Transferring Objects: Joint Inference of Container and Human Pose

Hanqing Wang[1] Wei Liang[1*] Lap-Fai Yu[2]
{whqueryk@gmail.com, liangwei@bit.edu.cn, craigyu@cs.umb.edu}
[1]Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing, China
[2]Graphics and Virtual Environments Laboratory, Department of Computer Science, University of Massachusetts, Boston, USA

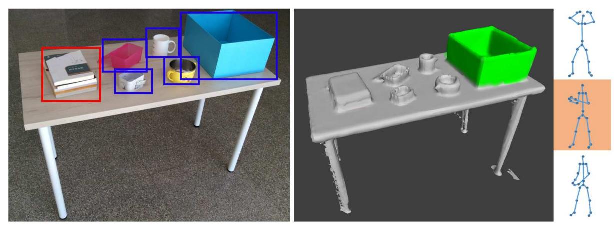
Abstract

*Transferring objects from one place to another place is a common task performed by human in daily life. During this process, it is usually intuitive for humans to choose an object as a proper container and to use an efficient pose to carry objects; yet, it is non-trivial for current computer vision and machine learning algorithms. In this paper, we propose an approach to jointly infer container and human pose for transferring objects by minimizing the costs associated both object and pose candidates. Our approach predicts which object to choose as a container while reasoning about how humans interact with physical surroundings to accomplish the task of transferring objects given visual input. In the learning phase, the presented method learns how humans make rational choices of containers and poses for transferring different objects, as well as the physical quantities required by the transfer task (*e.g.*, compatibility between container and containee, energy cost of carrying pose) via a structured learning approach. In the inference phase, given a scanned 3D scene with different object candidates and a dictionary of human poses, our approach infers the best object as a container together with human pose for transferring a given object.*

1. Introduction

Given a set of containees (red in Figure 1(a)), which object to serve as a container, and what pose is proper to transfer those containees to another place? When transferring an object from one place to another, a person will consider the physical quantities during the transfer task, *e.g.*, compatibility between container and containees, energy cost of carrying pose, *etc*. In this paper, we propose an approach to learn how humans make rational choices and reason about a proper container (green) and a pose (orange) for transferring objects from one place to another, as shown in Figure 1(b).

The overview of our approach is illustrated in Figure 2.



(a) task (b) solution

Figure 1. Which container and what pose is proper to transfer the given objects to another place? (a) Given a scanned scene as the input, we first detect all the objects: the objects bounded by the red rectangle are the *targeted containees* and the ones bounded by the blue rectangles are the *candidate containers*. (b) Our approach infers the best object as a container (green) as well as choosing the best pose (orange) from a pose dictionary.

Our approach takes a 3D scanned scene as the input, and uses a learned structured SVM to analyze the compatibility between containees, container and pose. By assuming the human judgments are near-optimal, we formulate the presented study as a ranking problem, and infer the best container and pose to carry out the transfer task.

Solving this inference problem will allow computers to predict and reason about how humans perform transfer tasks in everyday environments, and hence achieve better understanding and visual perception of the affordance of our physical surroundings.

This paper makes the following three major contributions:

- We propose a new task by joint inference of optimal container and human pose for transferring objects from one place to another.
- We present a ranking-based framework capable of reasoning about and selecting the best container and human pose to perform a transfer task.
- We propose a 3D scanned objects dataset, on which we perform experiments to validate the effectiveness of our framework, demonstrating that the performance of our approach is close to human judgment.

* Corresponding author (liangwei@bit.edu.cn).

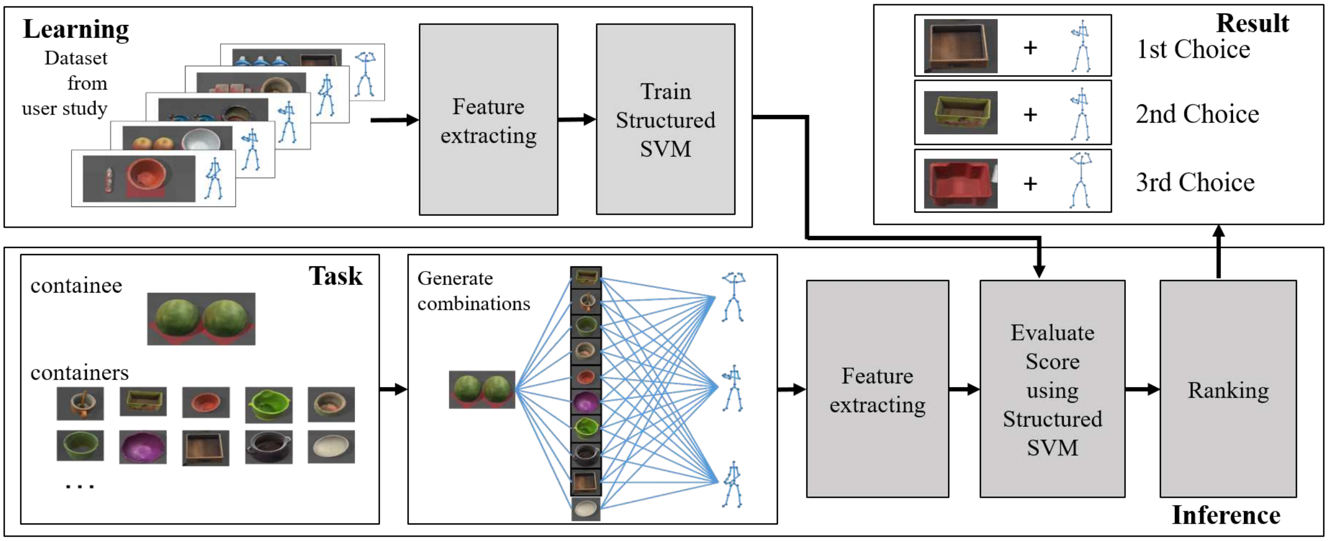


Figure 2. Overview of our approach. The input consists of the target containees and a set of candidate containers scanned from a real-world scene for the transfer **task**. In the **learning** phase, our approach learns how humans transfer different objects with different containers and poses. In the **inference** phase, our approach reasons about the best container and pose for performing the given transfer task.

