An Intuitive, Affordances Oriented Telemanipulation Framework for a Dual Robot Arm Hand System: On the Execution of Bimanual Tasks

Gal Gorjup, Anany Dwivedi, Nathan Elangovan, and Minas Liarokapis

Abstract—The concept of teleoperation has been studied since the advent of robotics and has found use in a wide range of applications, including exploration of remote or dangerous environments (e.g., space missions, disaster management), telepresence based time optimisation (e.g., remote surgery) and robot learning. While a significant amount of research has been invested into the field, intricate manipulation tasks still remain challenging from the user perspective due to control complexity. In this paper, we propose an intuitive, affordances oriented telemanipulation framework for a dual robot arm hand system. An object recognition module is utilised to extract scene information and provide grasping and manipulation assistance to the user, simplifying the control of adaptive, multi-fingered hands through a commercial Virtual Reality (VR) interface. The system's performance was experimentally validated in a remote operation setting, where the user successfully performed a set of bimanual manipulation tasks.

I. INTRODUCTION

Robot arm hand systems are reaching remarkable levels of speed and accuracy, making them invaluable for applications that require precise and repetitive manipulation in unstructured environments. Apart from working independently, such systems may also be configured to serve as extensions of human operators in scenarios where autonomous operation is infeasible or undesired. In such teleoperation frameworks, full or partial control of the robot agent is granted to the remote user, allowing them to accomplish their intent through the agent. Thus, human guidance enables the agent to solve complex tasks that would not have been feasible to attempt through autonomous system development.

Teleoperation has proven its effectiveness in a wide range of applications that require human skill without their physical presence. For operations in hazardous or inaccessible environments, robot deployment is considerably safer and more practical, compared to human expeditions. In NASA’s Robonaut project [1], for example, a teleoperated robot was developed for the purposes of space station maintenance. After the Fukushima Daiichi nuclear disaster in 2011, the plant environment was too dangerous for recovery through direct human involvement and several remotely operated robots were considered for the missions [2]. Apart from environmental factors, teleoperation may also be desired for the objective of reducing travel-related time delays and increasing operator efficiency. An application with high impact in healthcare is remote surgery, where expert knowledge can be applied over great distances, with no time spent on transport [3], [4]. Other examples also include teleoperation for more commercial purposes such as communication, telepresence, or mobile manipulation in home environments [5]. Opposed to purely practical applications, remote robot operation is also frequently used in research, namely as a means of providing examples in various learning frameworks [6], [7]. In this context, the data is recorded on the robot itself while the human operator compensates for differences in kinematics and dynamics of the platform. The learning framework is therefore not burdened with considering additional embodiment mappings, which generally results in shorter training times and better results.

In manipulation tasks, it is most intuitive for the user to operate a bimanual system that maps to their left and right hand. With two arms, the ability of the system to execute complex tasks drastically increases, especially when equipped with appropriately dexterous hands. While mapping human motion to robot arms in 3D space is relatively straightforward for most applications, issues often arise when it comes to controlling hands and grippers with several degrees of freedom (DoF). Various mapping strategies have been proposed to tackle this challenge [8]–[11], but most of them require human finger pose data captured through expensive sensory systems. Alternatively, when using commercial, widely accessible controllers, managing dexterous hands often becomes difficult, tedious or both. Attempting to directly map individual hand actuators to the limited interface inputs is not feasible and would most likely confuse the user.

On the other hand, defining a fixed set of grasp primitives requires the operator to comb through the grasp types in order to find one appropriate for the target task.

This work presents an assistive, affordance oriented teleoperation framework aimed at simplifying remote bimanual manipulation with dexterous hands using commercial virtual reality controllers. Instead of defining a static library of hand-specific grasp types, the framework relies on visual object recognition to propose a set of affordances to the user. Based on the desired task and object in focus, the system determines an appropriate grasp type that enables successful task execution with minimal input from the user. The framework was tested and experimentally validated to determine its suitability in practical applications.

