role in quantizing human perceptual space via 	extit{containment relations}. By 	extit{containers}, we refer to any general objects in a scene that can contain other objects, for example, fridge, mug, box, lunch bag, envelop and so on. The 	extit{containment relations} between containers and contained objects, i.e. 	extit{containees}, may change over time by agents.

In this paper, we propose a probabilistic approach to infer containment relations from RGB-D videos. Consider the example shown in Fig. 1. The containers and containees are tracked in a 3D scene and highlighted in colored bounding boxes in the top panel. The inferred containment relations are constructed in the bottom panel, pointing from containees to the corresponding containers, and the numbers on edges denote the frames when the containment relations occur. It is worth noting that the containment relations are time varying, and can be changed by human actions. The presented containment relation inference method is aim to address the following two tasks.

Recovering hidden objects with severe occlusions. Severe occlusions frequently happen in daily scene. Reasoning about containment relations helps to recover objects from tracking failures when objects are partially occluded or even completely unobservable. For instance, as shown in Fig. 1, although we only observe a person taking a pizza from a pizza...
1. We propose a probabilistic approach to infer containment graphs (Fig. 2(c)).

2. The containment relations are represented by a containment graph; in time, containment relation changes are proposed based on human actions. To find the optimal interpretation across the full sequence, a dynamic programming algorithm is adopted to globally optimize both spatial and temporal space, resulting in the optimal sequence of containment graphs. The containment graph sequence defines both spatial containment relations at each frame (blue edges in the graph) and temporal containment relation changes over time (changes of the blue edges, highlighted by red dashed arrows) caused by human actions.

Inferring subtle human actions. Because there are self-occlusions or occlusions by other objects in a scene, subtle human actions that involve small and local movements, such as placing a phone in a bag, are difficult to detect. If a change of objects’ status (i.e. a containment relation change) is observable, it is natural to reason about that some human actions occurred. The ability of inferring human subtle actions by goals instead of observing and matching detailed action trajectories provides possibilities for a robot to understand the intentions of agents.

Fig. 2 illustrates the framework of the proposed method. Given a RGB-D video captured by a consumer depth camera (Fig. 2(a)), our method first tracks the objects of interest and the human skeletons (Fig. 2(b)). Then at each frame $t$, the containment relations are represented by a containment graph; in time, containment relation changes are proposed based on human actions. To find the optimal interpretation across the full sequence, a dynamic programming algorithm is adopted to globally optimize both spatial and temporal space, resulting in the optimal sequence of containment graphs (Fig. 2(c)).

This paper makes three major contributions:

1. We propose a probabilistic approach to infer containment relations from videos over time. A dynamic programming algorithm is applied to solve the ambiguities on both containment relations in space and containment relation changes in time, providing a globally optimized solution.

2. We propose a dynamic graph representation for containment relations over time (shown in Fig. 1). The dynamic graph quantizes 3D scene space and provides a qualitative way for tracking objects with heavy occlusions.

3. We model the containment relation changes in time by assuming that human actions are the only cause to change the containment relations. This constraint in return helps to recover hidden objects and infer time varying containment relations at each frame.

1.1 Related Work

Cognitive studies [Hespos and Spelke, 2007] have shown that infants can understand containers and containment relations as early as 3.5 months old. Strickland and Scholl [Strickland and Scholl, 2015] suggest that infants can detect containment before understanding occlusion. Liang et al. [Liang et al., 2015] evaluates human cognition of containing relations for adults through a series of experiments using physical-based simulations. In computer science, the problem of containers has been studied from various perspectives in the fields of AI, robotics and computer vision.

AI. Qualitative Spatial Representation / Reasoning (QSR) has been extensively studied in the AI community. Cohn and Hazarika [Cohn and Hazarika, 2001] provided a survey of key ideas and results in QSR literature. Some typical methods include using ontology [Hudelot et al., 2008; Grenon and Smith, 2004], topology [Gerevini and Renz, 2002; Li, 2007], metric spatial representation [Frank, 1992; Papadias and Sellis, 1994], and other approaches [Hedau et al., 2010; Sokeh et al., 2013; Renz, 2012]. Since 1980s, the AI community began to study containers as a typical example for qualitative reasoning using symbolic input [Bredeweg and Forbus, 2003; Frank, 1996]. In particular, Collins and Forbus [Collins and Forbus, 1987] used containers to reason about liquid by introducing a new technique, namely molecular collection ontology. A knowledge base for qualitative reasoning about containers was developed by Davis et al. [Davis
Figure 3: Temporal assumption. An apple is taken out of a bowl by a person, resulting a containment relation change. (a) Tracked skeletons during the time interval \([t, t + m]\). (b) At time \(t\), the red bowl was the container of the apple. (c) At time \(t + m\), the person became the container of the apple.