1.1. Related Work

Understanding Affordance and Tool: Understanding object affordance [6, 8] from images and videos is a challenging task in computer vision. Instead of focusing on the appearance or the geometry of a given object [11, 25, 33], the concept of affordance tries to recognize objects based on their functionality and the end states of the objects, making understanding tool-use a perfect topic in the field of affordance. For instance, a hammer could be used to change the location of nails; a saw or an axes may be used to change the appearance of wood; containers are used to change the organization of their contained objects. Recently, physics-based reasoning approaches have been successfully applied in computer vision to reason about affordances [31, 12, 35, 36, 37, 39, 33, 30] given visual inputs.

Container can be viewed as half-tool [1] in which containability is the key affordance. Human cognition on containers has been extensively studies in the field of cognitive science, including some recent work on containability [22] with rigid body simulation, basin experiment [3, 19] and pouring prediction [18] with fluid simulation. In contrast, the problem of container has been rarely studied in the field of computer vision. Some recent notable work tried to integrate simulation [34], reasoning about containability and containment relations [23, 29, 28].

In this paper, we analyze how compatible a container and a pose are with respect to transferring the targeted containees. Different from previous work, we not only define a transfer task and several attributes encoding geometry features and physical concepts, but also define human energy cost to carry out the task.

Human Pose Prior: Analyzing the pose during the interactions with an object is another effort to understand the affordance of objects. In computer vision, recent work tried to use human pose prior to analyze the functionality of objects or scenes [2, 7, 9, 13, 14, 15, 17, 38, 26]. The human pose is an important prior. Human pose inference and object understanding can also reinforce each other in analyzing interaction activities. Kim et al. proposed a data-driven approach to infer the human pose in using an object based on geometry features [15]. Yao et al [31] used human poses to discover the functionality of an object for computer vision tasks such as object detection. They further demonstrated that pose estimation can be inferred from the functionality of the object [32]. Moreover, the specific human pose and object form a unique human-object interaction for a certain functionality of the object.

In comparison with previous approaches, our approach infers human pose and containers jointly during transferring objects. We consider not only the appearance of the human poses, but also the semantic meaning and physical cost of the poses.

2. Problem Formulation

We define an object transferring task $T(O)$ to transfer the targeted containees O from one place to another place. The goal of our approach is to infer an optimal container c^* to contain the containees and an optimal pose p^* to carry the container. The solution of the transfer task is represented by a tuple $s^* = (c^*, p^*)$.

2.1. Ranking Function

We formulate the optimization of s as a ranking problem. The ranking function is defined as

$$R(s_{ij}) = \langle \omega, \Psi(s_{ij}, T(O)) \rangle, \tag{1}$$

where $\Psi(\cdot)$ is a joint feature vector defined by the task $T(O)$ and the possible solution $s_{ij} = (c_i, p_j)$, and ω is the coefficient vector of feature vector $\Psi(\cdot)$. Here, $c_i \in \{c_1, c_2, \cdots, c_I\}$ represents a candidate container, I is the number of candidate containers, and $p_j \in \{p_1, p_2, p_3\}$ represents a candidate pose. In this paper, we consider three common poses: carrying around waist p_1, carrying around chest p_2 and carrying above head p_3. The dictionary of these poses can be extended easily.

The joint feature $\Psi(\cdot)$ models the relations among task, container and pose. We decompose it into two terms:

$$\Psi(s_{ij}, T(O)) = \psi(O, c_i) + \phi(\hat{c}_i, p_j), \tag{2}$$

where $\psi(O, c_i)$ models the compatibility between containees in the given task and the candidate container, $\phi(\hat{c}_i, p_j)$ models the compatibility between the container and the pose when the pose is taken to carry a container $\hat{c}_i$, and $\hat{c}_i$ represents the container c_i contains the containees, which has different attributes from original c_i, *e.g.* mass and height.

2.2. Compatibility of Containee and Container

$\psi(O, c_i)$ is a joint feature of containees and container, evaluating the compatibilty between them. We consider three factors: containability $\psi_c(O, c_i)$, efficiency $\psi_e(O, c_i)$ and stability $\psi_s(O, c_i)$. $\psi(O, c_i)$ is defined by the sum of these three terms:

$$\psi(O, c_i) = \underbrace{\psi_c(O, c_i)}_{\text{containability}} + \underbrace{\psi_e(O, c_i)}_{\text{efficiency}} + \underbrace{\psi_s(O, c_i)}_{\text{stability}}. \tag{3}$$

Containability $\psi_c(O, c_i)$ models the compatibility between the container and containees from the perspective of volume. We define a volume ratio: $\eta = \frac{V_O}{V_{c_i}}$m where V_O and V_{c_i} represent the volume of containees and container, respectively. Then we have

$$\psi_c(O, c_i) = \begin{cases} e^{-\frac{(\eta-\mu)^2}{2\delta^2}} & \eta \le 1 \\ 0 & \eta > 1 \end{cases}, \tag{4}$$

where μ is the mean of the best ratio and δ is the coefficient, which are learn from human study.