The rest of this paper is organised as follows: Section II introduces the related work in this field, Section III presents the developed framework and the experimental setup, Section IV presents the obtained results, while Section V concludes the work and discusses future directions.
II. RELATED WORK

As one of the earliest aspects of robotics, the field of teleoperation has seen significant development since its introduction in the early 1950s. In the context of remote manipulation with arm hand systems, teleoperation was in the early stages explored mostly from a control theoretic perspective [12], [13]. Interfaces were generally implemented in the form of classic I/O devices (joysticks, keyboards and monitors) or master/slave robots, where the master module often kinematically resembled the slave to provide intuitive control to the user. As technology progressed, the operated system functionality expanded through increased robot hand dexterity, arm range of motion and number of manipulators. To maximise control efficiency and convenience, alternative interfacing and control options were explored and adapted to the novel frameworks.

The standard solution for intuitively guiding a robot arm is mapping operator motion to the tool position and orientation. Examples employ inertial [14] or magnetic [15] motion sensors to track the human arm and achieve stable arm control. The above works also recorded elbow pose of the human arm and used it to grant a degree of anthropomorphism to the robot motion. Attempts with no sensors mounted on the arm were also explored, but the human motion data obtained through a vision system was noisy and not as reliable [16]. Concerning end-effector control, a trivial parallel gripper may be managed by a simple button or switch accessible by the operator. For frameworks employing dexterous, possibly anthropomorphic hands, a popular choice are data gloves which may track finger poses, forces and wrist orientation [1], [17]. A solution that combines user hand pose tracking and a rich robot hand control potential came in the form of Virtual Reality (VR) interfaces, which were promptly considered for telemanipulation [18]. Some recent examples include utilising the Leap Motion tracking system to obtain hand and finger poses and use them in a gesture-based, VR powered teleoperation framework [19]. In [20], the authors used the commercial HTC Vive VR system to collect teleoperation data in a learning framework for a PR2 robot.

The natural inconsistencies and imperfections of human motion propagate through the tracking system to produce a jittery and, depending on the choice of sensor, noisy signal. If left uncompensated, these errors lead to unstable robot motion which can hinder execution of precise tasks and frustrate the user. To account for this, the concept of shared control and assistive teleoperation was introduced, where the robot operates with a degree of autonomy to reduce user effort. Early assistive frameworks assumed to know the user’s intent [21], which later evolved into classifying motion into a predefined set of paths or behaviours to aid with execution of the task [22], [23]. More recent approaches focused on predicting arbitrary goals in real-world environments which expanded the application range [24], [25]. It is worth noting that the bulk of work in assistive teleoperation targeted movement compensation and neglected any support with grasping and manipulation. Even though the topic of grasp detection and synthesis has received much attention in the robotics research sphere, it has yet to be successfully integrated in remote operation frameworks. This paper moves towards manipulation aid in teleoperation by presenting a framework with an incorporated affordance system that simplifies control of multi-DOF hands with commercial controllers.

Affordances can be described as the sum of "all action possibilities" for a given object [26]. They are characterised by the properties of the object and robot to determine the possible interactions between them [27]. Affordances can either be defined explicitly, where only the object attributes are considered, or implicitly, where information about the action and outcomes is incorporated (object-robot interaction). Studying explicit affordances, the work of [28] extended the affordance attributes to identify visual cues (such as handles) as suitable interaction points. A later work [29] proposed a system for detection of functional object attributes (such as "liftable") based on affordance cues ("handle") from image features. In [27], the authors introduced visual object grouping (balls, boxes, etc.) as an intermediate step that enabled scalability of their affordance methodology. Such explicit affordances only offer object attributes, without considering robot capabilities. Contrary to this, studies such as [30], [31] utilised the implicit affordance representation that involves mapping object and robot affordances to predicted outcomes that can be directly used for planning and control. Authors of [31] proposed object-action complexes to map representational differences between the high level intelligence planning and low level robot control. This was achieved by pairing the transition states of object-action pairs as an instantaneous state transition fragment (ISFT). In [30], the relationships between the actions, objects and effects were encoded in affordances, whose learning and usage was then discussed in detail.

III. METHODS

A. Hardware and Framework

Concerning hardware, the system was based on two 6-DoF serial manipulators by Universal Robots (UR5 and UR10), equipped with adaptive robot hands developed by the New Dexterity research group at the University of Auckland [32]. The control interface was implemented with a commercial virtual reality system (HTC Vive).

The software side of the framework was created within the Robot Operating System (ROS) [33], which provided the necessary communication, testing and visualisation utilities. The framework layout is presented in Fig. 1, where the blocks conceptually correspond to the implemented ROS node architecture. For clarity, the diagram only presents the architecture corresponding to a single arm hand system. For bimanual operation, the arm and hand control capability was efficiently extended.