et al., 2013], which was expressed in a first-order language of time, geometry, objects, histories, and events. However, existing methods for spatial and temporal reasoning in the AI literature are mostly based on logic formulas, which are difficult to apply on processing real sensory inputs. The ability to handle noisy visual signal as inputs by introducing probability to model qualitative spatial relations makes our method different from previous work.

Robotics. Localizing objects using spatial relations has received considerable interests in the robotics community, e.g. [Aydemir et al., 2011; Wong et al., 2013; Feddema et al., 1997; Aksoy et al., 2010; Nitti et al., 2014]. Alper et. al [Aydemir et al., 2011] utilized spatial relations to describe topological relationships between objects. Wong et al. [Wong et al., 2013] studied occluded objects inside containers, and presented a novel generative model for representing container contents by using object co-occurrence information and spatial constraints. Feddema et al. [Feddema et al., 1997] applied two methods to control the surface of liquid in an open container which was moved by a robot. Most of the existing methods can only reason about containment relations in a known structured environment. In contrast, the proposed method aims to address the problem in arbitrary environments, where the number of objects are not fixed and relation changes occur more frequently.

Computer Vision. Two streams of studies are closely related to the present work: object affordance and tracking objects using context information. In recent literature, there is growing interest in understanding scenes and objects by their affordances and functionalities [Grabner et al., 2011; Gupta et al., 2011; Zhao and Zhu, 2013; Zhu et al., 2015; 2016] and their possible interactions with human poses [Satkin et al., 2012; Koppula et al., 2013; Zhu et al., 2014; Wang et al., 2014a; Wei et al., 2013; Pei et al., 2013]. Using context information has also been extensively explored in human-object interaction and multi-object tracking. For instance, Yang et al. [Yang et al., 2009] proposed to track multiple interacting objects by mining auxiliary objects, and [Wang et al., 2014b] formulated the interacting objects tracking as a network-flow Mixed Integer Program problem. More recently, a multiple objects tracking algorithm [Yoon et al., 2015] was proposed by maintaining spatial relations between objects using a Relative Motion Network. In comparison, our method utilizes the interactions between human and objects as temporal constraints to infer the explicitly modeled containment relations and their changes, resulting a probabilistic approach to recover objects with heavy occlusions.

2 Problem Definition
We use \(\Omega = \{O^i | i = 1 \cdots N\}\) to denote all the objects of interest in a scene, where \(O^i\) denotes the \(i\)th object. At each frame, we define the following variables:

- A containment indicator function is denoted by \(C_t(\cdot) \in \Omega\).
- A containment relation \(R^i_t = (O^i, C_t(O^i))\) is an ordered pair representing the containment relation between \(O^i\) and \(C_t(O^i)\). The set of all containment relations at time \(t\) is denoted as \(A_t = \{R^i_t | i = 1, \cdots, N\}\).
- A containment graph is denoted as \(G_t = (\Omega, A_t)\), where \(\Omega\) is the set of vertices and \(A_t\) is the set of directed edges.

To make the inference process tractable, we make the following assumptions about the properties of \(G_t\).

Spatial Assumptions. i) Each object must be contained by one and only one container, except the root node of \(G_t\), i.e., the “room”, which does not have its container. ii) There is no loop in \(G_t\), that is, object cannot contain itself. The nested containment relations do not form loops. iii) A person becomes a container when he or she holds an object.

Temporal Assumption. We assume that all containment relation changes ought to be caused by a person, which means a scene does not have external disturbance other than human. In other words, if we know there is no human actions, the containment relations should not change. This temporal assumption couples containment relations in space with containment relation changes over time. Fig. 3 gives an example: a person takes an apple out of a bowl, during which the person breaks the containment relation between the apple and the bowl, and establishes a new containment relation between the apple and the person.

