AN ABSTRACT OF THE THESIS OF

Sai Krishna Allani for the degree of Master of Science in Robotics presented on August 31, 2016.

Title: Analyzing Human Gaze Patterns During Grasping Tasks To Advance Robotic Grasping.

Abstract approved: ____

Ravi Balasubramanian

There is a strong interest in the robotics community in learning how humans grasp and manipulate objects, partly because robots need to operate in human environments and partly because humans are currently much better in physical interaction tasks than robots. This thesis seeks to identify the human heuristics for grasping by analyzing human gaze patterns when humans perform grasping tasks using their own hand and when they use a robotic hand. This thesis uses a human-subject experiment to analyze the participant's eye-gaze for finding what features people think are important for grasping objects. The features included where the fingertips settle down on the object relative to the object's edges, center of mass, etc. It was found that while gaze patterns on the objects are similar whether the human used the robot hand or the human hand, participants spent substantially more time gazing at the robotic hand then their own, particularly the wrist and finger positions. In a subsequent study, it was also shown that choosing camera angles that clearly display the features participants are interested in enables the participants to more effectively determine the effectiveness of a grasp from images. This thesis's findings are relevant both for robotic grasp planning algorithms (where visual cues are important for analyzing objects for potential grasps) and for designing tele-operation interfaces (how best to present the visual data to the remote operator).

Analyzing Human Gaze Patterns During Grasping Tasks To Advance Robotic Grasping

by
Sai Krishna Allani

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented August 31,2016
Commencement June 2017

Master of Science thesis of Sai Krishna Allani presented on August 31,2016

APPROVED:

Major Professor, representing Robotics

Head of the School of Mechanical, Industrial, and Manufacturing Engineering

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Sai Krishna Allani, Author

ACKNOWLEDGEMENTS

The author thanks everyone involved in this research and who has helped me in successful completion of this degree. First and foremost, I would like to express my deep gratitude to my advisor, Dr. Ravi Balasubramanian who has accepted me as one of his graduate student, giving me a platform to hone my skills and also provided the opportunity to participant in DARPA Robotics Challenge. I thank him for guiding me, motiving me and always willing to help and spending time with his students. In this two years of graduate degree I have learned a lot as his student. I also thank Dr. Cindy Grimm for her help and guidance which has helped me a lot in coming up with hypothesis and in designing surveys which has finally led to completion of my thesis.

I have to acknowledge all my colleagues in Robotics and Human Control Systems Lab for their assistance in data capture and in many other aspects. They are Saurabh Dixit and Jackson Carter. Thanks to all my friends and family for their support and encouragement.

TABLE OF CONTENTS

Page

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF TABLES

Chapter 1 Introduction

As the domain of robotics advances, people have started using robots in various domains ranging from factory automation to surgical robots. In factories, robots are being used in bin picking and for doing repetitive tasks. When it comes to the medical field, sophisticated robots are being used to do complex surgeries where surgeons can't reach. Robots are used in disaster and rescue situations and also in defense where it is dangerous for humans to work. Robotics-inspired prostheses are used for assisting physically disabled people. Across all these fields, to complete the task with success robot must be able to grasp and manipulate objects in the environment successfully. However, this has proven to be a hard task. Research shows that one in every four grasps fail even under structured environment. So there is a huge need for improvement in robotic grasping.

It is also expected that humans and robots will work together to perform complex physical interaction tasks in various environments. A wide spectrum of possible human-robot interaction scenarios exists to enable this (see Figure.1.1). At one end of the spectrum, the human and robot are physically separate from each other. In this scenario the human sends remote commands to the robot and the robot provides visual feedback. These kinds of robots are used in disaster rescue scenarios, where it is unsafe for humans to enter[1], [2]. At the other end of the spectrum, the human and robot are in contact physically. For example, a neuroprosthetic hand attached to the human provides touch and force information directly to the human (either directly through physical contact or through the peripheral nervous system) and in turn receives commands through the neural system[3]. Across this spectrum, it is important to understand how humans process both visual and touch information when performing interaction tasks, such as manipulating objects in the environment [4].

The information available to the human when teleoperating a robot to perform a physical interaction task is diverse. It could include direct 3D views, 3D point clouds provided by a laser scanner, 2D video or images provided by a still-camera or video feed, and tactile information provided through sensors attached to the robot. Prior work has been done in how to present the information to operators to get the quickest response time as well as the best decision from the operator[1], [5]–[9]. But very little work has been done in understanding (1) how the visual information provided to the operator is processed when performing physical interaction tasks such as grasping, and (2) how humans might compensate for missing tactile cues using visual ones.

Understanding operator gaze is important in a grasping task because it provides information about how humans perceives the object and the task environment, and what visual cues are important in completing that task. Specifically, eye gaze provides information about how important different features or descriptors of the object — such as the object's silhouette, surface, center of mass, and center line — are when performing the task with either a robot hand or a human hand. More subtly, *changes* in visual cues between using their own hand and the robotic one provides information about which features of a hand's position, such as finger location and wrist orientation, humans use when performing a task. For example, humans rarely look at their own hand when grasping an object.