A VR interface module provides connectivity to the HTC Vive system, tracking the controller poses and button states, in addition to offering haptic feedback functionality through controller vibration. The tracked controller position, orientation and button states are passed to the Pose Mapping and...
inverse kinematics node which maps them to the robot end-effectors and computes the corresponding inverse kinematics (IK). A safety mechanism is incorporated into the module and the controller pose is only forwarded to the robot if an "enable" button is held on the controller. The robot tool position follows the relative controller offset with respect to the initial state where tracking was enabled through the safety switch, while orientation tracking is absolute. The computed robot configuration $q$ is forwarded to the Arm Controller, which is based on the UR modern driver [34] and ROS Control [35] packages. To enable smooth and responsive real-time pose tracking, a closed-loop velocity controller for the UR robots was implemented as the joint speed commands provide best performance with the used robot models and control boxes [36]. The controller loops at 125 Hz, which corresponds to the maximum frequency allowed by the UR system real-time interface.

Affordance analysis is based on object detection performed on a 2D video stream of the teleoperated system workspace. Images from an HD webcam are passed to the Object Detection module employing a pre-trained deep Convolutional Neural Network (CNN) on common household items (Tensorflow ssd_mobilenet_v1_coco model [37]) to extract the labels and bounding boxes for objects of interest. These are processed with methods described in Section III-A to produce a grasp appropriate for the task that the user wishes to execute. The Hand Controller translates the obtained grasp into positions and velocities of motors used in the robot hands. To provide the user with a sense of executed grasp strength, hand motor current is mapped to vibration of the VR controllers. The webcam feed with overlaid object bounding boxes and affordance lists is streamed to a remote screen visible to the operator. A simple selection mechanism was implemented within the affordance analysis module, allowing the user to loop through and select the desired task through buttons on the VR controllers. The selected option is highlighted in the affordance list, where the default selection was for every object set to a neutral rest state to avoid accidental triggering of the grasping motion.

Fig. 1. Framework architecture for the proposed methodology. The robot arm trajectories are controlled using the Vive controllers. The objects in proximity to the robot arms are identified using a camera-based environment sensing module. The grasp affordances of the objects closest to the end-effector are displayed on the visual interface. A particular grasp is selected and executed by the user.

Fig. 2. Example subset of the object set used for affordance analysis.
Functional Affordance
Power
Move
Power
Pinch
Power
Move
Pinch
Power
Move
Pinch
Power
Move
Pinch
Power
Move
Pinch
Power
Move
Pinch
Power
Move
Pinch
Power
Move
Pinch
Power
Move
Pinch
Power

Fig. 3. Bimanual telemanipulation with the developed framework. The user guides robot arms with VR controllers, receiving visual and haptic feedback. The framework offers grasping and manipulation assistance in the form of an affordance menu, where the user selects the option corresponding to the desired task.

<table>
<thead>
<tr>
<th>Objects</th>
<th>Functional Affordance</th>
<th>Grasp Affordance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple</td>
<td>Move - Side</td>
<td>Power</td>
</tr>
<tr>
<td></td>
<td>- Top</td>
<td>Pinch</td>
</tr>
<tr>
<td>Cup</td>
<td>Move - Side</td>
<td>Power</td>
</tr>
<tr>
<td></td>
<td>- Top</td>
<td>Pinch</td>
</tr>
<tr>
<td></td>
<td>Drink</td>
<td>Power</td>
</tr>
<tr>
<td>Sports Ball</td>
<td>Move - Throw</td>
<td>Tripod</td>
</tr>
<tr>
<td></td>
<td>Click</td>
<td>Spherical</td>
</tr>
<tr>
<td>Bottle</td>
<td>Move - Side</td>
<td>Power</td>
</tr>
<tr>
<td></td>
<td>- Top</td>
<td>Pinch</td>
</tr>
<tr>
<td></td>
<td>Drink</td>
<td>Power</td>
</tr>
<tr>
<td>Mouse</td>
<td>Move - Click</td>
<td>Palm</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Palm</td>
</tr>
</tbody>
</table>

TABLE I  FUNCTIONAL AND GRASP AFFORDANCES OF THE EXAMPLE OBJECT SET.

