Enhancing Daily Life Through an Interactive Desktop Robotics System

Yuhang Zheng[1], Qiyao Wang[2], Chengliang Zhong[2,3(✉)], He Liang[2], Zhengxiao Han[4], and Yupeng Zheng[5]

[1] Beihang University, Beijing, China
[2] Tsinghua University, Beijing, China
zhongcl19@mails.tsinghua.edu.cn
[3] Xi'an High-Tech Research Institution, Xi'an, China
[4] Beijing University of Chemical Technology, Beijing, China
[5] Institute of Automation, Chinese Academy of Sciences, Beijing, China

Abstract. In this demo, we develop an intelligent desktop operating robot designed to assist humans in their daily lives by comprehending natural language with large language models and performing a variety of desktop-related tasks. The robot's capabilities include organizing cluttered objects on tables, such as dining tables or office desks, placing them into storage cabinets, as well as retrieving specific items from drawers upon request. This paper provides the design, development, and functionality of our robotics system, highlighting its advanced language understanding capabilities, perception algorithms, and manipulation techniques. Through real-world experiments and user evaluations, we demonstrate the effectiveness and practicality of our robotic companion in assisting individuals with everyday desktop tasks.

Keywords: Table organization · Natural language processing · Robotic perception and manipulation

1 Introduction

In recent years, there has been an increasing interest in the development of robotic systems aimed at providing assistance to humans in their daily lives, thereby enhancing convenience, productivity, and overall well-being [1–4]. In this demonstration, our focus lies on the task of table rearrangement: an intelligent robot designed to assist individuals in moving every object on the table to its specific location. This concise paper offers a comprehensive overview of the design, development, and functionality of our robotic companion, aiming to

Y. Zheng and Q. Wang—Contribute equally to this work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-981-99-9119-8_8.

L. Fang et al. (Eds.): CICAI 2023, LNAI 14474, pp. 81–86, 2024.
https://doi.org/10.1007/978-981-99-9119-8_8

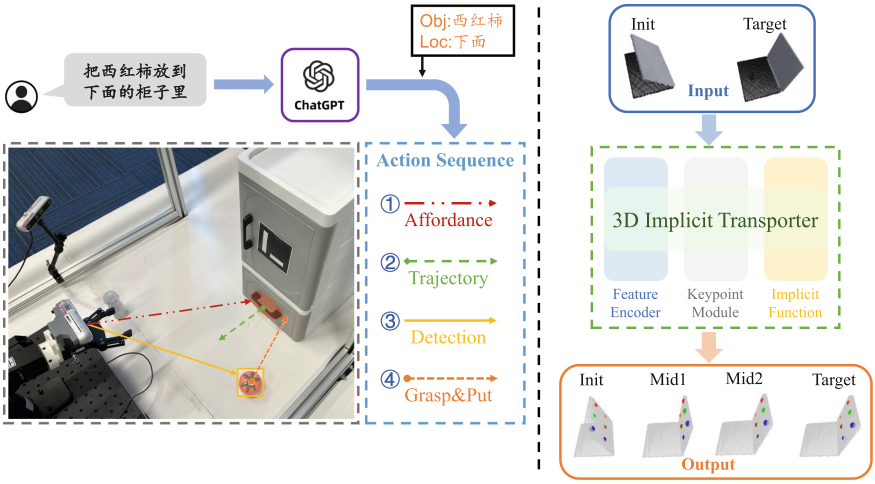

Fig. 1. The architecture of our robotics system. Our speech recognition module analyzes input speech to identify the objects to be manipulated and their target places. Using this information, our robot arm executes a sequence of actions to complete the organization task. The most challenging aspect lies in predicting the trajectory of opening the cabinet door, for which we introduce a 3D Implicit Transporter network to effectively manipulate the articulated object, as depicted in the right panel.

revolutionize how we interact with our immediate workspace and elevate the quality of our lives.

One of the key considerations in designing this robotic system is ensuring its user-friendliness. Humans can effortlessly communicate with the robot, issuing commands and requesting specific actions to be performed. Additionally, the inherent clutter often encountered on desks, be it in home or office settings, presents considerable challenges in terms of automated organization and task management [5]. Furthermore, the system is required to rapidly adapt to novel scenarios, such as encountering new tables and unfamiliar objects.

