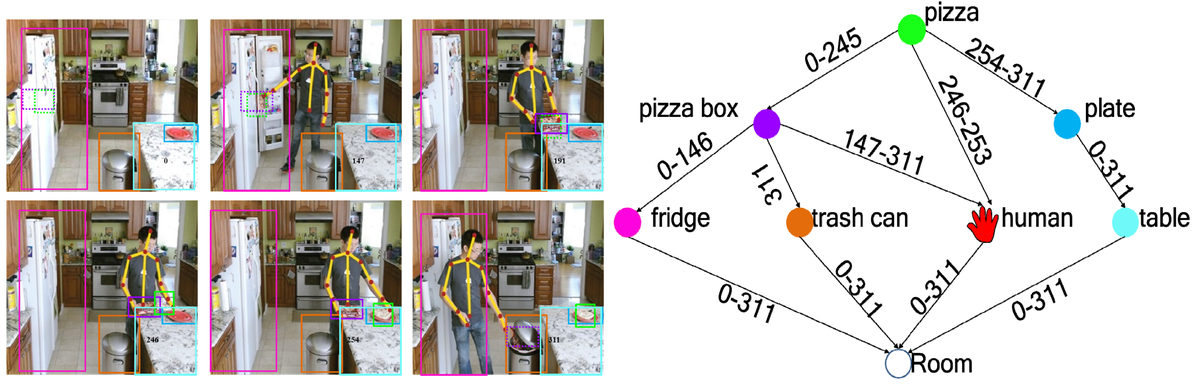
What is Where: Inferring Containment Relations from Videos

Wei Liang*♦ Yibiao Zhao* Yixin Zhu* Song-Chun Zhu*
♦ Beijing Institute of Technology, China *University of California, Los Angeles

Motivation

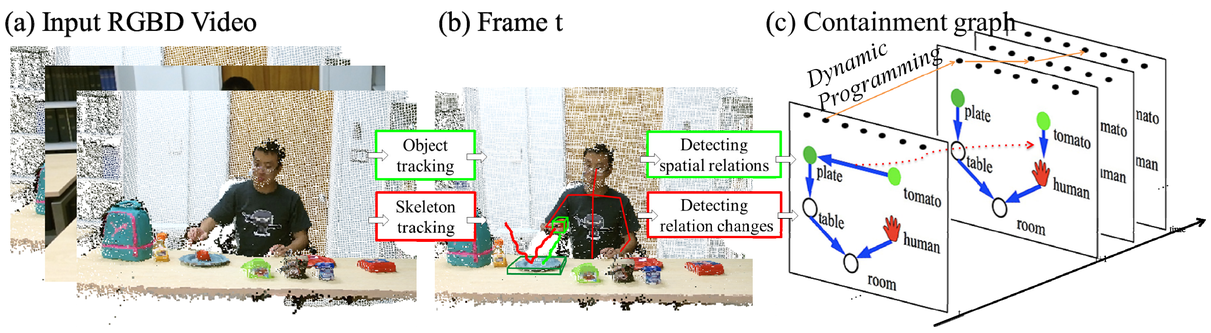
Where did the girl get the pizza? | *Who is holding the yellow ball?* | *Where are the red shoes?*

We present a probabilistic approach to explicitly infer containment relations between objects in 3D scenes. The proposed method is aim to address two tasks:
(a) Recovering hidden objects with severe occlusions.
(b) Inferring subtle human actions.



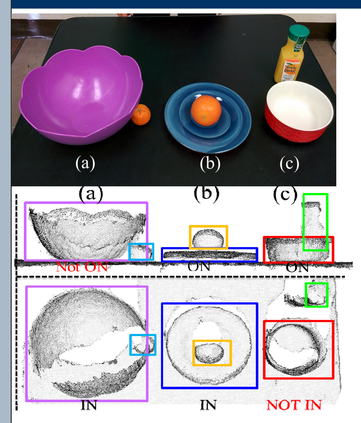
Structured, qualitative and abstract interpretation of containment relations over time in a scene. The goal is to answer "what is where over time". (Right) The inferred containment relations. The numbers on edges denote the frames when the containment relations occur.

Framework



(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.