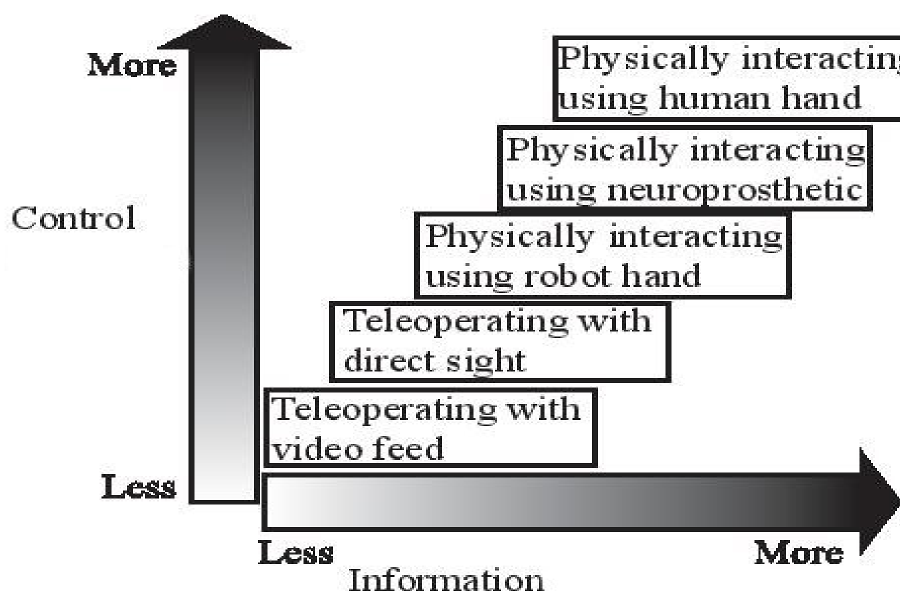


Figure 1.1 Human-robot interaction possibilities

In the long term, this information is useful for both autonomous robots and teleoperated robots. Specifically, when developing control algorithms for robot hands, it is useful to know which features to focus on in order to prune the decision space of where the robot hand should grasp the object. For tele-operation environments, knowing which visual cues are important to the human enables the robot to determine which views and data to send to the remote operator, as well as improve the design of the human-robot interface to simplify the manipulation task.

This work answers two important questions (1) what features people think are *important* for grasping, and (2) how effectively can humans use these features in robot grasp images to determine grasp quality. In order to answer these questions, we explore differences in eye gaze patterns between two different points on the manipulation spectrum. In this study we use the human hand as an example of an “ideal robotic tool”, where the human has the best information about the object and the manipulator and best control over the manipulator as well. As a counter example of tele-operation, we use the human physically positioning a robotic hand. Here, the operator has full, natural control, and complete visuals. We analyzed the eye gaze data difference between the two conditions in three different stages of the manipulation task: pre-grasp, during manipulation, and post-manipulation grasp evaluation. Differences between eye gaze patterns can help distinguish extrinsic (visual) cues from intrinsic (touch, proprioception) ones, and more importantly, which visual cues humans use as substitutes for missing intrinsic cues. We use this information to design a camera view-point algorithm[10] and a scoring metric for view-points, which uses our identified features to determine the best view for grasp evaluation from images.

Chapter 2 explores prior work done related to robotic grasping, analyzing human eye-gaze patterns and optimizing camera viewpoints. Chapter 3 explains physical study protocol and how we analyzed eye gaze data. Chapter 4 presents the results from eye-gaze data analysis. Chapter 5 explains how on-line study is conducted and presents the results from the on-line study. Chapter 6 provides conclusion for this work.

Chapter 2 Background

2.1 Robotic Grasping

The domain of robotic grasping and manipulation has seen significant progress in terms of hardware [11],[12],[13] and software development [14],[15],[16],[17]. However, there is a strong need to improve the ability of robots to robustly physically interact with the environment. Specifically, prior work has shown that even in a laboratory environment with almost perfect information for grasp planning, robotic grasping performance only succeeds about 75% of the time, i.e., one in four grasps fail [18]. The primary reason for this poor performance is that even small differences in object shape or object position cause the object to, for example, slip out during the grasping process. There have been significant efforts with mixed success to address these issues using physics-based heuristics and brute-force search algorithms to find more robust grasps[18],[19],[20],[21].

Prior work has explored "learning from demonstration", where humans teach robots [22],[23],[24]to advance robot performance. However, most previous approaches for gathering data are time-intensive [25]. Prior work has explored crowd-sourcing where they employed images or video of the grasps to receive human input [26]. That work showed that humans are likely to over-estimate how successful the grasp will be. Despite this over-estimate, humans are still more accurate than learning approaches that use heuristics such as center of grasp, center of mass for certain subsets of grasp types. Other work in the context of learning from demonstration also revealed a novel heuristic that humans use for improving grasp quality, namely, "skewness" where the human aligns the robot's wrist to the object's principal axis [18]. Other studies also showed that integrating human demonstrated grasps with existing grasp planning softwares tends to increase the performance of grasping[27]. It also showed that researchers can learn the placement of thumb and index fingers from human grasps which can be used to replicate more human-like grasps in robotics .