3 Problem Formulation
The objective of our work is to interpret the observed video as the optimal sequence of containment graphs \(\{G_t\}^*\) from a given RGB-D video \(\mathcal{V}_{[1,T]} = \{V_t | t = 1, \cdots, T\}\).

3.1 Containment Relations in 3D Space
At each frame \(t\), a container \(O^i\) is contained by a container \(C_t(O^i)\) if and only if it satisfies all following relations defined in terms of an energy function \(\Phi\) with the three components:

- IN relation defined by the energy term \(\phi^{IN}\);
- ON relation defined by the energy term \(\phi^{ON}\);
- AFFORD relation defined by the energy term \(\phi^{AFF}\).

IN relation describes containment relations from the top view:

\[
\phi^{IN}(R^i_t, V_t) = \Gamma(O^i)/[\Gamma(O^i) \cap \Gamma(C_t(O^i))],
\]

3420
where $\Gamma(O^i)$ is the projected area of containee $O^i$ along the gravity axis, and $\Gamma(O^i) \cap \Gamma(C_t(O^i))$ is the overlapped area between containee $O^i$ and its container $C_t(O^i)$ projected in 2D from the top view.

The bottom right of Fig. 4 shows three examples of IN relation. If a containee is contained by its container, the boundary of the containee should be inside the contour of the container from the top view.

**ON relation** describes containment relations from the front view:

$$\phi^{ON}(R^i_t, V_t) = D(Z_b(O^i), Z_g(C_t(O^i)))$$  \hspace{1cm} (2)

where $Z_b(O^i)$ is the bottom coordinates of the containee projected to 2D. $Z_g(C_t(O^i)) = [Z_t(C_t(O^i)), Z_b(C_t(O^i))]$ is the interval of the container’s top and bottom coordinates projected to 2D, and $D$ is a distance function which measures how well the bottom of the containee $Z_b(O^i)$ falls into the intervals between the top and the bottom of the container $Z_g(C_t(O^i))$. If a containee is contained by its container, the bottom of the containee has to contact the container, and the containee should be above the container. Three examples of ON relation are illustrated in the top right of Fig. 4.

**AFF relation** $\phi^{AFF}(R^i_t, V_t)$ measures the ability of a container to afford a containment relation with containee at frame $V_t$, which is a pair-wise term. The containment relation is subject to a set of physical and geometric constraints. For example, a porous basket can neither contain a containee bigger than itself, nor smaller containees like beads. In this paper, we only consider the relative volume between the container and the containee.

**Energy of containment relations** is defined as:

$$\phi(G_t, V_t) = \lambda_1 \cdot \phi^{IN} + \lambda_2 \cdot \phi^{ON} + \phi^{AFF}$$  \hspace{1cm} (3)

where $\lambda_1$ and $\lambda_2$ are the weights of the energy terms, obtained through cross-validation during the training phrase.

### 3.2 Containment Relation Changes in Time

The containment relation change between frame $t$ and $t+1$ is denoted as $\Delta R^i_t$. We classify the changes based on human actions into the following four categories, shown in Fig. 5.

**Move-in** is defined as $\Delta R^i_t : (O^i, H_t) \rightarrow (O^i, C_{t+1}(O^i))$, which describing a containee $O^i$ moves from a person $H_t$ at frame $t$ to another container $C_{t+1}(O^i)$ at frame $t+1$.

**Move-out** is the opposite change to move-in, defined as $\Delta R^i_t : (O^i, C_t(O^i)) \rightarrow (O^i, H_{t+1})$.

**No-change** implies there is no containment relation change between frame $t$ and $t+1$, defined as $C_t(O^i) = C_{t+1}(O^i)$.

**Paranormal-change** refers to the change that is in conflict with the temporal assumption that a person is the only cause that leads to containment relation changes, and thus is ruled out through reasoning. Formally, if the containment relation $R^i_t$ changes, i.e. $C_t(O^i) \neq C_{t+1}(O^i)$, but no person $H$ is involved, i.e. $C_t(O^i) \neq H_t \& C_{t+1}(O^i) \neq H_{t+1}$, such changes are defined as paranormal-change.

