What is Where: Inferring Containment Relations from Videos

Wei Liang[1,2] **(liangwei@bit.edu.cn), Yibiao Zhao**[2] **(ybzhao@ucla.edu),**
Yixin Zhu[2] **(yixin.zhu@ucla.edu), Song-Chun Zhu**[2]**(sczhu@stat.ucla.edu)**
[1]School of Computer Science, Beijing Institute of Technology (BIT), China
[2]Center for Vision, Cognition, Learning, and Autonomy, University of California, Los Angeles (UCLA), 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, paranomal-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 1326 video clips taken in 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.

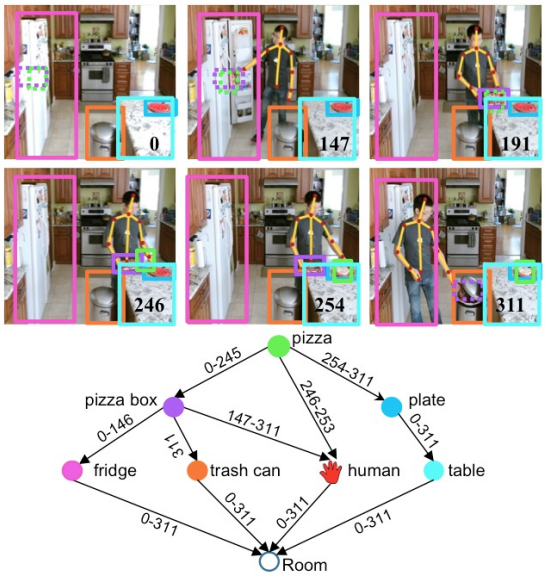

Figure 1: (Top) Structured, qualitative and abstract interpretation of containment relations over time in a scene. The goal is to answer "what is where over time". (Bottom) The inferred containment relations. The numbers on edges denote the frames when the containment relations occur.

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 *containers* play an important role in quantizing human perceptual space via *containment relations*. By *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 *containment relations* between containers and contained objects, i.e. *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

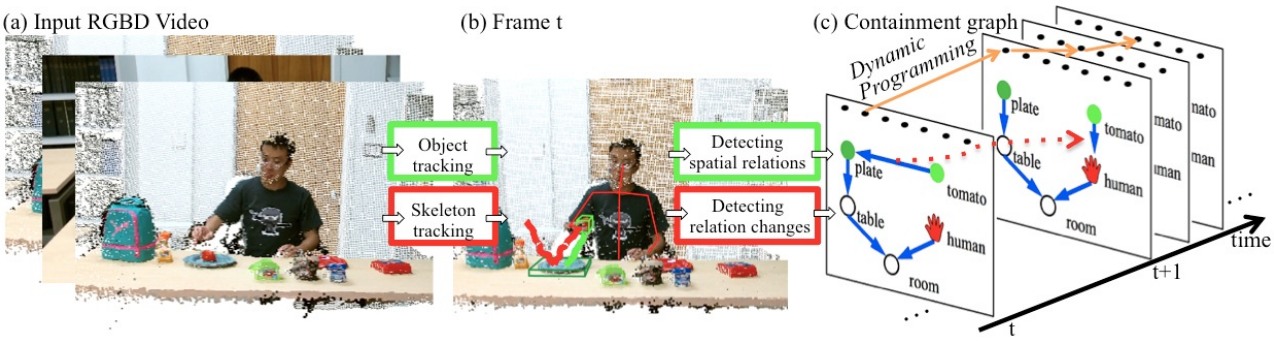

Figure 2: An overview of the proposed approach. (a) Given a RGB-D video, we first track objects and human skeletons in 3D space. (b) At each frame, the tracked 3D bounding boxes are used to construct containment relations, whereas tracked human skeletons are used to detect containment relation changes. (c) Across the video, a joint spatial-temporal inference method is used to find 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.

box at frame 246, we are able to infer that the pizza was contained by the pizza box from frame 1 to 245, during which period the pizza was unobservable. By simple commonsense that a containee shares the same position as its container when the containment relation between them holds, we are able to recover the hidden containee with severe occlusions or even without actually seeing it. This capability provides a potential solution to build a system (e.g. an assistive robot) to answer "what is where over time". For example, a robot can help to localize an object in a room if a person forgets where he or she left it.