John D.Sweeney and Rod Grupen [28] used human demonstration techniques to learn grasp affordance preshapes —the pose of the hand and fingers relative to the object just prior to initiating a grasping action—from demonstration data, and use those affordances to generate preshape hypotheses for novel objects based on visual appearance. A grasp affordances [29] is a way of grasping an object to achieve a particular function, i.e., reaching a coffee cup to drink and reaching a coffee cup to transport. This research showed improved grasping performance on novel objects. Hyun et.al [30] has used appearance-based visual cues to improve prediction of grasp affordances.

Research has been done in improving the performance of robotic grasping by using image based visual servoing (IBVS) technique in the shared autonomy grasping system, in which visual input is given to control a dynamic system [31] . In this method, visual input plays a vital role in deciding where the finger lands on the object, so providing a clear visual input with few occlusions and more information about the scene is important. Research has also shown that using predictive displays where there is a huge delay in sending and receiving commands can improve performance of grasping [32]. By using predictive displays operators can "lead" the task and take larger steps with confidence, while performing a teleoperation task. In order for a grasp to be successful, predictive displays should show where the fingers will land on the object in the near future with high accuracy.

Human-in-the loop is another technique researchers have explored where they combine human cognitive skills with autonomous tools to increase the performance of grasping [33]. Human-in-the loop (HitL) has shown that subjects grasped more number of objects with less collisions in cluttered environment when there is more level of autonomy on the robot which decreases effort on human operator. This research also pointed out that operators have concerns with comfort level with GUI and other properties such as positioning a virtual camera while performing a manipulation task. One way of unloading effort on humans performing the task is to automate the camera positioning such that it shows what the operator wants to see to perform the task. Paul Michelman and Peter Allen[34] has used shared autonomy in hand teleoperation system. This research showed that the teleoperation tasks are completed 50% more quicker when force sensations are used.

Researchers have studied gaze patterns for evaluating static images of grasps [35], which showed that participants use many of the same cues as they do for grasping, and that participants are similarly likely to overestimate the effectiveness of grasps that look “human”. However, no prior work has studied human eye gaze in 3D when controlling a robot arm in a physical interaction task.

The teleoperation literature has shown that many factors can affect the performance of a remote teleoperator like viewpoint, perception and time delay. But there is little to no work done in understanding how humans visually process the task environment while performing a task with a physical robot or how they perceive views shown on a user interface while teleoperating. Most research has shown that operators have concerns with how the data is presented to them and in understanding the presented data. Research showed that many operators spend a lot of time rotating the camera angle in the user interface in order to find the best view to understand the scene[33]. In order to take the load off of the operator and to get a quick response from them, we have to present data to the operator which shows useful information. One way of doing this is first to understand what features or descriptors people think are important when performing a manipulation task. In order to find these features, we analyzed eye gaze data of humans when they were performing a manipulation task. Our research will help in improving how the data is presented to the teleoperator, visualizing contact points on user interface and in pruning the decision space. All these factors will help in getting a quick response from the operator.

2.2 Analysis of eye gaze data

There is a growing body of prior work on where humans look when performing grasps using their own hands [36],[37],[38]. This work showed that people's gaze patterns are a mix of tracking the object's center of mass, looking at the top of the object, and looking at where the forefinger will make contact with the object (which in their case was the top of the object). Varying the task [38] or asking the participants to do the grasp from memory [37] changed the ratios of which regions were gazed at, and in what order, but did not substantially change the types of regions. In [35] we see that these same patterns hold for the robotic grasping task, but that participants also spend substantial time looking at the fingers, wrist, and other contact points.

2.3 View-point selection

Viewpoint selection is a very hard task for two main reasons: first, there are many factors which must be taken into account to generate a good view, ranging from perceptual issues to geometric analysis of the objects in the scene; second, the "best" view heavily depends on what we want to see, and thus, it is both a use and application-dependent task.

There is very little work done in identifying what features are important in an object to represent them in the best possible way, so that useful information is captured in that view [39],[40]. There are several existing techniques for camera viewpoint selection [10], [41] which evaluate a large number of visually salient features and artistic guidelines. Most of the features which they evaluate are domain specific. Placement of camera is very important in order to see features without any occlusions. In GUI's, effective camera placement and control is fundamental for the user to understand the virtual environment and be able to effectively accomplish the intended task. Ineffective camera placements could cause the user to miss important visual details and thus make incorrect assumptions of the environment while performing the task. In most of the currently used GUI's, users directly position the camera using an input device through a tedious and time-consuming process requiring a succession of "place the camera" and "check the result" operations. In recent years, researchers have come up with methods to automatically position the camera so that it shows useful information with less occlusions[42],[43],[44]. These new methods relieve the user from direct control. Camera placement optimization is done by considering multiple features of the input image. Most of these features depends on the domain of the problem, so it is crucial to identify which features are important in a specific domain.