Our robot has been meticulously engineered to tackle these challenges, offering a reliable and intuitive solution to handle a wide range of desktop-related tasks. Through harnessing advancements in natural language processing, computer vision, and robotic manipulation, this innovative robot is capable of seamlessly understanding user commands, accurately identifying and categorizing objects, and proficiently executing specified actions to enhance desktop organization. We evaluate our system quantitatively on a real-world task and find that our robot achieves a success rate of 87.3% on grasp and correctly places 90.0% of objects.

2 System Architecture

As shown in Fig. 1, the robot system consists of two parts: (1) speech recognition and (2) perception and manipulation. Detailed information on each component is described below.

2.1 Speech Recognition

Taking human speech as input, the speech recognition module first converts voice to text and then outputs the names of organized objects and the parts or locations of the cabinets. In the speech recognition module, we choose Microsoft Azure's speech recognition and speech synthesis cloud services and select ChatGPT [6] as the dialogue interaction robot to achieve the integration of the intelligent interaction system. To enable the robot to flexibly answer various questions related to organizing objects on the desktop, it is necessary to send a human-written prompt to ChatGPT for online training after initializing the interaction system, and then start the dialogue and Q&A. The content of the prompt mainly includes **1)** the robot's responsibilities and main tasks to be completed; **2)** names of objects and the parts or locations of the cabinets; **3)** several example scenarios that illustrate how the robot answers human questions.

2.2 Perception and Manipulation

Affordance Prediction. Affordance refers to the potential uses or actions that a particular object or feature in the environment can offer, such as a door handle affording the action of grasping. The concept of affordance provides valuable semantic information for robot agents, as it enables them to understand how they can interact with their surroundings to perform various tasks. In particular, we leverage the notion of affordance to guide our robot arm's grasping actions. To achieve this, we employ AffCorrs, a one-shot transfer method [7], to generate a grasping affordance map based on a source image containing labeled regions for grasping and an image sequence to be labeled. AffCorrs outputs the grasping region for each frame, which we then use in conjunction with aligned depth maps and camera intrinsic parameters to reconstruct a point cloud representing the location of the grasped object. This approach enables our robot arm to accurately locate and grasp objects in its environment.

Grasp Pose Planning. This function determines the best way for the gripper to hold and lift an object. To grasp the object, we need to obtain its position and estimate the grasping pose of the gripper. Here, we adopt the method of object detection. The system takes RGBD images captured by Realsense D455 as input and a YOLO-v7 network [8], trained with self-collected data, is utilized for detection. For grasping, we combined the detection results with the depth frames got previously to obtain the grasping pose.

Articulated Object Manipulation. Articulated objects play a ubiquitous role in our everyday lives, serving as storage units for various items [9]. Manipulating such objects presents a significant challenge due to the inherent shape variations and dynamic changes in their topology over time. In order to tackle this challenge, we have developed a novel network that leverages temporally consistent keypoints to infer the kinematic structure and movement of different parts. This network builds upon our previous work, denoted as 3D Implicit Transporter.

Our network operates on two input point clouds that capture the dynamic movement of object parts. The primary objective of our method is to reconstruct the shape of the target state by transporting explicit feature grids from the source state based on the predicted 3D keypoints. This process is carried out in a self-supervised manner. By leveraging temporally aligned keypoints, we are able to predict the direction of movement for the articulated object. In contrast to employing a single-step action to reach the desired target, our approach generates a series of sequential actions over an extended timeframe, gradually transforming the articulation state. The effectiveness of these long-horizon sequential actions is demonstrated in our accompanying demo.

In our system, we utilize MoveIt [10], the most popular motion planning framework, to control the movement of the robotic arm. To minimize any potential planning failures, we ensure that the robotic arm returns to its preset home position before every manipulation.

3 Results

We demonstrate the effectiveness of our system in both perception and manipulation tasks. For the object detection module, we achieve 98.2% map@50% on the collected dataset. Additionally, the average relative repeatability of the keypoint prediction is 83.1%. The robot can compute the moving direction with the temporally aligned keypoints, which enables it to perform the correct action to manipulate articulated objects and achieve a success rate of 87% in the Pybullet emulator.