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 (shwon 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] has 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 reasoningusing 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

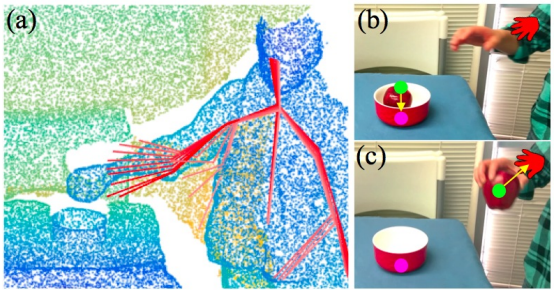

Figure 3: Temporal assumption. An apple is taken out of a bowl by a person, resulting a containment relation change. (a) Tracked skeletons during the time interval $[t, t+m]$. (b) At time t, the red bowl was the container of the apple. (c) At time $t+m$, the person became the container of the apple.

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 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 ith object. At each frame, we define the following variables:

- A containment indicator function is denoted by $C_t(\cdot) \in \Omega$. If O^j contains O^i directly, then $O^j = C_t(O^i)$, where O^i is the containee and O^j is the container.
- A containment relation $\mathcal{R}_t^i = \langle O^i, C_t(O^i) \rangle$ 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 $\Lambda_t = \{\mathcal{R}_t^i | i = 1, \cdots, N\}$.
- A containment graph is denoted as $\mathcal{G}_t = (\Omega, \Lambda_t)$, where Ω is the set of vertices and Λ_t is the set of directed edges.

To make the inference process tractable, we make the following assumptions about the properties of $\mathcal{G}_t$.

Spatial Assumptions. i) Each object must be contained by one and only one container, except the root node of $\mathcal{G}_t$, i.e., the "room", which does not have its container. ii) There is no loop in $\mathcal{G}_t$, 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 $\{\mathcal{G}_t\}^*$ 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 containee 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 Φ with the three components:

- IN relation defined by the energy term ϕ^{IN};
- ON relation defined by the energy term ϕ^{ON};
- AFFORD relation defined by the energy term ϕ^{AFF}.

IN relation describes containment relations from the top view:

$$\phi^{\text{IN}}(\mathcal{R}_t^i, V_t) = \Gamma(O^i) / [\Gamma(O^i) \cap \Gamma(C_t(O^i))], \qquad (1)$$

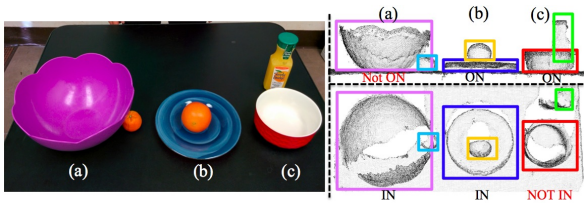

Figure 4: Containment relations in 3D space. (Left) A RGB image of a desktop scene. (Bottom right) Depth images from the top view. (Top right) Depth images from the front view. (a) and (c) violate ON relation and IN relation, respectively. Only (b) is considered to satisfy both IN and ON relations.

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:

$$\phi^{\text{ON}}(\mathcal{R}_t^i, V_t) = D(Z_b(O^i), Z_g(C_t(O^i))), \tag{2}$$

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 $\phi^{\text{AFF}}(\mathcal{R}_t^i, V_t)$ 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:

$$\phi(\mathcal{G}_t, V_t) = \lambda_1 \cdot \phi^{\text{IN}} + \lambda_2 \cdot \phi^{\text{ON}} + \phi^{\text{AFF}}, \tag{3}$$

where λ_1 and λ_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 $\Delta\mathcal{R}_t^i$. We classify the changes based on human actions into the following four categories, shown in Fig. 5.