Efficiency $\psi_e(O, c_i)$ models the efficiency of the container choice. It is intuitive that when a person tries to accomplish an object transferring task, they prefer to choose a lighter-weighted container rather than a heavier one, resulting in spending less extra work in carrying. $\psi_e(O, c_i)$ is defined as:

$$\psi_e(O, c_i) = \frac{1}{1 + M_O/M_{c_i}}, \tag{5}$$

where M_O is the mass of containees, and M_{c_i} is the mass of container.

Stability $\psi_s(O, c_i)$ models the stability of containees in a container. Considering the case in which a higher mass center of containees increases the risk of spill out, we model $\psi_s(O, c_i)$ by the height of mass center:

$$\psi_s(O, c_i) = \begin{cases} 1 - \frac{1}{1+H_O/H_{c_i}} & H_O \le H_{c_i} \\ 1 & H_O > H_{c_i} \end{cases}, \tag{6}$$

where H_O is the height of containees' mass center, and H_{c_i} is the height of the container's mass center.

2.3. Compatibility of Container and Pose

$\phi(\hat{c}_i, p_j)$ is a joint feature of container and pose, where $\hat{c}_i$ represents the container with updated attributes when it is containing the containees. We adopt two terms to model the compatibility between container and pose: convenience $\phi_c(\hat{c}_i, p_j)$ and energy cost $\phi_e(\hat{c}_i, p_j)$.

$$\phi(\hat{c}_i, p_j) = \underbrace{\phi_c(\hat{c}_i, p_j)}_{\text{convenience}} + \underbrace{\phi_e(\hat{c}_i, p_j)}_{\text{energy}}, \tag{7}$$

$\phi_c(\hat{c}_i, p_j)$ evaluates the convience of the pose which is taken to carry the container. $\phi_e(\hat{c}_i, p_j)$ is the energy cost when a person carries container $\hat{c}_i$ with pose p_j.

Convenience $\phi_c(\hat{c}_i, p_j)$ models the compatibility between the container and the pose. In the results reported by Knapik *et al*. [16], it is suggested that lower load placement is preferred for stability; people prefers to carry objects on the hands because this pose is more convenient with high movement freedom. However, the load location is also restricted by the appearance of the object. Higher load placement occupies more spaces than the space occupied using lower load placements. Thus, there is a trade-off between the convenience and afforhdance. According to Knapik's study, we define $\phi_c(\hat{c}_i, p_j)$ as

$$\phi_c(\hat{c}_i, p_j) = \lambda(H_{p_j} - W_{\hat{c}_i} + a) + (1-\lambda)\frac{b}{H_{p_j} - W_{\hat{c}_i} + c}, \tag{8}$$

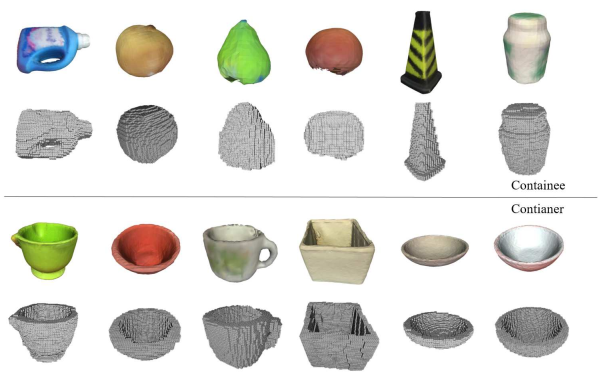


Figure 3. Examples of voxelization. For each pair, the top row shows the original meshes, and the bottom shows the voxelization results.

where a, b, c are the coefficients learned by cross-validation, λ is a trade-off parameter, H_{p_j} is the height of the load location with the pose p_j, and $W_{\hat{c}_i}$ is the width of the container if it is containing the containees.

Energy $\phi_e(\hat{c}_i, p_j)$ models how much work a person does when they take a pose to carry the container. We adopted the results reported in Knapik's study [16] to estimate the energy cost of carrying a container. Assuming the carried container is near the center of the person who takes the carrying pose, the basic energy cost of the pose is:

$$\begin{aligned} M =& 1.5W + 2(W + M_{\hat{c}_i})(\frac{M_{\hat{c}_i}}{W})^2 \\ &+ N(W + M_{\hat{c}_i})(1.5V^2 + 0.35VG), \end{aligned} \tag{9}$$

where W is the weight of the person (set as 65 kg), $M_{\hat{c}_i}$ is the mass sum of the container and containees, V is the walking velocity (set as 4.2 km/h), and G is the slope or grade (set as 1). In this paper, we assume that this energy cost will not change over time.

A ratio is applied to approximate real energy costs for different poses. The ratio is calculated by the distance to the mass center of a person. In our experiments, we use the ratio of 1.2, 1.5, and 1.9 for carrying around chest p_2, carrying around waist p_1 and carrying above head p_3, respectively. Thus, $\phi_e(\hat{c}_i, p_j)$ is defined as

$$\phi_e(\hat{c}_i, p_j) = \gamma_{p_j} M, \tag{10}$$

where γ_{p_j} is the ratio of pose p_j.