B. Object Affordance Analysis

The implicit approach to affordance analysis (refer to Section II) is used when linking the object functionality and robot grasp affordances to find actions that can be performed on the object by the robot hand. For this proof of concept, the object and grasp affordances were defined manually and an example subset for the objects examined in Fig. 2 is presented in Table I.

Following object detection through methods presented in Section III-A, the objects are filtered to find the ones with which the user most likely intends to interact. This is achieved by projecting the current robot end-effector pose onto the camera image and comparing the distances of the detected objects to the tool projection (Algorithm 1). Once the closest objects for each hand have been found, the database is queried for the appropriate set of functional affordances. These are presented in a tree structure to the user, who chooses the desired functionality through the interface. The offered options also include information regarding the corresponding grasp type so the user can account for it when deciding on the angle of hand approach. For example, if the robot recognises an open water bottle, the bottle-cap and a glass, it might provide the functionality affordances of "drink", "pour" or "close" to the user, as well as their corresponding grasp types. If the user chooses to "close", the framework selects the power grasp to hold the bottle with the left arm, while the pinch grasp is proposed for the right arm picking up the cap.

Algorithm 1: Affordance analysis: Linking detected objects to hands

```plaintext
objects = FindObjects(video_stream, threshold);
for hand in [left_hand, right_hand] do
    hand_pos = ProjectToolToImage(tool_pose);
    closest = objects[0];
    d_min = Distance(closest, hand_pos);
    for object in objects do
        d = Distance(object, hand_pos);
        if d < d_min then
            closest = object;
            d_min = d;
    end
    affordances = Map(closest, database[hand]);
    Display(affordances, video_stream);
end
```

C. Experiments

In the first stage, the pose tracking capabilities of the developed framework were investigated by recording and examining the desired and actual pose of the robot end-effector. The operator first performed a one-dimensional motion with varying frequency which was used to characterise system lag. This was followed by arbitrary 2D and 3D motion which was examined in terms of tracking error. The second experiment was aimed at testing the affordance analysis and object-to-hand linking. A selection of objects from the set presented in Table I was placed on the table, in clear view of the camera. One of the hands was then hovered over each of the objects while observing its matching affordance list. Through this, the object recognition, robot tool-to-image projection and affordance linking was verified.

The final set of experiments utilised the entire framework in a remote manipulation setting. The VR system and visual interface were set up in a remote location, streaming the VR controller data and receiving the video feed. The operator was in control of both robot arms and hands, monitoring them on the feedback screen (Fig. 3). Two tasks were performed in this context:

1) Pick and place: A ball was placed at an arbitrary position on the table and the operator was asked to pick it up choosing a grasp from the different available affordance options.
2) Pouring: A bottle filled with small loose components and a cup were placed in the workspace and the
Fig. 4. Tracking of the controller motion by the UR robot. Subfigure A shows the target and actual position for one-dimensional periodic motion. Subfigure B shows the tracking for a two-dimensional trajectory, while subfigure C shows the trajectory tracking in a three-dimensional motion.

Fig. 5. Linking object affordances to current robot end-effector pose. Robot tool pose is projected to the camera image and the affordances of the closest detected object are presented in the selection menu.

The clip highlights the advantages of using an assistive, affordance based telemanipulation framework for simple bimanual tasks. The tasks were successfully completed with minimal effort from the operator’s side.

V. CONCLUSION

This paper presented an affordances oriented, assistive telemanipulation framework implemented on a dual robot arm hand system. A commercially available VR system was used for guiding the arms and interfacing with the affordance selection menu which provided grasping and manipulation assistance. The framework was successfully tested in a remote manipulation setting, where a number of bimanual tasks were executed by the operator.

Regarding future work, several aspects of the concept can be expanded and improved. Currently, the system relies on a manually defined, static affordance database and cannot provide the affordances of completely new objects. This could be solved by implementing a scalable knowledge base that employs a set of inference rules to estimate affordances of previously unseen objects. System control could be improved by including assistance in the form of local motion correction for object grasping, effectively shrinking the dimensionality of the user control space. The user experience could also be enhanced by enforcing anthropomorphism of the robot.
motion and by embedding the control interface into the video stream using augmented reality approaches. Further work should also be invested into system validation and comparison with alternative solutions from an ergonomic viewpoint. A user experience survey or an appropriate benchmark would provide valuable information regarding the system’s ease of use and its intuitiveness of operation.

REFERENCES