Move-in is defined as $\Delta\mathcal{R}_t^i : \langle O^i, H_t \rangle \rightarrow \langle O^i, C_{t+1}(O^i) \rangle$, 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 $t+1$.

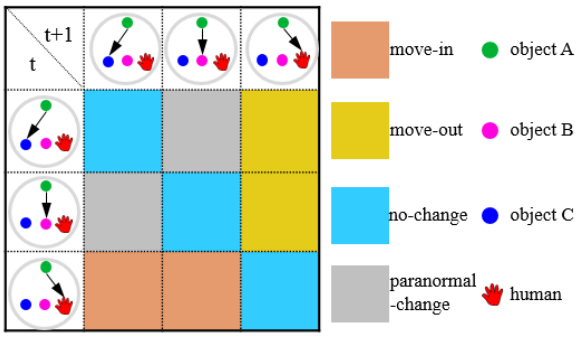

Figure 5: Transition matrix of containment relation changes for object A from frame t to $t+1$. Move-in: the container of A changes from a person to an object. Move-out: the container of A changes from an object to a person. No-change: the container of A does not change. Paranormal-change: the containment relation changes without human intervention violate the *temporal assumption*, thus are ruled out.

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

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 $\mathcal{R}_t^i$ 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

$$\psi(\mathcal{G}_t, \mathcal{G}_{t+1}, V_{[t-\epsilon,t+\epsilon]}) = \langle \omega_{\mathcal{L}_j}, \theta \rangle, \tag{4}$$

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

For θ, 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 $\mathcal{F}_m^h$, the relative movements between the human hands and the object $\mathcal{F}_m^r$, and the movements of the object $\mathcal{F}_m^o$. Suppose that the sliding window size is 2ϵ, the feature vector sequence at time t is $\mathcal{F}_m = (\mathcal{F}_m^h, \mathcal{F}_m^r, \mathcal{F}_m^o), m \in [t-1-\epsilon \quad t-1+\epsilon]$. $\mathcal{F}_m^h$ is the relative distance of all skeletons to three base points (two shoulders and a spine point). $\mathcal{F}_m^r$ is the distance between the human hands and the position of the object. $\mathcal{F}_m^o$ is the object position changes during the time interval 2ϵ. A wavelet transform is then applied to $\mathcal{F}_m$. The coefficients at the low frequency are kept as the interaction feature θ. 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 $\{\mathcal{G}_t\}^*$ is defined as:

$$\{\mathcal{G}_t\}^* = \underset{\{\mathcal{G}_t\}}{\operatorname{argmin}} \, \mathrm{E}(\{\mathcal{G}_t\},\{V_t\}) \tag{5}$$

$$= \underset{\{\mathcal{G}_t\}}{\operatorname{argmin}} \left[\mu \sum_{t=1}^{T} \phi(\mathcal{G}_t, V_t) + \sum_{t=1}^{T-1} \psi(\mathcal{G}_t, \mathcal{G}_{t+1}, V_{[t-\epsilon,t+\epsilon]}) \right],$$

where ϕ is the data term which models the energy of containment relations in space, ψ is the smooth term that models the containment relation changes in time, and μ is the trade-off parameter between the spatial-temporal cues.

4 Inference by Dynamic Programming

The goal of the inference process is to find the optimal sequence of containment graphs $\{\mathcal{G}_t\}^*$ for the input RGB-D video by optimizing the energy function defined in Eq. 5.

The time complexity of searching the entire solution space is $O(N^{(N-1)\cdot T})$, where N is the number of objects and T is the video length. It is impractical to brute-force search the entire space.