2.4. Physical Attributes Estimation

In this section, we introduce how to estimate volume of container and containees in the task, which is used in the ranking function.

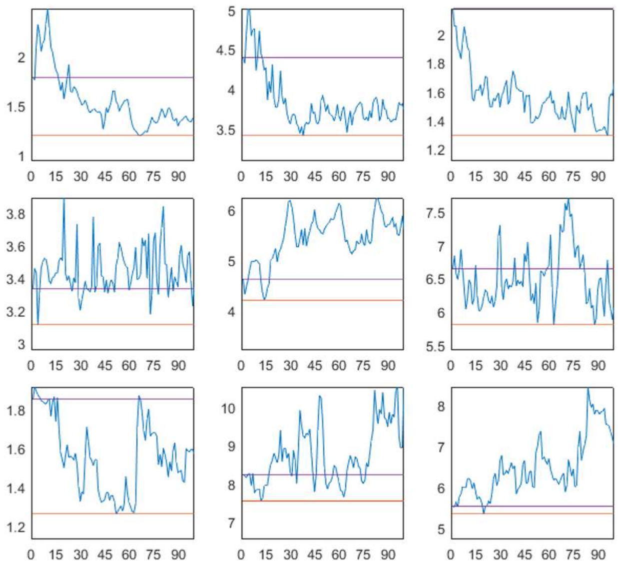


Figure 4. 9 examples of simulation. The plot shows the energy changes during the simulation. The x-axis refers to the sequence of the simulation, the y-axis refers to the energy of the system, the purple line indicates the initial energy, and the orange line indicates the lowest energy.

Volume of container. We apply voxelization to estimate the volume of 3D model. The raw input of our approach is reconstructed 3D models using a depth camera. Inspired by Yu's work [33], we voxelize the input 3D mesh and fill up the inside space. Figure 3 shows some examples. Then we count the number of voxels as the estimation of the container's volume. For each container c_i, we define the volume $V_{c_i} = \sum_{l=0}^{L} v_l$, where v_l is the unit volume of voxel, L is the number of voxel which is filled in the container.

Volume of containees. Since each container may be able to contain more than one containees, we estimate the volume of containees using a physics-based simulation approach. One simple way is to randomly put containees into a container, and count the volume of the objects after reaching the stable state, resulting in an estimated volume of containees. However, such estimation may not be accurate enough to reflect the volume in real-world, as the objects may be accidentally stable supported by the container, making the estimated volume larger than the expected.

Inspired by the intuitive physics theory [27, 20], we further add the disturbance during the simulation, preventing accidental stable events. Specifically, we put all the containees into one container while shaking the container. All configurations are recorded through the shaking process. Lower is the potential energy, more stable is the system. When the potential energy of the configuration goes beyond the adjacent peak, it slips to another local optimal configuration. The minimal space occupied in the simulation process

Figure 5. Some 3D mesh examples of containees (left) and containers (right) in our dataset.

is used as the volume of containees. As shown in Figure 4, a minimal energy is reached during the simulation which is marked with an orange line.

3. Learning Human Utilities

3.1. Rational Choice Assumption

Rational choice assumption means that human choices are rational and near-optimal [4, 5, 10, 24]. In this case, when a person chooses a container and a pose to transfer objects from a place to another, the choice obeys the rule of minimizing the transfer cost, considering the attributes of containees, container and pose.

Under the rational choice assumption, we consider the choices made by human are near-optimal. Assuming that the rational configuration is $s^* = (c^*, p^*)$, for a random configure $s_{ij} = (c_i, p_j)$, it will have lower score than the rational choice in one task. That is, in one task $T(O)$, for all i, j, $s^* \neq s_{ij}$, we have

$$R(s^*) > R(s_{ij}). \tag{11}$$

3.2. Learning and Inference

Learning the coefficient vector ω on training data is solved by a structured learning appraoch. The optimization function is

$$\begin{aligned} \min \quad & \frac{1}{2}\omega \cdot \omega + \lambda \sum_k \xi_k^2, \qquad (12) \\ s.t. \quad & \forall s \in C \times P \setminus s^*, \\ & \langle \omega \cdot \Psi(s^*, T(O_k)) \rangle - \langle \omega \cdot \Psi(s, T(O_k)) \rangle > 1 - \xi_k^2, \\ & \xi_k \geq 0, \end{aligned}$$

where ξ is the slack variable to avoid overfitting, λ is the trade-off parameter to keep the balance between maximizing the margin and satisfying the constraints, k is the number of tasks in training dataset, $C = \{c_1, c_2, \cdots, c_I\}$, and $P = \{p_1, p_2, p_3\}$.

In the inference phase, we reason about the optimal container and pose by maximize our ranking function:

$$s^* = \operatorname*{argmax}_s \langle \omega \cdot \Psi(s, T(O)), \rangle \tag{13}$$

where $s \in C \times P$.

4. Experiments

In this section, we first introduce our dataset. Then we evaluate our approach from four aspects: (i) accuracy of our approach on different scale dataset; (ii) validation of features; (iii) containability of object; and (iv) expansibility on depth data.

4.1. Dataset

We collect a 3D object dataset for our experiment, including 302 scanned 3D objects, ranging from typical tools, household objects, to large pieces of furniture. All meshes are captured by consumer-level RGB-D sensors, and are divided into containee and container based on geometry and category. Some examples are shown in Figure 5.