To present the data in best possible way, researcher have also explored how to calculate the view-point quality and describe how this information can be used to get the quick response from the operator. Computation of good view-points is important in several fields: computation geometry, visual servoing, robot motion, etc. There is no consensus about what a good view means. In many areas the quality of a view-point is intuitively related to how much information it gives us about the scene. A good view must help us to understand as much as possible about the object or scene represented. Bourque and Dudek [45] define an interesting point in an image as the one different from the surrounding context. These regions are the ones on which the human attention would focus. The features which are used to calculate view-point quality are domain dependent, i.e., in games, features are showing characters without occlusions, while in grasping, features are showing objects and fingertips without occlusions in cluttered environment.

Researchers have come up with new metrics to calculate quality of view-points so that they can select good views. Kamada and Kawai [46] have proposed a method to compute a point of view, which minimizes the number of degenerated edges of a scene. They consider a viewing direction to be good if parallel line segments are as far away one from another as possible at the screen. This means minimizing an angle between a direction of view and a normal of a considered face. Or, for a complex scene, this means minimization of the maximum angle deviation for all faces of the scene.

Colin [47] has proposed a method, initially developed for scenes modeled by octrees. The method is to compute a good viewpoint for an octree. The main principle of the method can be described as follows:

1) Choose three best directions of view d_1 , d_2 and d_3 among the 6 directions corresponding to 3 coordinate axes passing through the center of the scene.

2) Compute a good direction in the pyramid defined by the 3 chosen directions, taking into account an importance of each of the chosen directions. The view-point is considered to be good if it shows high number of voxels.

Plemenos and Benayada [43] have proposed a heuristic that extends the definition given by Kamada and Kawai. Their heuristic considers a viewpoint to be good if it minimizes maximum angle deviation between direction of view and normals to the faces and gives a high amount of details. Vazquez et.al[49],[50],[51],[52],[53], used viewpoint entropy/Shannon's entropy as a metric or information measure to obtain good views. This entropy considers both the projected area and the number of visible faces from the 3D model normal dispersion. Satoshi et.al [54] used novel information quantity of Fencher type based on Fencher's law in psychophysics to select good views. Dmitry Sokolov and Dimitri Plemenos [55] has used curvature of a surface as a metric to choose good views.

Francisco et.al [56] have used ambient occlusion technique to choose good views. Thales et.al [57] has used machine learning algorithms to learn good views using intelligent galleries.

Contributions

As mentioned in the previous sections, there is little to no research done in understanding human gaze in 3D. Most research in the context of teleoperating robots for grasping tasks has shown that teleoperators have concerns with the comfort level of virtual interface. Most of the existing grasping softwares takes considerable amount of time to come up with a grasp which might not work if there is any change in the position of the object. One way to improve is to prune the decision space. Our research will fill the gap in robotic grasping by understanding how humans perceive task environment while performing the grasping tasks and what features humans think are important for grasping. We will also focus on how this information can be leveraged to present visual data in best possible way to a teleoperator.

Our contribution from this research is a method for picking important features and providing a quality metric to choose good views for the more specific task of grasp evaluation. Also we have conducted an on line study to verify how effective are the features we chose in determining the quality of grasp.

Chapter 3 Experimental Methods

To understand eye gaze patterns better we decided to run a user study where we collect eye gaze data of participants while they are performing the task. In this chapter we will explain how experimental study is conducted, what objects were used in the study, what tasks were performed during the experimental study and how data is collected and analyzed. In high level, participants were asked to lift the object from the table and also perform an object-specific task. They did both these tasks with their own hand and a robotic hand. The order of the two phases were randomized to prevent learning effects. The participants' eye-gaze was tracked throughout the task, and they were also asked to "think aloud" — verbally evaluating their choice of grasp.

3.1 Objects and tasks

We used 15 everyday objects for the study. A photo of the objects is show in the Figure 3.1. For each object, participants were asked to perform two tasks. The first task was to pick up the object from the table and the second task was object-specific (see Table 3.1). The object-specific tasks are tasks or actions that are associated naturally with each object, such as pouring water out of a jug, throwing a ball, or pressing a button on the remote.

3.2 Phases

Human grasping and human-driven grasping were captured during our study. The study features a training phase and two distinct capture phases: in the first capture phase the participants use their own hands to grab an object, while in the second, the participants physically position the robotic arm and hand to grasp the object (order randomized). Data was not captured during training phase. In this phase, participants were asked to familiarize themselves with the hand by moving it around adjusting the fingers. Although there was a gravity compensation mode for the arm, it did not adjust well when the hand was opened and closed, so participants were also given instructions to ask for help in supporting the hand if needed.

When participants were doing the task with their own hand (human hand), they were asked to use only three fingers, i.e., thumb and first two fingers, to replicate the three fingers of the robotic hand.