Fortunately, at each frame, the container of the containee O^i is independent of the container of the containee O^j. 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:

$$\{\mathcal{R}_t^i\}^* = \underset{\{\mathcal{R}_t^i | t=1,\cdots,T\}}{\operatorname{argmin}} \left[\mu \sum_{t=1}^{T} \phi(\mathcal{R}_t^i, V_t) + \sum_{t=1}^{T-1} \psi(\mathcal{R}_t^i, \mathcal{R}_{t+1}^i, V_{[t-\epsilon,t+\epsilon]}) \right]. \tag{6}$$

By aggregating $\{\mathcal{R}_t^i\}^*$ from each object, we can recover the full sequence of containment graphs $\{\mathcal{G}_t\}^*$ 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

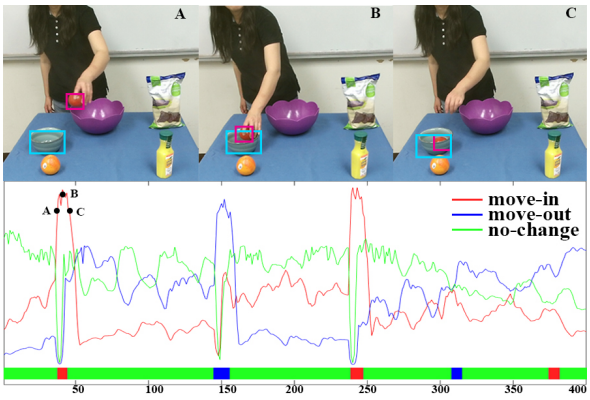

Figure 6: Probability of three different containment relation changes over time between two objects (in red and cyan bounding boxes). The ground truth is shown in the bottom.

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.

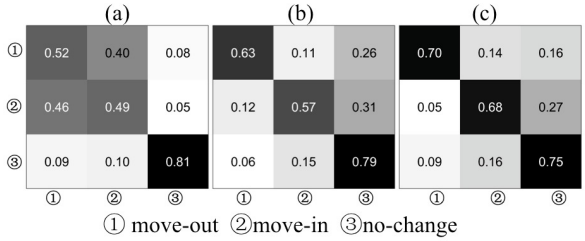

Figure 7: Confusion matrix of relation change recognition. (a) Human actions only. (b) Human actions with object context. (c) Joint inference using the proposed method.

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.

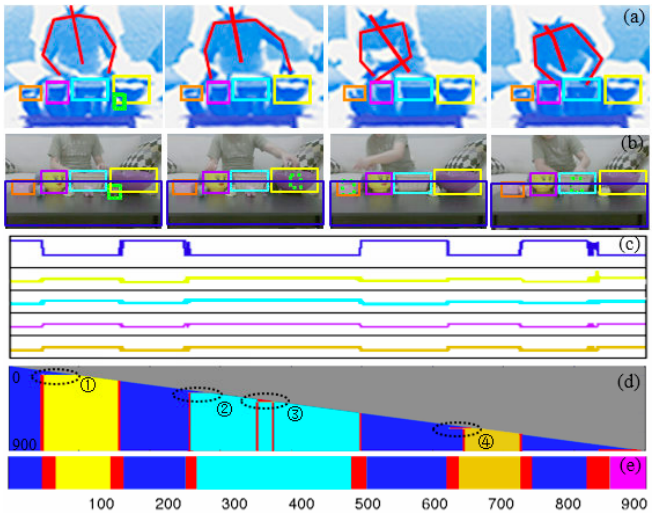

Figure 8: Inference of containment relations for the object in green bounding box. Each color denotes a different object. (a) The tracked objects and the human skeletons in 3D scene. (b) Refined tracking results. We recover the positions of hidden containee using the positions of its container. (c) The probability of the containee contained by each possible container in space. (d) The inference result matrix given different length of the same video. The results are corrected as more information provided ①-④. (e) Ground truth.

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 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.