Using this dataset, we design a transferring task dataset with a collection of 400 tasks, each of which is to move given objects from one place to another. Those tasks are generated randomly. Each task includes targeted containees, twelve candidate containers and three poses.

To annotate those tasks, we build a questionnaire system to collect human data, where users were shown the image of a task. They were asked to choose the best container to hold containees and a proper pose to carry the container. We collect the data from over 200 people whose ages range from 18 to 56. We select the most frequent answer as the

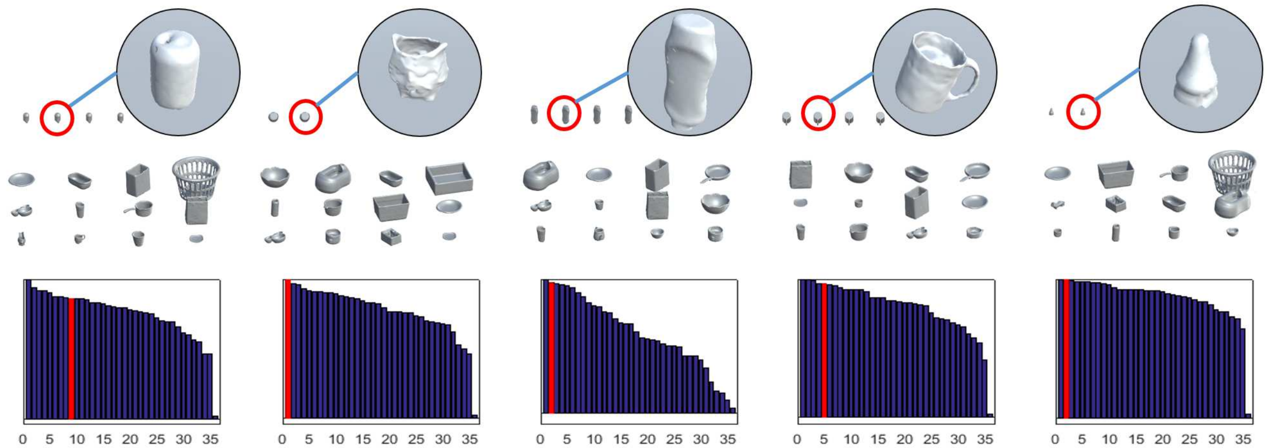


Figure 6. Ranking results using the proposed approach. Top: the target containees. Middle: the candidate containers. Bottom: rankings of configurations, in which the red bar is the position of the ground truth.

ground truth for each task. We use 268 tasks as training data and 132 tasks as testing data.

4.2. Inference of Container and Pose

Since each task includes 12 candidate containers and 3 poses in our dataset, there are 36 potential candidate configurations in total in each task. The goal of the inference is to rank those configurations and evaluate the results by comparing with ground truth. Figure 6 showcases five examples of the configuration ranking and the comparison with ground truth. Most ground truths fall in the top 10 configurations.

Since human judgments have variations and the choices of human are near-optimal, we evaluate our results of prediction using a top-12 criteria: if the ground truth is one of the top 12 configurations of the predicted objects, we consider the prediction as a correct prediction. Under this evaluation, the accuracy of our approach is 66.67%.

To evaluate the influence of container and pose during ranking, we analyze the top 3 containers with a fixed pose. Figure 7 illustrates the scores of the top 3 containers together with the score of ground truth (highlighted) of two cases. From the third column of the first case and the first two columns of the second case, we find that our algorithm learned a diverse human poses choosing for different containers.

In the first case (left), the configuration of the ground truth get the highest score. It is interesting that the ground truth pose (p_1) is not the most energy-saving pose compared with p_2. The reason is that carrying with p_1 will decrease

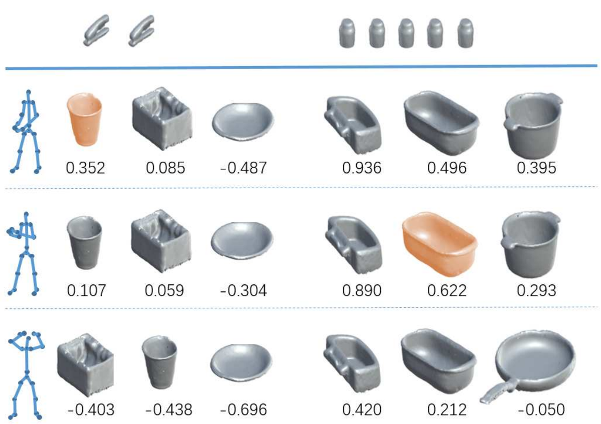


Figure 7. The top 3 container under different human poses in two tasks (left and right). First row show the containees, and the remaining rows are inferred container and the human pose. The number below a container is the score of the configuration estimated by our approach. The highlighted configuration is the ground truth.