Figure 3.1 Objects used in the study

3.3 Prompts and think-aloud

We asked them to think aloud to provide insight into what they were thinking of while performing the grasping tasks. For the move-the-object task, participants were asked to actually move the object using the robotic hand and arm. For the other tasks, they were not required to perform the task, but simply needed to position the hand. They were given explicit permission to pick up the object, position it how they wanted, and use their other hand if needed. At the end of each grasp, participants were asked: “Is this grasp exactly what you wanted? Or are the finger placements slightly different than what you were intending? (How so?)". This prompt is aimed at determining how much the robotic hand limitations affected the participants' grasp choice. Also it has helped us in interpreting the results.

Table 3.1 Object-specific tasks and number of grasps captured (including pick-up task).

Object	**Natural Task**	**Total Grasps**
Water Pitcher	Pour water out of pitcher	11
Spray Bottle	Pull trigger to spray	14
Margarita Glass	Drink out of glass	14
Cereal Box	Pour cereal out of box	12
Cracker Box	Pour crackers out of box	15
Television Remote	Press power button on remote	11
Toy Plane	Pretend to fly plane around	13
Food Clip	Open clip as if using it to close bag	10
Soap Dispenser	Press down on nozzle to dispense soap	10
Foam Cylinder	Throw object overhand	16
Bison Plush Toy	Hand toy to someone	5
Plush Ball	Throw ball underhand	19
Sock Doll	Hand doll to someone	16
Decorative Cord	Hand cord by its metal ring	5
Tape Roll	Support tape roll so that another hand can be used to rip tape off	11
Total Grasps/object		182 12.1

3.4 Data Capture: equipment and procedure

The equipment used for this study included a pair of SMI Eye Tracking Glasses 2.0 to collect eye-gaze data and a Barrett WAM Arm with BH280 BarrettHand to perform the robotic grasping. We also instrumented the working space with spatial calibration patterns (see Figure. 3.2) and the capture procedure to ensure calibration between data sources (eye-gaze, Kinect sensor, and BarrettHand).

3.4.1 Eye tracking

The SMI glasses recorded both where the user was looking and what they were looking at. The data was recorded as a 960*720 video stream at 30 Hz, plus an eye gaze location for each video frame (as x, y image coordinates). The eye gaze data also included other information such as pupil diameter, fixations, and saccades. The eye tracker had to be fit to the person's head (similar to goggles) using two nose pieces and calibrated to their eyes. To perform the calibration, the participant was asked to sit down in front of the table and fixate on a red dot on the table (see Figure.3.2). This one-point calibration was performed using SMI software. We checked the calibration at the end of each grasp trail by having the participant focus on the red dot again.

3.4.2 Arm and hand tracking

We used a Barrett WAM and Barrett Hand (BH-280) in the study. The arm is back drivable and gravity compensated; that is, the arm location can be physically adjusted with ease. However, the BarrettHand's fingers cannot be physically adjusted from external forces; it can only be done through motors. We used a physical set of three sliders to control how much each finger was closed, and a knob to control the spread of the fingers. Note that the two joints of each finger are controlled with one actuator.

3.4.3 Audio and temporal alignment

The eye-tracker recorded audio with the video. In addition to recording what the participant said, we also used this information to temporally align the eye-tracker to the arm data streams using a generated beep. All other data alignment was through the Robot Operating System(ROS) toolkit.

