A Commonsense Knowledge-based Object Retrieval Approach for Virtual Reality

Haiyan Jiang* [1]   Dongdong Weng⊠† [1]   Xiaonuo Dongye [1]   Nan Zhang [2]   Luo Le [1]

[1]Beijing Institute of Technology; [2]China North Advanced Technology Generalization Institute

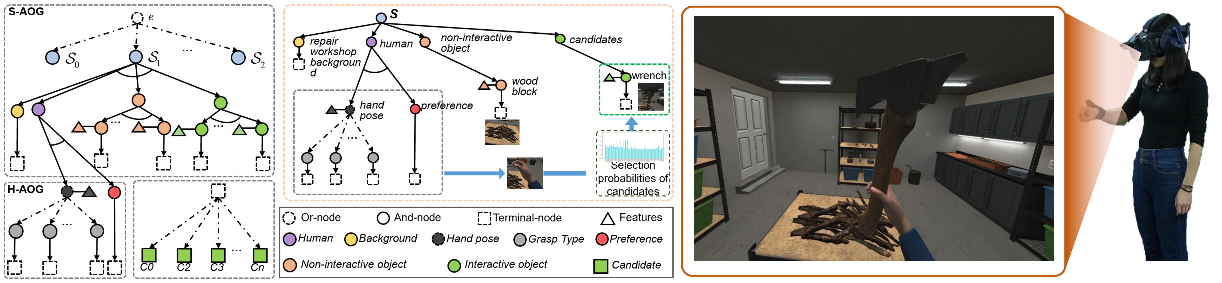

Figure 1: *Left*: Illustration of the SH-AOG. “--→” are used to connect Or-nodes and their child nodes and “→” for And-nodes. Each event graph $\mathcal{G}_e$ is rooted at its corresponding event node e expanding into different scenes which is expanded into a background, non-interactive objects, interactive objects, and a human. The root of H-AOG is a human expanded into preference and hand pose which is expanded into 33 grasp types. Each terminal node of SH-AOG is expanded into all candidates for retrieval. *Right*: In a virtual room, a user performs a grasping gesture. The expected virtual object will be inferred using the parse graph and then generated to the user's virtual hand with an appropriate pose. *Middle*: Instantiating the parse graph. *Bottom in the middle*: legend of graphs.

Abstract

Out-of-reach object retrieval is an important task in many applications in virtual reality. Hand gestures have been widely studied for object retrieval. However, the one-to-one mapping metaphor of the gesture would cause ambiguity problems and memory burdens for the retrieval of plenty of objects. Therefore, we proposed a grasping gesture-based object retrieval approach for out-of-reach objects based on a graphical model called And-Or graphs (AOG), leveraging the scene-object occurrence, object co-occurrence, and human grasp commonsense knowledge. This approach enables users to acquire objects by using natural grasping gestures according to their experience of grasping physical objects. Importantly, users could perform the same grasping gesture for different virtual objects and perform different grasping gestures for one virtual object in the virtual environment.

Index Terms: Human-centered computing—Human computer interaction (HCI)—Interaction techniques—Gestural input.

1 Introduction

Object retrieval is a common and basic task in many applications for virtual reality (VR). Object retrieval could be achieved by bare hands through grasping metaphors like in the physical world. However, even for user-led gestures which have good discoverability, users also need to remember plenty of gestures for different objects [1]. Gestures and objects are one-to-one in the mapping mechanism of these approaches. However, there are many cases users would perform different grasp gestures for one object [1] without training, while they also would perform a similar gesture to grasp different objects which have similar shapes or part shapes.

Studies have focused on understanding the way humans grasp objects [2], reaching a consensus about the range of grasp types that humans commonly use. There is a kind of commonsense knowledge that how users grasp physical objects using different grasp gestures. Moreover, when performing object retrieval, users usually are situated in a specific environment (*e.g.*, car repair training scene, kitchen scene), where the selected object is related to the environment's background. There is also a kind of commonsense knowledge in our life experience about where an object would usually appear [3]. For example, tools for repair (*e.g.*, screwdriver and hammer) are usually used in maintenance environments, while kitchen utensils (*e.g.*, pots and pans) are usually used in kitchen environments. In addition, another kind of commonsense knowledge is that objects could co-occur, which has been widely used for scene synthesis [3]. For example, when there is a chopping board with a carrot, kitchen knives are more likely to appear than screwdrivers.

Therefore, we propose a commonsense knowledge-based object retrieval approach, leveraging the commonsense knowledge of object occurrence, object co-occurrence, and human grasp. A graphical model was used to model the aforementioned commonsense knowledge and represent the object retrieval using a spatial-human And-Or graph (SH-AOG) which is composed of a spatial And-Or graph (S-AOG) and a human And-Or graph (H-AOG). This is inspired by a lot of works in which the graphical models of And-Or graph (AOG) [4] are used to parse the 3D scene and human-object interaction. When users perform object retrieval, the optimal parse graph is utilized to get the acquired candidate which has the highest probability. The proposed approach enables users to acquire virtual objects by performing the grasp gestures based on their grasp experience in the physical world without the one-to-one mapping limitation. With this approach, users could perform different grasp gestures for one object(*e.g.*, a mug could be grasped by the handle and cup body) and perform one grasp for different objects (*e.g.*, the same grasp gesture could grasp the knife, pot, and spatula).