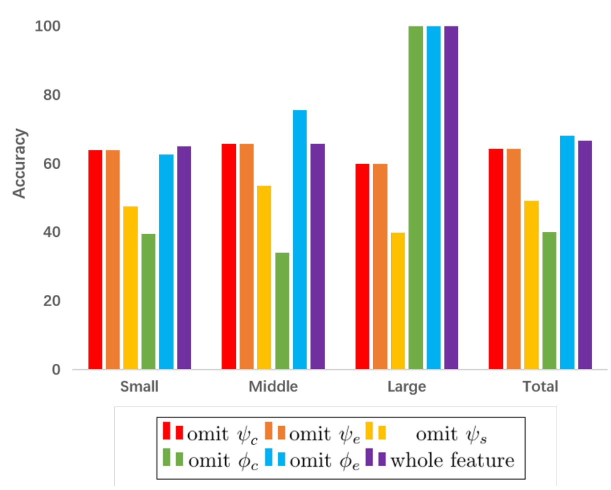


Figure 8. Accuracy of different features omitted in Small, Middle, Large and Total sets, respectively.

Model	Testing Set	Top 3	Top 6	Top 9	Top 12	Top 15	Top 18	Top 24
Omit ψ_c term	Small	20.93%	44.19%	51.16%	63.95%	70.93%	79.07%	93.02%
	Middle	21.95%	26.83%	56.10%	65.85%	75.61%	80.49%	92.68%
	Large	20.00%	20.00%	40.00%	60.00%	60.00%	60.00%	80.00%
	Total	21.21%	37.88%	52.27%	64.39%	71.97%	78.79%	92.42%
Omit ψ_e term	Small	16.28%	33.72%	48.84%	63.95%	76.74%	80.23%	96.51%
	Middle	12.20%	41.46%	51.22%	65.85%	68.29%	75.61%	90.24%
	Large	60.00%	60.00%	60.00%	60.00%	60.00%	80.00%	100.00%
	Total	16.67%	37.12%	50.00%	64.39%	73.48%	78.79%	94.70%
Omit ψ_s term	Small	10.47%	19.77%	33.72%	47.67%	61.63%	75.58%	86.05%
	Middle	4.88%	12.20%	26.83%	53.66%	60.98%	65.85%	80.49%
	Large	40.00%	40.00%	40.00%	40.00%	80.00%	80.00%	100.00%
	Total	9.85%	18.18%	31.82%	49.24%	62.12%	72.73%	84.85%
Omit ϕ_c term	Small	9.30%	17.44%	24.42%	39.53%	50.00%	59.30%	81.40%
	Middle	7.32%	14.63%	19.51%	34.15%	56.10%	65.85%	80.49%
	Large	40.00%	60.00%	60.00%	100.00%	100.00%	100.00%	100.00%
	Total	9.85%	18.18%	24.24%	40.15%	53.79%	62.88%	81.82%
Omit ϕ_e term	Small	17.44%	34.88%	50.00%	62.79%	72.09%	82.56%	91.86%
	Middle	12.20%	39.02%	63.41%	75.61%	78.05%	82.93%	90.24%
	Large	60.00%	60.00%	80.00%	100.00%	100.00%	100.00%	100.00%
	Total	17.42%	37.12%	55.30%	68.18%	75.00%	83.33%	91.67%
Whole feature	Small	17.44%	34.88%	51.16%	65.12%	73.26%	81.40%	94.19%
	Middle	17.07%	46.34%	58.54%	65.85%	70.73%	70.73%	87.80%
	Large	40.00%	40.00%	60.00%	100.00%	100.00%	100.00%	100.00%
	Total	18.18%	38.64%	53.79%	66.67%	73.48%	78.79%	92.42%

Table 1. Results using different models tested in different datasets. **Top n** indicates the ratio of the ground truth ranked in the first n configurations.

the cost of convenience. We can also observe the similar results on the other containers. For example, the container of the second column with pose p_1 has a higher score than the other two poses.

In the second case (right), the configuration of the ground truth get the third high score. The container achieved the highest score has less volume than the third highest score container. Human may think that the first container has no enough volume to contain those containees due to noise of perception. In such situation, people tend to choose the container with a little surplus space for object transferring task.

4.3. Validation of Features

To analyze the usefulness of each term of the feature in our model, we compare the accuracy of the model by turning off some terms.

In this experiment, we designed four testing set: "Small", "Middle", "Large" and "Total". The containees whose diameter are smaller than 15 cm are clustered as the "Small" set, the containees whose diameter are larger than 15 cm and smaller than 65 cm are clustered as the "Middle" set, the containees whose diameter are larger than 65cm are clustered as the "Large" set, and "Total" is the set that includes all of the testing data in different scales. We test the model with all the features, and compare to the models with one feature omitted. A bar plot is shown in Figure 8.

The more detailed analysis of the ranking accuracy is listed in Table 1. Both the model that omits ϕ_c and the model that omits ψ_s have a marked performance drop in accuracy, indicating the importance of this two feature terms. The model that omits ϕ_e achieves a higher accuracy than the whole feature model in the "Middle" set except for the evaluations of **Top 3** and **Top 6**. The reason is that human is not sensitive to the energy cost when the differences of energy changes are not significant.

4.4. Containability of Object

The 3D meshes in our dataset are manually divided into containee set and container set. However, in reality, many objects are multi-functional, *i.e.*, an object can be served as both containee and container based on different contexts information. We try to use our approach to infer the affordance, more specifically, the containability of objects. In this experiment, the candidate container set is not labeled. We merge the target objects set and the container set as the candidate container set. For each task, we use the highest score among the scores of a certain object in different car-

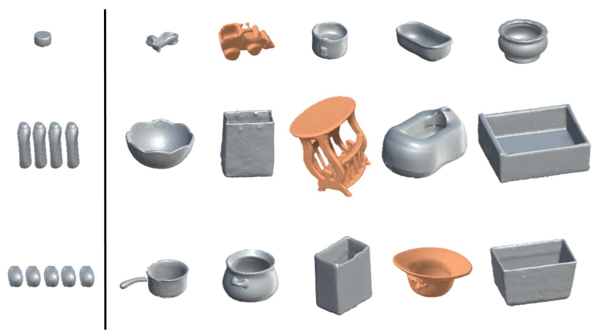


Figure 9. Each row illustrates the top 5 objects used as the ”container” in a task. The objects on the left of the vertical line are the containees. The highlighted objects are in ”containee” sets.

rying poses to represent the score of this object and rank according to the scores.