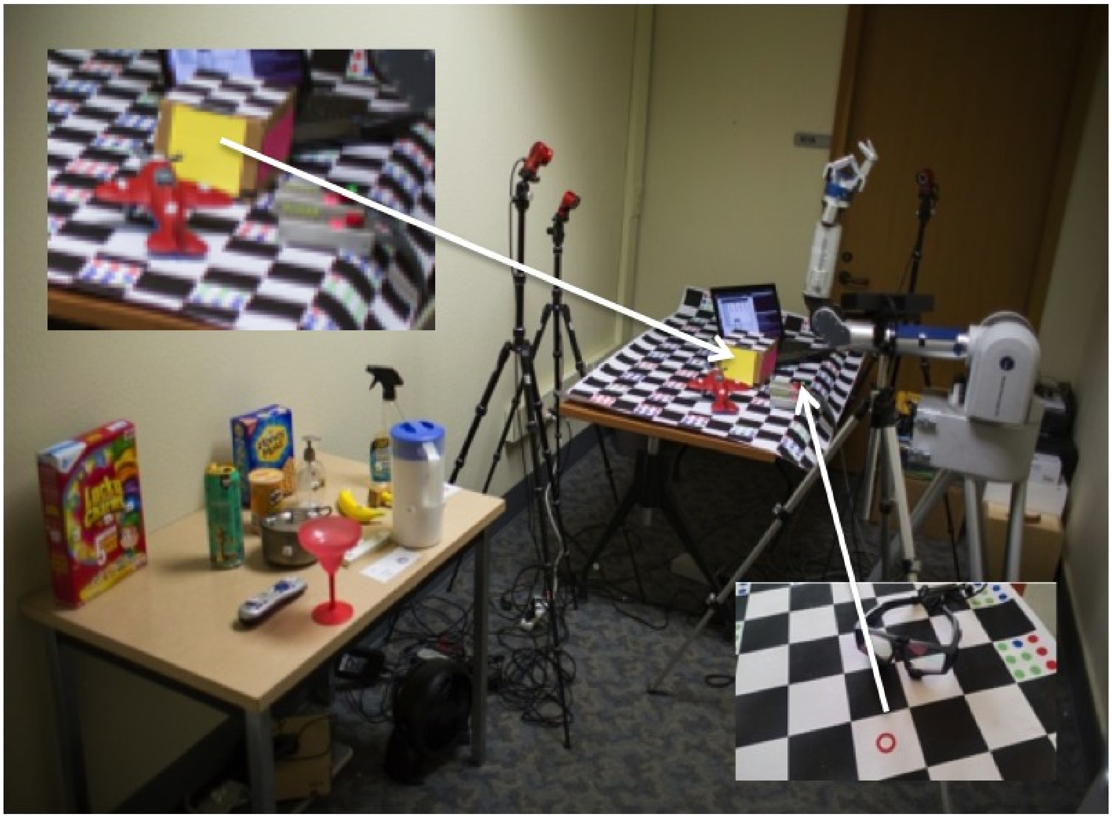


Figure 3.2 Study set up. The table included a checkerboard pattern for further calibration. The red circle was used to calibrate the eye tracker. The box on the table was used in the object placement tasks. Participants were seated at the table.

3.5 Protocol management and flow

The study is designed to be run by two researchers. One researcher handled the Ubuntu Linux PC running ROS and the arm, the other handled the SMI eye tracking laptop. Both researchers were involved in explaining the study and talking to the participant.

The average time for a data collection session was an hour and a half, covering two grasps each for three or four objects. The maximum time was restricted to two hours due to eye strain generated by the eye tracking glasses, as well as general fatigue from performing the experiment. New participants went through a training session to familiarize themselves with the robot arm and hand before starting data collection.

The general flow of the study can be seen in Figure 3.3 and is also outlined in the list below.

1. Subject enters room and signs consent form.
2. Brief training session with a test object.
3. Eye tracking calibration performed.
4. Study tasks explained to participant.
    (a) Object placed on table, and participant told to use robot hand or their own hand (order randomized) to perform pick up task.
        i. Pick up task performed.
        ii. Repeat until no new grasps.
    (b) Natural task explained
        i. natural task performed.
        ii. repeat until no new grasps.
    (c) Object-tasks repeated with human hand or robot hand
    (d) Eye tracking recording stopped, and re-calibration is done to check if there is any drift in eye-gaze.
5. Repeat a-d with as many objects as possible
6. Eye tracking recording stopped, all other data collection ended.

3.6 Participants

We recruited 13 participants, ranging in age from 16 to late 50's, all with normal or corrected to normal with contacts vision. It is not possible to wear regular eye glasses with SMI eye-gaze glasses. On average participants specified 4.5 (maximum 8) grasps per object across the two tasks.

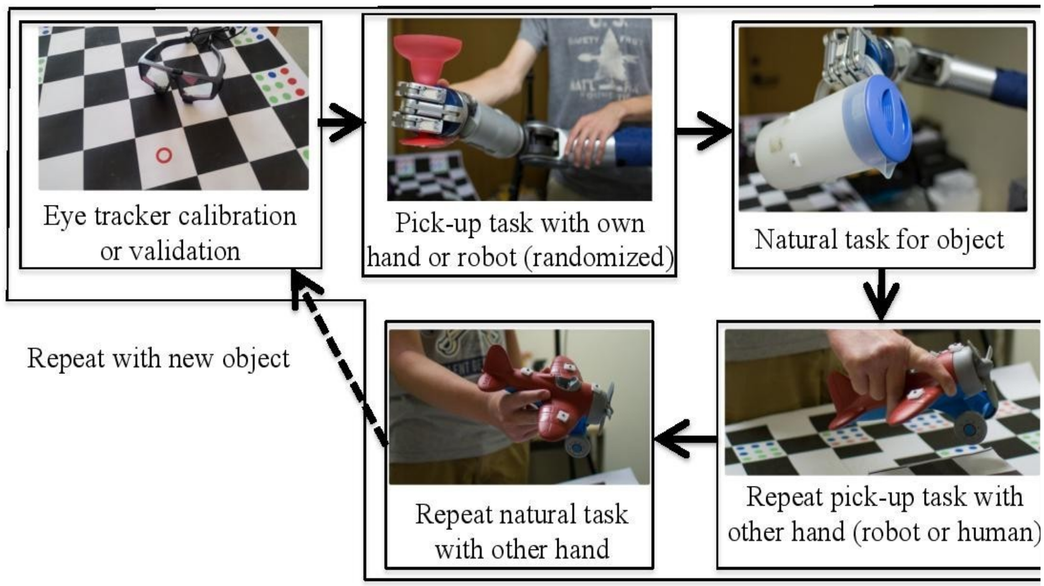


Figure 3.3 Flow chart of the study procedure

3.7 Analysis of eye-tracking data

We performed two types of analysis on the eye-gaze data. The first analysis focused on labeling what the participants were looking at before, during and after grasping the object to perform the manipulation task (see Figure 3.4) . The second analysis focused on differences in fixation points for different objects. Example frames from the video are show in Figure 3.5.