**Energy of containment relation changes** is defined as

$$\psi(G_t, G_{t+1}, V_{t-\epsilon, t+\epsilon+\epsilon}) = \omega_{L_j} \cdot \theta,$$  \hspace{1cm} (4)

where $\omega_{L_j}$ is the template parameter for four types of containment relation changes $L_j, j \in \{1, 2, 3, 4\}$, $[t-\epsilon, t+\epsilon]$ is the time interval of the sliding windows for temporal feature extractions, and $\theta$ is the extracted feature vector.

For $\theta$, we introduce objects context during feature extractions [Wei et al., 2013]. Three kinds of features are considered in a sliding window on the time axis: the human pose $F^h_m$, the relative movements between the human hands and the object $F^r_m$, and the movements of the object $F^n_m$. Suppose that the sliding window size is $2\epsilon$, the feature vector sequence at time $t$ is $F_m = (F^h_m, F^r_m, F^n_m)$, $m \in [t-1-\epsilon, t-1+\epsilon]$, $F^h_m$ is the relative distance of all skeletons to three base points (two shoulders and a spine point). $F^r_m$ is the distance between the human hands and the position of the object. $F^n_m$ is the object position changes during the time interval $2\epsilon$. A wavelet transform is then applied to $F_m$. The coefficients at the low frequency are kept as the interaction feature $\theta$. The window sizes and sliding steps are both multi-scale.

### 3.3 Joint Spatial-Temporal Energy

By combining both the energy of containment relations in space at each frame (Eq. 3), and the energy of containment
relation changes between adjacent frames (Eq. 4), the optimal containment graph \( \{G_t\}^* \) is defined as:
\[
\{G_t\}^* = \arg\min_{\{G_t\}} E(\{G_t\},\{V_t\})
\]
which is independent of the container of the containee. By assuming such property that the container of each containee is independent, we can optimize the solution for each object separately. Hence, Eq. 5 can be rewritten in terms of containment relations with respect to each object:
\[
\{R_{ti}\}^* = \arg\min_{\{R_{ti}\}|t=1,\ldots,T} \left[ \sum_{t=1}^{T} \mu \phi(R_{ti},V_t) + \sum_{t=1}^{T-1} \psi(R_{t-1,i},R_{ti},V_{[t-\epsilon,t+\epsilon]}) \right]
\]

By aggregating \( \{R_{ti}\}^* \) from each object, we can recover the full sequence of containment graphs \( \{G_t\}^* \) of the scene. A dynamic programming is adopted to find the optimal solution of Eq. 6 with the time complexity \( O(N^2 \cdot T) \).

5 Experiments
5.1 Dataset
We collected a RGB-D video dataset with diverse actions to evaluate the proposed method. Our dataset consists of 1326 video clips in 9 scenes captured by a Kinect sensor, in which 800 clips are used to train our model and the remaining clips are for testing. For each video clip, RGB and depth images, 3D human skeletons as well as point cloud information are used as the input of the proposed method. Our dataset is unique, compared with traditional ones in the following aspects: i) it focuses on containers and containment relations; ii) it includes partially and completely occluded objects, such as an apple in a bowl (partial occlusion) and a laptop in a backpack (complete occlusion); iii) it includes diverse containment relation changes in different scenarios, such as throwing, picking up, opening lid, zipping zipper and so on.

5.2 Detection of Containment Relation Changes
Consider the case shown in Fig. 6, where a person moves a containee from one container to another, during this process the containee changes directions, scales, and views. Severe occlusions by hands and other objects also occurred. We show the probabilities of three relation changes by detection: move-in, move-out and no-change between two objects highlighted with bounding boxes. At each frame, we compare the category which has the highest probability among three candidate categories with the ground truth (bars at the bottom). The detection of relation changes works well in most cases, but it fails in certain situations: i) when skeletons, the containee and the container are occluded at the same time (frame 54-70), the algorithm cannot distinguish the relation change of move-in or move-out from no-change; ii) some skeletons or the containee are occluded partly (frame 380-390), which causes difficulties in distinguishing move-in from move-out.

We quantitatively compare the results of containment relation changes between our method with two baseline methods, as shown in Fig. 7: (a) recognition by human actions only, and (b) recognition by both human actions and object context. Both of them are trained by multi-class SVM on the same training data. There are obvious confusions between move-in and move-out in (a). By introducing object context, the proposed method improve the accuracy as shown in (b). The proposed method achieves the best performance in (c). The reason is that our method is able to correct some temporal detection errors.
5.3 Inferring Containment Relations

Fig. 8 demonstrates the inference process. Firstly, we track all objects of interest by state-of-the-art RGB-D trackers [Song and Xiao, 2013]. The objects are bounded by boxes with different colors in Fig. 8 (a). Take the object in the green box as an example, we infer the containment relations it involves in.