Figure.9 shows three examples of this experiment. We list the top 5 candidate ”containers”. Our approach inferred not only the ordinary containers but also some normal objects labelled as containee in our previous experiments which annotate container by geometry and category of an object. In the first task, a toy car with a crate is inferred as a good container. In the second task, a stool is inferred. In the third task, a hat is inferred. The common ground of those objects is that they have the functional basis which is able to contain the containees, further, our approach inferred the affordance of the object in containing task.

4.5. Testing on Depth Input

To test the performance of our approach with different kind of input, we use the depth of the task scene as the input of our approach. Given a RGB-D scene, as shown in Figure.10, we segment the objects and re construct them using the default functionalities provided by the Structure Sensor SDK. After that,we normalize the scale according to the depth of objects. The target objects are labelled manually. After that, we retrive the segmented objects in our dataset to find the most similar 3D model[21]. Then we use the depth of the objects to recover the scale of each 3D model. The last step is to use our approach to estimate score of all solutions.

We test 30 scenes and the accuracy is about 63.33%, close to the global accuracy described in Section 4.2. We find that some bad matching from the depth to the 3D model may lead to the failure.

5. Limitations and Future work

In this paper, we propose an approach to jointly infer container and human pose for transferring objects. We formulate the optimization of container and pose inference as a ranking problem, considering the compatibility of con-

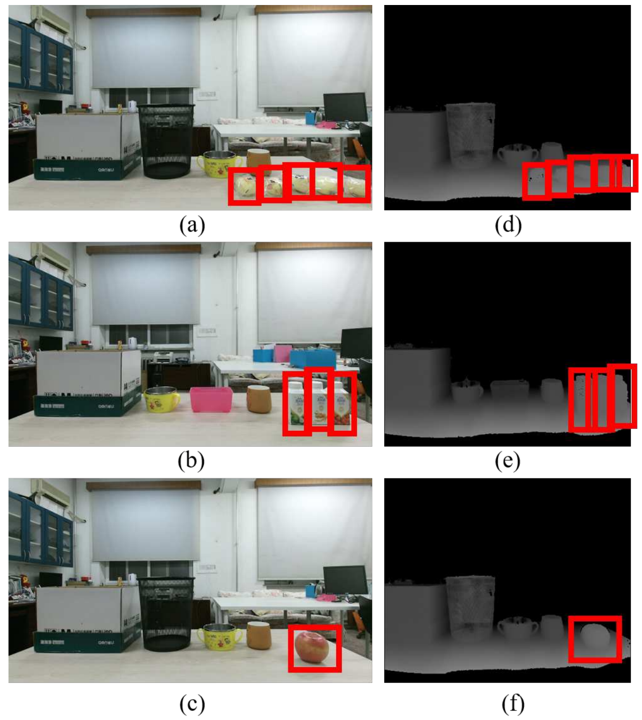


Figure 10. Three examples of our tests on the depth input. (a)(b)(c) are the task scene captured by RGB camera. (d)(e)(f) is the corresponding scene captured by Kinect2 depth camera. The objects in the red bounding box is the target objects.

tainee, container and pose. Our current work has several limitations that we will address in future research.

Currently, the input of our approach is the labeled 3D scene. In the future, we would like to recognize the task scene in an unsupervised fashion. In addition, extending the presented work using 2D information instead of 3D would be an interesting directions. Furthermore, current objects in the dataset only includes rigid objects; incorporating liquid, sand, deformable objects would also make a promising future direction.

Our approach also has some limitations in human pose recognition. Currently, we do not incorporate the grasping pose during the interactions with containers. In the future, it would make a finer-grained recognition if we could generate the proper pose while taking grasping into consideration.

Acknowledgment

This research is supported by the Joseph P. Healey Research Grant Program provided by the Office of the Vice Provost for Research, Strategic Initiatives & Dean of Graduate Studies of UMass Boston, a Natural Science Foundation of China (NSFC) grant No.61472038 and No.61375044, a National Science Foundation 1565978. We acknowledge NVIDIA Corporation for graphics card donation.

References

[1] C. Baber. *Cognition and tool use: Forms of engagement in human and animal use of tools*. CRC Press, 2003.

[2] E. Bar-Aviv and E. Rivlin. Functional 3d object classification using simulation of embodied agent. 2006.

[3] C. Bates, P. Battaglia, I. Yildirim, and J. B. Tenenbaum. Humans predict liquid dynamics using probabilistic simulation. In *CogSci*, 2015.

[4] G. S. Becker. Crime and punishment: An economic approach. In *Essays in the Economics of Crime and Punishment*, pages 1–54. NBER, 1974.

[5] L. E. Blume and D. Easley. Rationality. In *The New Palgrave Dictionary of Economics*. Palgrave Macmillan Basingstoke, UK, 2008.

