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 humans 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 with both object and pose candidates. Our approach predicts which object to choose as a container while reasoning about how humans interact with the 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 containee, 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.

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.
1.1. Related Work

Understanding Affordance and Tool: Understanding object affordance from images and videos is a challenging task in computer vision. Instead of focusing on the appearance or the geometry of a given object, 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 given visual inputs. Container can be viewed as half-tool in which containability is the key affordance. Human cognition on containers has been extensively studied in the field of cognitive science, including some recent work on containability with rigid body simulation, basin experiment and pouring prediction. In contrast, the problem of container has been rarely studied in the field of computer vision. Some recent notable work tried to integrate simulation, reasoning about containability and containment relations.

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. 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. Yao et al. 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. 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

\[
R(s_{ij}) = \langle \omega, \Psi(s_{ij}, T(O)) \rangle,
\]

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 \( \omega \) 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:

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

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 Containeer and Container

\( \psi(O, c_i) \) is a joint feature of containees and container, evaluating the compatibility 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 the three terms:

\[
\psi(O, c_i) = \psi_c(O, c_i) + \psi_e(O, c_i) + \psi_s(O, c_i).
\]

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

\[
\psi_c(O, c_i) = \begin{cases} 
    e^{-\frac{(\eta-\mu)^2}{2\delta^2}} & \eta \leq 1 \\
    0 & \eta > 1
\end{cases},
\]

where \( \mu \) is the mean of the best ratio and \( \delta \) 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:

\[
\psi_e(O, c_i) = \frac{1}{1 + M_O/M_{c_i}},
\]

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:

\[
\psi_s(O, c_i) = \begin{cases} 
    1 - \frac{1}{1 + H_O/H_{c_i}} & H_O \leq H_{c_i} \\
    1 & H_O > H_{c_i}
\end{cases},
\]

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 \( \phi_e(\hat{c}_i, p_j) \).

\[
\phi(\hat{c}_i, p_j) = \phi_c(\hat{c}_i, p_j) + \phi_e(\hat{c}_i, p_j),
\]

\( \phi_c(\hat{c}_i, p_j) \) 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. [15], 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 affordance. According to Knapik’s study, we define \( \phi_c(\hat{c}_i, p_j) \) as

\[
\phi_c(\hat{c}_i, p_j) = \lambda(H_{p_j} - W_{\hat{c}_i} + \alpha) + (1 - \lambda) \frac{b}{H_{p_j} - W_{\hat{c}_i} + c},
\]

where \( \lambda \) is the coefficient.
where $a$, $b$, $c$ are the coefficients learned by cross-validation, $\lambda$ 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:

$$M = 1.5W + 2(W + M_{\hat{c}_i})(\frac{M_{\hat{c}_i}}{W})^2 + N(W + M_{\hat{c}_i})(1.5V^2 + 0.35VG),$$

(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

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

(10)

where $\gamma_{p_j}$ 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.

**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...
3. Learning Human Utilities

3.1. Rational Choice Assumption

Rational choice assumption means that human choices are rational and near-optimal \cite{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

\[
R(s^*) > R(s_{ij}).
\]  

(11)

3.2. Learning and Inference

Learning the coefficient vector \( \omega \) on training data is solved by a structured learning approach. The optimization function is

\[
\min \frac{1}{2} \omega \cdot \omega + \lambda \sum_k \xi_k^2,
\]  

(12)

s.t. \( \forall s \in C \times P \setminus s^* \),

\[
\langle \omega \cdot \Psi(s^*, T(O_k)) \rangle - \langle \omega \cdot \Psi(s, T(O_k)) \rangle > 1 - \xi_k^2;
\]

\( \xi_k \geq 0 \),

where \( \xi \) is the slack variable to avoid overfitting, \( \lambda \) 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, \( C = \{c_1, c_2, \ldots, c_I\} \), 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:

\[
s^* = \arg\max_s \langle \omega \cdot \Psi(s, T(O)), \rangle
\]  

(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.