2 Commonsense Knowledge-based Approach

A commonsense knowledge-based object retrieval approach considering the situated scene and human grasp gesture and preference is presented in this section. An And-Or graph (AOG) represents the hierarchical decompositions from a retrieval event (top-level) to object candidates for retrieval (bottom level) by a set of terminal and non-terminal nodes. There are four types of nodes in the And-Or graph, including And-nodes (solid circles), Or-nodes (dashed circled), and terminal nodes (dashed squares), which are Or-nodes as the root expanded into terminal nodes with solid squares. The terminal nodes with solid squares represent our object candidates for retrieval. The non-terminal nodes $V_{NT} = V_{\text{And}} \cup V_{Or}$ encode the graph rules. For an **And-node** $v \in V_{And}$, an **And rule** is defined as $v \to v_1 \cdot v_2 \cdots v_n$. An **Or-node** V_{Or} represents the possibilities of

*E-mail: jianghybit@163.com

†Corresponding author, E-mail: crgj@bit.edu.cn

alternative choices (*e.g.*, the user selects a knife or an apple). For an **Or-node** $v \in V_{Or}$, an **Or rule** is defined as $v \rightarrow v_1|v_2|\cdots|v_n$, with probabilities $p_1|p_2|\cdots|p_n$. A parse graph pg is derived from the AOG by selecting the switches related to Or-nodes.

Particularly, we represent the object retrieval task structure using a SH-AOG, as shown in figure 1. The SH-AOG can be decomposed into two parts: the S-AOG and H-AOG. The root of S-AOG is one Or-node expanded into different scenes, which are expanded into the scene background, the non-interactive object set, the interactive object set, and the human. Thereinto, the non-interactive object set is And-node expanded into the non-interactive objects which would co-occur with object candidates; the interactive object set is an And-node expanded into the interactive objects which are candidates displayed in scenes and could be chosen, and would also co-occur with other candidates. The root of H-AOG is human expanded into hand pose and preference, where hand pose is expanded into 33 grasp types. Each terminal node (dashed squares) of the SH-AOG is one Or-node expanded into different object candidates (terminal nodes with solid squares) based on commonsense knowledge.

Formally, let $\mathcal{G}_e =< e, V_{NT} \cup V_T, \mathcal{R}, \mathcal{P} >$ denote an AOG, where e is the root node of an event, V_{NT} is the set of the non-terminal nodes including the scene' background labels $\mathcal{B}$, the non-interactive object labels $\mathcal{N}$, interactive object labels $\mathcal{I}$, grasp type labels $\mathcal{T}$, and the human preference labels $\mathcal{L}$. V_T is the set of the terminal, $\mathcal{R}$ stands for the production rules, and $\mathcal{P}$ is the probability model defined on the AOG. For an object retrieval event, we extract features of hand poses Γ_H, affordance features of non-interactive objects Γ_{An} and interactive object Γ_{Ai}, and preference features Γ_L. We use ${\Gamma_{An}}^k$ to denote the feature of the k-th non-interactive object and ${\Gamma_{Ai}}^k$ to denote the feature of the k-th non-interactive object.

2.1 Probabilistic Formulation and reference

In this section, we introduce the probabilistic model defined by the SH-AOG. Given the extracted features, the posterior probability of a parse graph sequence pg is defined as:

$$\begin{aligned} p(pg|\Gamma, \mathcal{G}_e) &\propto p(\Gamma|pg)p(pg|\mathcal{G}_e) \\ &= p(\Gamma_H|pg)p(\Gamma_{An}|pg)p(\Gamma_{Ai}|pg)p(\Gamma_L|pg)p(pg|\mathcal{G}_e). \end{aligned} \tag{1}$$

The first four terms are likelihood terms for grasp type, the affordance of non-interactive objects, the affordance of interactive objects, and the preference given a parse graph pg. The last term is a prior probability of the parse graph given the graph $\mathcal{G}_e$.

$$p(\Gamma_H|pg) = p(\Gamma_H|\mathcal{T}) = \frac{p(\mathcal{T}|\Gamma_H)P(\Gamma_H)}{P(\mathcal{T})} \propto p(\mathcal{T}|\Gamma_H), \tag{2}$$

$p(\mathcal{T}|\Gamma_H)$ is the detection probability of a grasp type, which could be gotten through the a neural network-based prediction model.