[6] J. J. Gibson. The theory of affordances. *Lawrence Erlbaum*, 1977.

[7] H. Grabner, J. Gall, and L. V. Gool. What makes a chair a chair? 2011.

[8] J. G. Greeno. Gibson's affordances. *Psychological Review*, pages 336–342, 1994.

[9] A. Gupta, S. Satkin, A. A. Efros, and M. Hebert. From 3d scene geometry to human workspace. 2011.

[10] P. Hedström and C. Stern. Rational choice and sociology. In *The New Palgrave Dictionary of Economics*, pages 872–877. Palgrave Macmillan Basingstoke, UK, 2008.

[11] R. Hu, O. Van Kaick, B. Wu, H. Huang, A. Shamir, and H. Zhang. Learning how objects function via co-analysis of interactions. *Acm Transactions on Graphics*, 35(4):47, 2016.

[12] Z. Jia, A. Gallagher, A. Saxena, and T. Chen. 3d-based reasoning with blocks, support, and stability. 2013.

[13] Y. Jiang, H. S. Koppula, and A. Saxena. Hallucinated humans as the hidden context for labeling 3d scenes. 2013.

[14] Y. Jiang, M. Lim, and A. Saxena. Learning object arrangements in 3d scenes using human context. In *ICML*, 2012.

[15] V. G. Kim, S. Chaudhuri, L. Guibas, and T. Funkhouser. Shape2pose: Human-centric shape analysis. *ACM Trans. Gr.*, 2014.

[16] J. Knapik. Loads carried by soldiers: Historical, physiological, biomechanical and medical aspects. 1989.

[17] H. S. Koppula and A. Saxena. Physically grounded spatio-temporal object affordances. 2014.

[18] J. Kubricht, C. Jiang, Y. Zhu, S.-C. Zhu, D. Terzopoulos, and H. Lu. Probabilistic simulation predicts human performance on viscous fluid-pouring problem. 2016.

[19] J. Kubricht, Y. Zhu, C. Jiang, D. Terzopoulos, S. Zhu, and H. Lu. Consistent probabilistic simulation underlying human judgment in substance dynamics. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, 2017.

[20] J. R. Kubricht, K. J. Holyoak, and H. Lu. Intuitive physics: Current research and controversies. *Trends in Cognitive Sciences*, 2017.

[21] Y. Li, A. Dai, L. Guibas, and M. Nießner. Database-assisted object retrieval for real-time 3d reconstruction. In *Computer Graphics Forum*, volume 34. Wiley Online Library, 2015.

[22] W. Liang, Y. Zhao, Y. Zhu, and S.-C. Zhu. Evaluating human cognition of containing relations with physical simulation. In *Proceedings of the 37th annual conference of the cognitive science society*, 2015.

[23] W. Liang, Y. Zhao, Y. Zhu, and S.-C. Zhu. What is where: Inferring containment relations from videos. In *25th International Joint Conference on Artificial Intelligence*, 2016.

[24] S. Lohmann. Rational choice and political science. In *The New Palgrave Dictionary of Economics*. Palgrave Macmillan Basingstoke, UK, 2008.

[25] A. Myers, C. L. Teo, C. Fermller, and Y. Aloimonos. Affordance detection of tool parts from geometric features. In *IEEE International Conference on Robotics and Automation*, pages 1374–1381, 2015.

[26] S. Qi, S. Huang, P. Wei, and S.-C. Zhu. Predicting human activities using stochastic grammar. 2017.

[27] T. D. Ullman, E. Spelke, P. Battaglia, and J. B. Tenenbaum. Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, 2017.

[28] K. M. Varadarajan and M. Vincze. Afnet: The affordance network. 2012.

[29] X. Wang, E. Türetken, F. Fleuret, and P. Fua. Tracking interacting objects optimally using integer programming. In *Computer Vision–ECCV 2014*, pages 17–32. Springer, 2014.

[30] J. Wu, I. Yildirim, J. J. Lim, B. Freeman, and J. Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in neural information processing systems*, pages 127–135, 2015.

[31] B. Yao and L. Fei-Fei. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1691–1703, 2012.

[32] B. Yao, J. Ma, and L. Fei-Fei. Discovering object functionality. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2512–2519, 2013.

[33] L. F. Yu, N. Duncan, and S. K. Yeung. Fill and transfer: A simple physics-based approach for containability reasoning. In *IEEE International Conference on Computer Vision*, pages 711–719, 2015.

[34] L.-F. Yu, N. Duncan, and S.-K. Yeung. Fill and transfer: A simple physics-based approach for containability reasoning. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[35] B. Zheng, Y. Zhao, J. C. Yu, K. Ikeuchi, and S.-C. Zhu. Beyond point clouds: Scene understanding by reasoning geometry and physics. 2013.

[36] B. Zheng, Y. Zhao, J. C. Yu, K. Ikeuchi, and S.-C. Zhu. Detecting potential falling objects by inferring human action and natural disturbance. 2014.

[37] Y. Zhu, A. Fathi, and L. Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *European conference on computer vision*, pages 408–424. Springer, 2014.

[38] Y. Zhu, C. Jiang, Y. Zhao, D. Terzopoulos, and S.-C. Zhu. Inferring forces and learning human utilities from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3823–3833, 2016.

[39] Y. Zhu, Y. Zhao, and S. C. Zhu. Understanding tools: Task-oriented object modeling, learning and recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2855–2864, 2015.