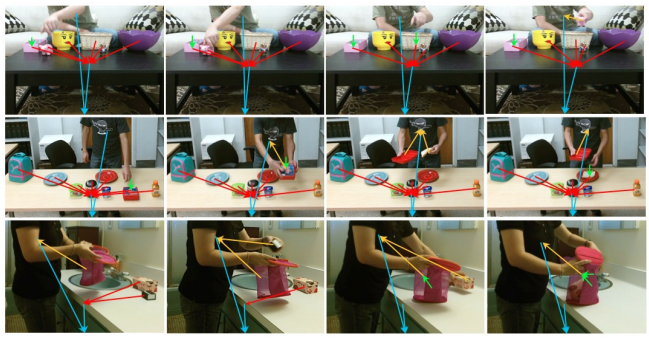

Figure 9: Some qualitative results. Each row shows the result of a specific scene. The arrows represent the containment relations between objects. Each arrow points from one containee to its container. The green arrows denote containment relations between the hidden containees and their containers.

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 %.

Method	no occlusion	partial occlusion	complete occlusion	overall
Baseline	0.75	0.21	0.08	0.37
Ours	0.86	0.64	0.43	0.59

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

[Aksoy *et al.*, 2010] Eren Erdal Aksoy, Alexey Abramov, Florentin Worgotter, and Babette Dellen. Categorizing object-action relations from semantic scene graphs. In *ICRA*, 2010.

[Aydemir *et al.*, 2011] Alper Aydemir, K Sjoo, John Folkesson, Andrzej Pronobis, and Patric Jensfelt. Search in the real world: Active visual object search based on spatial relations. In *ICRA*, 2011.

[Bredeweg and Forbus, 2003] Bert Bredeweg and Kenneth D Forbus. Qualitative modeling in education. *AI magazine*, 24(4):35, 2003.

[Cohn and Hazarika, 2001] Anthony G. Cohn and Shyamanta M. Hazarika. Qualitative spatial representation and reasoning: An overview. *Fundamenta informaticae*, 46(1-2):1–29, 2001.

[Collins and Forbus, 1987] John W Collins and Kenneth D Forbus. Reasoning about fluids via molecular collections. In *AAAI*, 1987.

[Davis *et al.*, 2013] Ernest Davis, Gary Marcus, and Angelica Chen. Reasoning from radically incomplete information: The case of containers. *Advances in Cognitive Systems*, pages 273–288, 2013.

[Feddema *et al.*, 1997] John T Feddema, Clark R Dohrmann, Gordon G Parker, Rush D Robinett, Vicente J Romero, and Dan J Schmitt. Control for slosh-free motion of an open container. *Control Systems, IEEE*, 17(1):29–36, 1997.

[Frank, 1992] Andrew U Frank. Qualitative spatial reasoning about distances and directions in geographic space. *Journal of Visual Languages & Computing*, 3(4):343–371, 1992.

[Frank, 1996] Andrew U Frank. Qualitative spatial reasoning: Cardinal directions as an example. *International Journal of Geographical Information Science*, 10(3):269–290, 1996.

[Gerevini and Renz, 2002] Alfonso Gerevini and Jochen Renz. Combining topological and size information for spatial reasoning. *Artificial Intelligence*, 137(1):1–42, 2002.

[Grabner *et al.*, 2011] Helmut Grabner, Juergen Gall, and Luc Van Gool. What makes a chair a chair? In *CVPR*, 2011.

[Grenon and Smith, 2004] Pierre Grenon and Barry Smith. Snap and span: Towards dynamic spatial ontology. *Spatial cognition and computation*, 4(1):69–104, 2004.

[Gupta *et al.*, 2011] Abhinav Gupta, Scott Satkin, Alexei A Efros, and Martial Hebert. From 3d scene geometry to human workspace. In *CVPR*, 2011.

[Hedau *et al.*, 2010] Varsha Hedau, Derek Hoiem, and David Forsyth. Thinking inside the box: Using appearance models and context based on room geometry. In *ECCV*. 2010.

[Hespos and Spelke, 2007] Susan J Hespos and ES Spelke. Precursors to spatial language: The case of containment. *The categorization of spatial entities in language and cognition*, pages 233–245, 2007.