The qualitative results are shown in Fig. 2. We show the key action of our robot arm after parsing the input speech. In addition, (a) and (b) shows the process of our robotic system searching for the target object and the grasping position respectively.

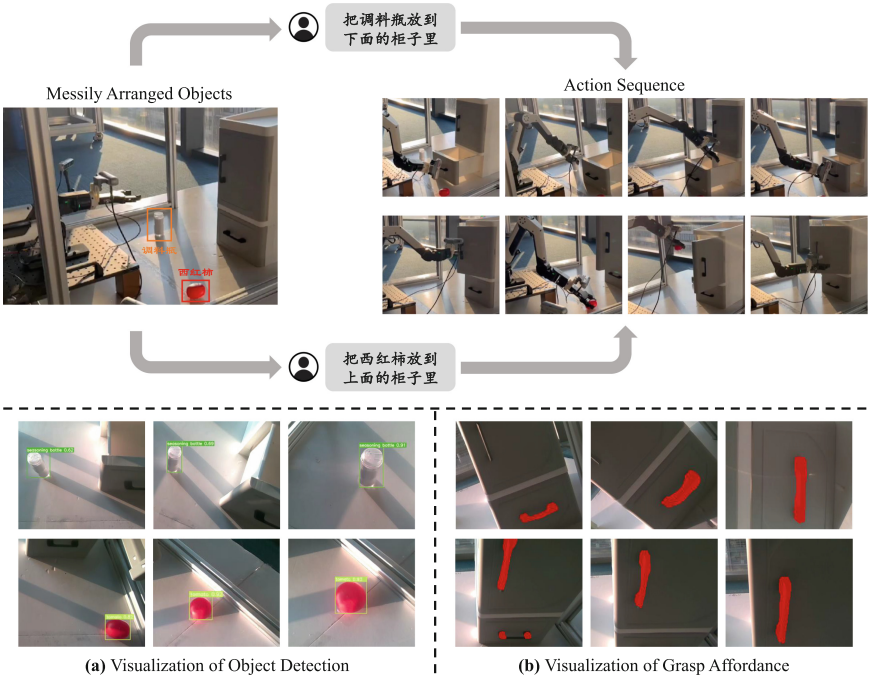

Fig. 2. The results of our robotics system. (a) and (b) visualize the process of our robotic system searching for the target object and the grasping position respectively.

References

1. Billard, A., Kragic, D.: Trends and challenges in robot manipulation. Science **364**(6446), eaat8414 (2019)
2. Shridhar, M., Manuelli, L., Fox, D.: Cliport: what and where pathways for robotic manipulation. In: Conference on Robot Learning. PMLR, pp. 894–906 (2022)
3. Wu, J., Antonova, R., Kan, A., et al.: Tidybot: personalized robot assistance with large language models. arXiv preprint arXiv:2305.05658 (2023)
4. Driess, D., Xia, F., Sajjadi, M.S.M., et al.: Palm-e: an embodied multimodal language model. arXiv preprint arXiv:2303.03378 (2023)
5. Liu, Z., Liu, W., Qin, Y., et al.: Ocrtoc: a cloud-based competition and benchmark for robotic grasping and manipulation. IEEE Robot. Autom. Lett. **7**(1), 486–493 (2021)
6. Ouyang, L., Wu, J., Jiang, X., et al.: Training language models to follow instructions with human feedback. Adv. Neural. Inf. Process. Syst. **35**, 27730–27744 (2022)
7. Hadjivelichkov, D., Zwane, S., et al.: One-shot transfer of affordance regions? affcorrs! In: Conference on Robot Learning. PMLR, pp. 550–560 (2023)
8. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7464–7475 (2023)

9. Jiang, Z., Cheng-Chun, H., Zhu, Y.: Ditto: building digital twins of articulated objects from interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
10. Coleman, D., Sucan, I., Chitta, S., et al.: Reducing the barrier to entry of complex robotic software: a moveit! case study. arXiv preprint arXiv:1404.3785 (2014)