![Figure 5. Some 3D mesh examples of containees (left) and containers (right) in our dataset.](image-url)

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.
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 the...
<table>
<thead>
<tr>
<th>Model</th>
<th>Testing Set</th>
<th>Top 3</th>
<th>Top 6</th>
<th>Top 9</th>
<th>Top 12</th>
<th>Top 15</th>
<th>Top 18</th>
<th>Top 24</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Small</td>
<td>20.93%</td>
<td>44.19%</td>
<td>51.16%</td>
<td>63.95%</td>
<td>70.93%</td>
<td>79.07%</td>
<td>93.02%</td>
</tr>
<tr>
<td></td>
<td>Middle</td>
<td>21.95%</td>
<td>26.83%</td>
<td>56.10%</td>
<td>65.85%</td>
<td>75.61%</td>
<td>80.49%</td>
<td>92.68%</td>
</tr>
<tr>
<td></td>
<td>Large</td>
<td>20.00%</td>
<td>20.00%</td>
<td>40.00%</td>
<td>60.00%</td>
<td>60.00%</td>
<td>60.00%</td>
<td>80.00%</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>21.21%</td>
<td>37.88%</td>
<td>52.27%</td>
<td>64.39%</td>
<td>71.97%</td>
<td>78.79%</td>
<td>92.42%</td>
</tr>
<tr>
<td>Omit $\psi_e$ term</td>
<td>Small</td>
<td>16.28%</td>
<td>33.72%</td>
<td>48.84%</td>
<td>63.95%</td>
<td>76.74%</td>
<td>80.23%</td>
<td>96.51%</td>
</tr>
<tr>
<td></td>
<td>Middle</td>
<td>12.20%</td>
<td>41.46%</td>
<td>51.22%</td>
<td>65.85%</td>
<td>68.29%</td>
<td>75.61%</td>
<td>90.24%</td>
</tr>
<tr>
<td></td>
<td>Large</td>
<td>60.00%</td>
<td>60.00%</td>
<td>60.00%</td>
<td>80.00%</td>
<td>80.00%</td>
<td>100.00%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>16.67%</td>
<td>37.12%</td>
<td>50.00%</td>
<td>64.39%</td>
<td>73.48%</td>
<td>78.79%</td>
<td>94.70%</td>
</tr>
<tr>
<td>Omit $\psi_s$ term</td>
<td>Small</td>
<td>10.47%</td>
<td>19.77%</td>
<td>33.72%</td>
<td>47.67%</td>
<td>61.63%</td>
<td>75.58%</td>
<td>86.05%</td>
</tr>
<tr>
<td></td>
<td>Middle</td>
<td>4.88%</td>
<td>12.20%</td>
<td>26.83%</td>
<td>53.66%</td>
<td>60.98%</td>
<td>75.61%</td>
<td>90.24%</td>
</tr>
<tr>
<td></td>
<td>Large</td>
<td>40.00%</td>
<td>40.00%</td>
<td>40.00%</td>
<td>80.00%</td>
<td>80.00%</td>
<td>100.00%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>9.85%</td>
<td>18.18%</td>
<td>31.82%</td>
<td>49.24%</td>
<td>53.79%</td>
<td>62.88%</td>
<td>84.85%</td>
</tr>
<tr>
<td>Omit $\phi_c$ term</td>
<td>Small</td>
<td>9.30%</td>
<td>17.44%</td>
<td>24.42%</td>
<td>39.53%</td>
<td>50.00%</td>
<td>59.30%</td>
<td>81.40%</td>
</tr>
<tr>
<td></td>
<td>Middle</td>
<td>7.32%</td>
<td>14.63%</td>
<td>19.51%</td>
<td>34.15%</td>
<td>56.10%</td>
<td>65.85%</td>
<td>80.49%</td>
</tr>
<tr>
<td></td>
<td>Large</td>
<td>40.00%</td>
<td>60.00%</td>
<td>60.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>9.85%</td>
<td>18.18%</td>
<td>24.24%</td>
<td>40.15%</td>
<td>53.79%</td>
<td>62.88%</td>
<td>81.82%</td>
</tr>
<tr>
<td>Whole feature</td>
<td>Small</td>
<td>17.44%</td>
<td>34.88%</td>
<td>51.16%</td>
<td>65.12%</td>
<td>73.26%</td>
<td>81.40%</td>
<td>94.19%</td>
</tr>
<tr>
<td></td>
<td>Middle</td>
<td>17.07%</td>
<td>46.34%</td>
<td>58.54%</td>
<td>65.85%</td>
<td>70.73%</td>
<td>70.73%</td>
<td>87.80%</td>
</tr>
<tr>
<td></td>
<td>Large</td>
<td>40.00%</td>
<td>40.00%</td>
<td>60.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>18.18%</td>
<td>38.64%</td>
<td>53.79%</td>
<td>66.67%</td>
<td>73.48%</td>
<td>78.79%</td>
<td>92.42%</td>
</tr>
</tbody>
</table>

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 $\phi_c$ and the model that omits $\psi_s$ have a marked performance drop in accuracy, indicating the importance of this two feature terms. The model that omits $\phi_e$ 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.

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 reconstruct 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 retrieve 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 containee, 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 direction. 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.