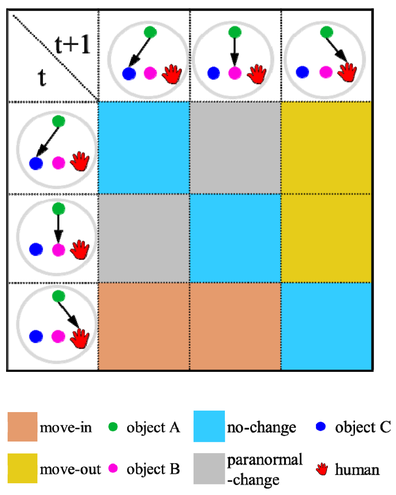
In Space



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

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

In Time



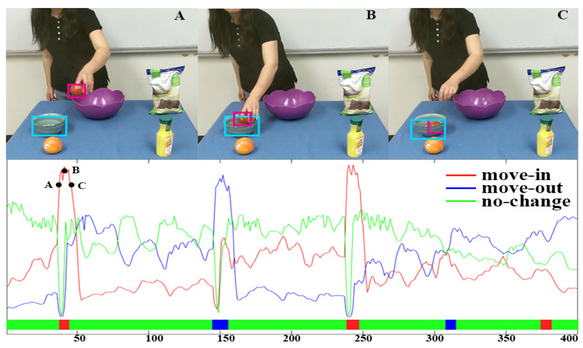
Containment relation changes for object A from frame t to t + 1.
Move-in: the container of A changes from a person to an object.
Move-out: the container of A changes from an object to a person.
No-change: the container of A does not change.
Paranormal-change: the containment relation changes without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

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

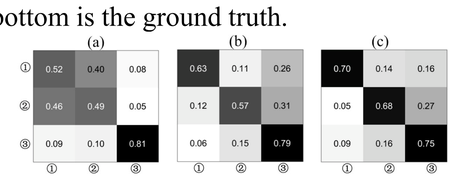
Energy Function

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

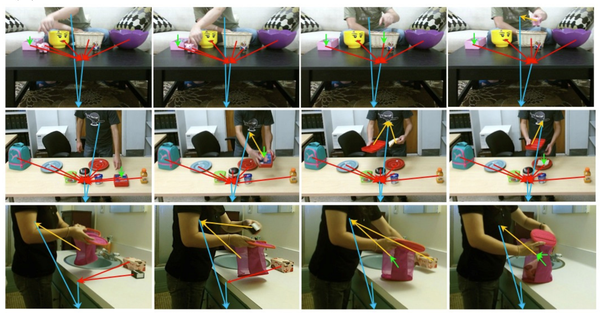
Experiment Results



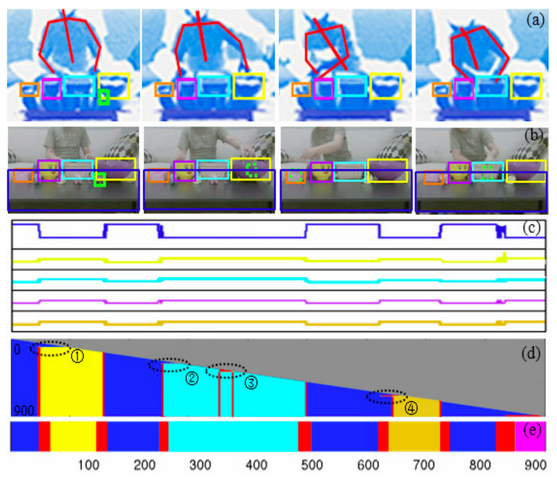
Probability of three relation changes over time between two objects (bounded by boxes). The bar in the bottom is the ground truth.



Confusion matrix of relation change recognition.
(a) Human pose sequence only.
(b) Human pose sequence with objects context.
(c) Joint inference in our method.



The arrows represent the containment relations between objects. Each arrow points from one object to its container.



Inference of containment relations for object in green bounding box. Each color denotes an object.
(a) The tracked objects and the human skeletons.
(b) Refined tracking results.
(c) The probability of the object 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.
(e) Ground truth.

Methods	no occlusion	partial occlusion	complete occlusion	overall
Baseline	0.73	0.21	0.08	0.37
Ours	0.82	0.76	0.54	0.65

Accuracy of containment relations in %. N-occlusion, P-occlusion and C-occlusion are no occlusion, partial occlusion and complete occlusion situation respectively.

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.