3.7.1 What did they look at?

We annotated the eye-gaze data using a two-step process. In the first step, we annotated the video into three phases: Pre-grasp (when the participant is looking at the object but has yet to touch it), during-grasp (when the participant closes the hand around the object) and post-grasp evaluation (when the participant evaluates the grasp in the think-aloud protocol) (see Figure 3.4). While manually annotating through the videos we found similar patterns in all the videos that the subjects looked at a particular set of features while performing the task with their own hand and robotic hand. The features we found that subjects are focusing on are same as the features found in other research [35]. So we decided to find how much time participants spent looking at each of these features while performing the task. So in the second step we labeled what features the participants were looking at. The view is split into five regions, three of which focuses on the object and two on the hand (see Table 3.3). Apart from these features subjects also looked at the control box which is used to control finger joints of the robotic hand while performing manipulation task.

We annotated the video data and produced the statistics using MaxQDA [max 1989-2015] (see Figure.3.6). To verify inter-coder reliability, we had a second coder repeat the coding for 1 participant; the code alignment was within 5%. For this analysis, the gaze points outside of the object and hand were ignored.

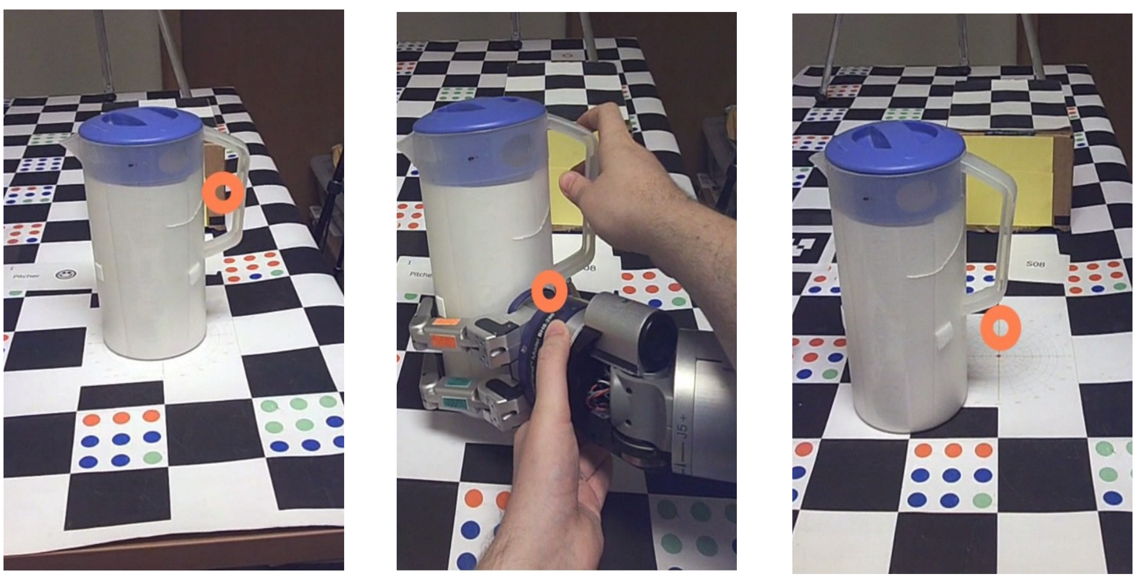


Figure 3.4 Three stage of grasping. Left to Right: before, during and after

Not all participants had all codes, most notably, very few participants had a pre-grasp gaze for the robotic hand, and there were also 2 participants who had no pre-grasp gaze for the human hand. We hypothesize two reasons for this:

1. Peripheral vision was sufficient in some cases for the participant to categorize the object.

2. If the participants were doing the robotic hand second, they had no need to look at the object again (10 of 11 subjects with no pre-grasp gaze).