$$\begin{aligned} p(\Gamma_{An}|pg) = p(\Gamma_{An}|\mathcal{N}) &= \frac{p(\mathcal{N}|\Gamma_{An})P(\Gamma_{An})}{P(\mathcal{N})} \\ &\propto p(\mathcal{N}|\Gamma_{An}) = \prod_{k=1}^{K} p(\mathcal{N}^k|{\Gamma_{An}}^k), \end{aligned} \tag{3}$$

and similarly

$$p(\Gamma_{Ai}|pg) \propto p(\mathcal{I}|\Gamma_{Ai}) = \prod_{k=1}^{K} p(I^k|{\Gamma_{Ai}}^k), \tag{4}$$

where

$$p(\mathcal{N}^k|{\Gamma_{An}}^k) = \frac{1}{d_{An}^k}, \qquad p(\mathcal{I}^k|{\Gamma_{Ai}}^k) = \frac{1}{d_{Ai}^k}. \tag{5}$$

d_{An}^k is the distance (unit: m) between the hand wrist and the k-th non-interactive objects and d_{Ai}^k is the distance between the hand wrist and the k-th interactive objects.

$$p(\Gamma_L|pg) = p(\Gamma_L|\mathcal{L}) \propto p(\mathcal{L}_{\mathcal{K}}|\Gamma_{L_K}). \tag{6}$$

Let $p(\mathcal{L}_{\mathcal{K}}|\Gamma_{L_K})$ denote probability by the ratio of selection times of the k-th candidate to all selection times.

The prior probability of a parse graph pg can be computed by:

$$\begin{aligned} p(pg|\mathcal{G}_e) &= p(pg|\mathcal{G}_{\mathcal{B}})p(pg|\mathcal{G}_{\mathcal{N}})p(pg|\mathcal{G}_{\mathcal{I}})p(pg|\mathcal{G}_{\mathcal{T}}) \\ &= p(pg|\mathcal{G}_{\mathcal{B}})p(pg|\mathcal{G}_{\mathcal{T}}) \prod_{k=1}^{K} p(pg|\mathcal{G}_{\mathcal{N}}^k) \prod_{j=1}^{J} p(pg|\mathcal{G}_{\mathcal{I}}^j), \end{aligned} \tag{7}$$

where $p(pg|\mathcal{G}_{\mathcal{B}})$ represents the probabilities of the candidates that occur in the scene, $p(pg|\mathcal{G}_{\mathcal{T}})$ represents the probabilities of the candidates that could be grasped by the grasp types, $p(pg|\mathcal{G}_{\mathcal{N}}^k)$ represents the probabilities of the candidates that co-occur with the k-th non-interactive object, and $p(pg|\mathcal{G}_{\mathcal{I}}^j)$ represents the probabilities of the candidates that co-occur with the j-th interactive object.

For an object retrieval event, the $pg*$ is the best explain based on the extracted features by maximizing the posterior probability:

$$pg* = \underset{pg}{argmax}\, p(pg|\Gamma, \mathcal{G}_e). \tag{8}$$

3 Object Retrieval System

Based on the SH-AOG, we proposed the object retrieval system. First, the hand pose is predicted by a neural network from an RGB image in real time. Then, another neural network is used for grasp types prediction by using the hand pose; Next, according to the highest probability based on the proposed approach, an object was retrieved. Finally, a new neural network was used to predict the object pose in the wrist coordinate system. As shown in figure 1, VEs are displayed with an HTC Vive Pro. A Logitech C930 Webcam is attached to the head-mounted display (HMD) for capturing hand images. An HTC Vive Tracker is fixed on the user's wrist to obtain the wrist pose which controls the pose of the virtual wrist. All of the virtual objects' 3D models and virtual scenes are obtained from Unity Asset Store or the YCB dataset [5].

4 Conclusion and future work

We model commonsense knowledge using an AOG for object retrieval, including object occurrence in the scene, object co-occurrence, and human grasping commonsense knowledge. The proposed object retrieval approach allows users to select or retrieve virtual objects using grasping gestures as grasping physical objects in the real world without the one-to-one mapping limitation. In the future, the performance of this approach should be investigated.

Acknowledgments

This work was supported by the National Key Research, the Development Program of China (No.2022YFF090230), the Beijing Municipal Science & Technology Commission, and Administrative Commission of Zhongguancun Science Park under Grant Z221100007722002.

References

[1] Yukang Yan, Chun Yu, Xiaojuan Ma, Xin Yi, Ke Sun, and Yuanchun Shi. Virtualgrasp: Leveraging experience of interacting with physical objects to facilitate digital object retrieval. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–13, New York, NY, USA, 2018. Association for Computing Machinery.

[2] Visar Arapi, Cosimo Della Santina, Giuseppe Averta, Antonio Bicchi, and Matteo Bianchi. Understanding human manipulation with the environment: A novel taxonomy for video labelling. *IEEE Robotics and Automation Letters*, 6(4):6537–6544, 2021.

[3] Francesco Giuliari, Geri Skenderi, Marco Cristani, Yiming Wang, and Alessio Del Bue. Spatial commonsense graph for object localisation in partial scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19518–19527, 2022.

[4] Song-Chun Zhu, David Mumford, et al. A stochastic grammar of images. *Foundations and Trends® in Computer Graphics and Vision*, 2(4):259–362, 2007.

[5] Yale-CMU-Berkeley. Ycb benchmarks – object and model set. `https://www.ycbbenchmarks.com/`.