[Hudelot *et al.*, 2008] Céline Hudelot, Jamal Atif, and Isabelle Bloch. Fuzzy spatial relation ontology for image interpretation. *Fuzzy Sets and Systems*, pages 1929–1951, 2008.

[Koppula *et al.*, 2013] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970, 2013.

[Li, 2007] Sanjiang Li. Combining topological and directional information for spatial reasoning. In *IJCAI*, 2007.

[Liang *et al.*, 2015] Wei Liang, Yibiao Zhao, Yixin Zhu, and Song-Chun Zhu. Evaluating human cognition of containing relations with physical simulation. CogSci, 2015.

[Nitti *et al.*, 2014] Davide Nitti, Tinne De Laet, and Luc De Raedt. Relational object tracking and learning. In *ICRA*, 2014.

[Papadias and Sellis, 1994] Dimitris Papadias and Timos Sellis. Qualitative representation of spatial knowledge in two-dimensional space. *The VLDB Journal*, 3(4):479–516, 1994.

[Pei *et al.*, 2013] M Pei, Z Si, B Yao, and SC Zhu. Video event parsing and learning with goal and intent prediction. *Computer Vision and Image Understanding*, 117(10):1369–1383, 2013.

[Renz, 2012] Jochen Renz. Implicit constraints for qualitative spatial and temporal reasoning. In *KR*, 2012.

[Satkin *et al.*, 2012] Scott Satkin, Jason Lin, and Martial Hebert. Data-driven scene understanding from 3d models. In *BMVC*, 2012.

[Sokeh *et al.*, 2013] Hajar Sadeghi Sokeh, Stephen Gould, and Jochen Renz. Efficient extraction and representation of spatial information from video data. In *AAAI*, 2013.

[Song and Xiao, 2013] Shuran Song and Jianxiong Xiao. Tracking revisited using rgbd camera: Unified benchmark and baselines. In *ICCV*, 2013.

[Strickland and Scholl, 2015] Brent Strickland and Brian J Scholl. Visual perception involves event-type representations: The case of containment versus occlusion. 2015.

[Wang *et al.*, 2014a] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning and recognition. In *CVPR*, 2014.

[Wang *et al.*, 2014b] Xinchao Wang, Engin Türetken, François Fleuret, and Pascal Fua. Tracking interacting objects optimally using integer programming. In *ECCV*. 2014.

[Wei *et al.*, 2013] Ping Wei, Yibiao Zhao, Nanning Zheng, and Song-Chun Zhu. Modeling 4d human-object interactions for event and object recognition. In *ICCV*, 2013.

[Wong *et al.*, 2013] Lawson LS Wong, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. Manipulation-based active search for occluded objects. In *ICRA*, 2013.

[Yang *et al.*, 2009] Ming Yang, Hua Gang, and Ying Wu. Context-aware visual tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(7):1195–1209, 2009.

[Yoon *et al.*, 2015] Ju Hong Yoon, Ming-Hsuan Yang, Jongwoo Lim, and Kuk-Jin Yoon. Bayesian multi-object tracking using motion context from multiple objects. In *WACV*, 2015.

[Zhao and Zhu, 2013] Yibiao Zhao and Song-Chun Zhu. Scene parsing by integrating function, geometry and appearance models. In *CVPR*, 2013.

[Zhu *et al.*, 2014] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *ECCV*, pages 408–424. 2014.

[Zhu *et al.*, 2015] Yixin Zhu, Yibiao Zhao, and Song Chun Zhu. Understanding tools: Task-oriented object modeling, learning and recognition. In *CVPR*, 2015.

[Zhu *et al.*, 2016] Yixin Zhu, Chenfanfu Jiang, Yibiao Zhao, Demetri Terzopoulos, and Song-Chun Zhu. Inferring forces and learning human utilities from videos. In *CVPR*, 2016.