Fig. 8 (c) shows the probability of containment relations between the object in the green box and its potential container, using the spatial cues only. Each row represents one possible container. The height of the line represents how likely this container contains the object in the green box. Due to severe occlusions, the spatial cues of objects are missing constantly. When the object is occluded, the probability for each possible container is evenly distributed.

Fig. 8 (d) shows the inference result matrix, in which the $n^{th}$ row represents the DP result of each frame given the first $n$ frames of the video. The grey area denotes the states that are not observed up to the $n^{th}$ frame. As more information is given, the DP algorithm gradually corrects results and gets closer to the ground truth. Take $\textcircled{1}$ for example, the inferred container from frame 53 to 90 does not change to the human hand until the 91st frame. The initial inferred container remains to be the table due to heavy occlusion. But as time goes, the temporal information accumulates and wins against the spatial cues. our method achieves good performance in comparison with the ground truth in Fig. 8 (e).

We also perform quantitative evaluations. For comparison, we transform the tracking results into containment relations as a baseline. Specially, we apply non-maximum suppression at each frame for all candidates, and the bounding box with the highest score is considered as the present position of the object. For each object $O_i$, its container is the object which is nearest to $O_i$, and satisfies both ON relation and IN relation.

Fig. 8 (d) shows the inference result matrix, in which the $n^{th}$ row represents the DP result of each frame given the first $n$ frames of the video. The grey area denotes the states that are not observed up to the $n^{th}$ frame. As more information is given, the DP algorithm gradually corrects results and gets closer to the ground truth. Take $\textcircled{1}$ for example, the inferred container from frame 53 to 90 does not change to the human hand until the 91st frame. The initial inferred container remains to be the table due to heavy occlusion. But as time goes, the temporal information accumulates and wins against the spatial cues. our method achieves good performance in comparison with the ground truth in Fig. 8 (e).

We also perform quantitative evaluations. For comparison, we transform the tracking results into containment relations as a baseline. Specially, we apply non-maximum suppression at each frame for all candidates, and the bounding box with the highest score is considered as the present position of the object. For each object $O_i$, its container is the object which is nearest to $O_i$, and satisfies both ON relation and IN relation.

We divide our dataset into three parts according to the visibility of objects: no occlusion, partial occlusion and complete occlusion. Partial occlusion is the situation that the containee or its container is observed partially, whereas complete occlusion means that the container or its container is occluded completely, such as a laptop is put into a backpack. We quantitatively evaluate the accuracy of containment relations on these situations. The results are shown in Table 1. In the completely occluded situation, both methods perform worse than in the other situations. The proposed method performs better by recovering some relations from complete occlusions. In the situation of no occlusion, there are some false positives because of the observation noise in the detection process. Our method is efficient to eliminate some false positives.

### Table 1: Accuracy of containment relation estimation in %.

<table>
<thead>
<tr>
<th>Method</th>
<th>no occlusion</th>
<th>partial occlusion</th>
<th>complete occlusion</th>
<th>overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.75</td>
<td>0.21</td>
<td>0.08</td>
<td>0.37</td>
</tr>
<tr>
<td>Ours</td>
<td>0.86</td>
<td>0.64</td>
<td>0.43</td>
<td>0.59</td>
</tr>
</tbody>
</table>

6 Conclusion

Containers and containment relations are ubiquitous in daily scenes and activities. They are useful not only to answer “what is where over time” for various AI tasks, but also for quantizing the perception of functional space, detecting and tracking hidden objects and heavily occluded objects, and reasoning about human subtle actions. The presented method achieves good performance in some challenging scenarios. However, it is still limited in the following aspects: i) IN and ON relation do not describe all containment relations, such as liquid or gas; ii) the objects with large deformation, such as plastic bags, are still difficult to solve.

Acknowledgment. The authors would like to thank the
support of a DARPA SIMPLEX project N66001-15-C-4035, a MURI project N00014-16-1-2007, and a NSF grant IIS1423305.

References