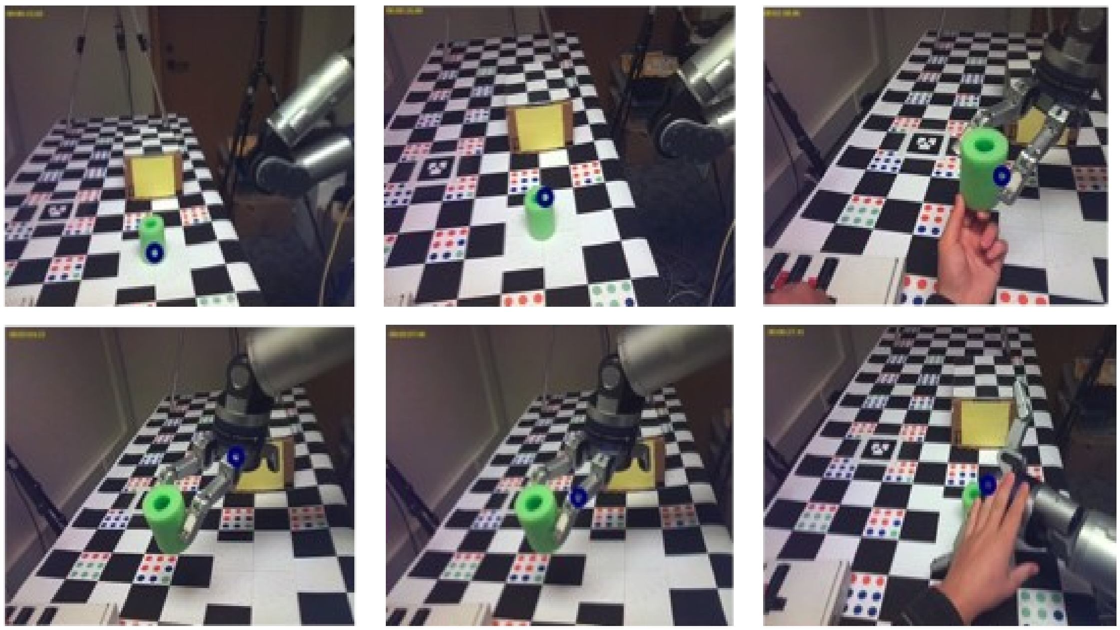


Figure 3.5 Example frames from the eye-tracking video showing the different eye gaze locations (center, top, side, finger, wrist).

3.7.2 Data analysis of eye-gaze data

Once eye-gaze data is annotated using MaxQDA we have performed data analysis to show that the data is statistically significant.

As described above, in the first stage of annotation we annotated the video into three different stages. Once we have completed annotating all the videos, we measured how much percentage of total time is spent in each of these stages for each participant. After that we did a two sample T-test at 5% significance level between all the three stage across both condition, i.e., human hand and robot hand. The box plot for the percentage of time spent in each stage in two different scenarios is show in Figure 4.2

In second stage of annotation, we focused on annotating what features participants were looking at in each of the stages. We annotated the features participant is looking at if the eye gaze was moving/constant (700ms - 3sec) at a certain region. Once second stage of annotation is done, we did a two sample T-test at 5% significance level between all the features across three stage for all participants. Before performing T-test we did feature scaling to bring data between the range 0 and 1. Feature scaling is done by using the formula below:

$$X' = (X - X_{min}) / (X_{max} - X_{min})$$

Following steps are performed on eye-gaze data:

1. Careful annotation of eye-gaze data into three stage, i.e., before, during and after.
2. Statistical analysis is performed.
3. While annotating we found similar patterns across all the videos that is, participants are focusing on certain features on the object and hand while performing the task. These features coincide with features mentioned in other research [35],[37],[36]. So we decided to find percentage of time spent on each feature at each stage.
4. Careful annotation of eye-gaze data to identify features.
5. Normalization of data between range 0 and 1.
6. Statistical analysis is performed.

Pseudo code for T-test:

```
H =[];
P =[];
temp_vec_h= [];
temp_vec_p = [];
for i = 1:size(data,2)
        temp_vec_h = [];
```

```
        temp_vec_p = [];
        for j = 1:size(data,2)
                a_mat = data(:,i);
                b_mat = data(:,j);
                [h,p] = ttest(a_mat,b_mat);
                temp_vec_h = [temp_vec_h, h];
                temp_vec_p = [temp_vec_p, p];
        end
        H = [H; temp_vec_h];
        P = [P; temp_vec_p];
end
```

where,

temp_vec_h = temporary list of h values of ttest

temp_vec_p = temporary list of p values of ttest

data = normalized time spent looking at each of the five features in three different stages across all participants

H = final list of h values from ttest

P = final list of p values from ttest

A subset of P-values for all the combinations between the data is shown in the Table 3.2 (see Appendix A.3 for complete set of P-values) . A box plot which shows normalized time spent on each feature at each stage is show in the Figure 4.1.

Table 3.2 P-values

	before_top	**during_top**	**after_top**	**before_wrist**
before_top	1	0.398038	0.333148	**0.01855**
during_top	0.398038	1	0.85671	**2.6E-006**
after_top	0.333148	0.85671	1	**3.5E-006**
before_wrist	**0.01855**	**2.6E-006**	**3.5E-006**	1

3.7.3 Fixations

While annotating the eye-gaze data we found another pattern in the data. We saw that subjects tend to focus on multiple locations on objects while performing the task. These locations or points on complex objects are more when compared to simple objects. In order to find these points, we implemented the EyeMMV fixation detection algorithm [58], which filters the coordinate sequences by applying a threshold of dispersion to the points. We used standard settings [59] for the algorithm: a 90 ms minimum fixation duration and a maximum fixation dispersion of 0.5 degree of visual angle (DVA), with a preliminary filter of 5 pixels greater than ½ DVA. We measured visual angle based on gaze frames where the participant was focused on the object. The algorithm produces a sequence of fixations, with each fixation centered at the average of the coordinates and lasting a given duration. Using fixations, over the raw coordinate data, both reduces processing time and removes saccades, where the viewer is essentially blind. We overlay the fixation counts with the annotations to produce the number of fixations on the object during each phase.

Table 3.3 Annotation codes

Regions	Codes
Regions (object)	Centerline Top Edges
Regions (hand)	Wrist Finger tips

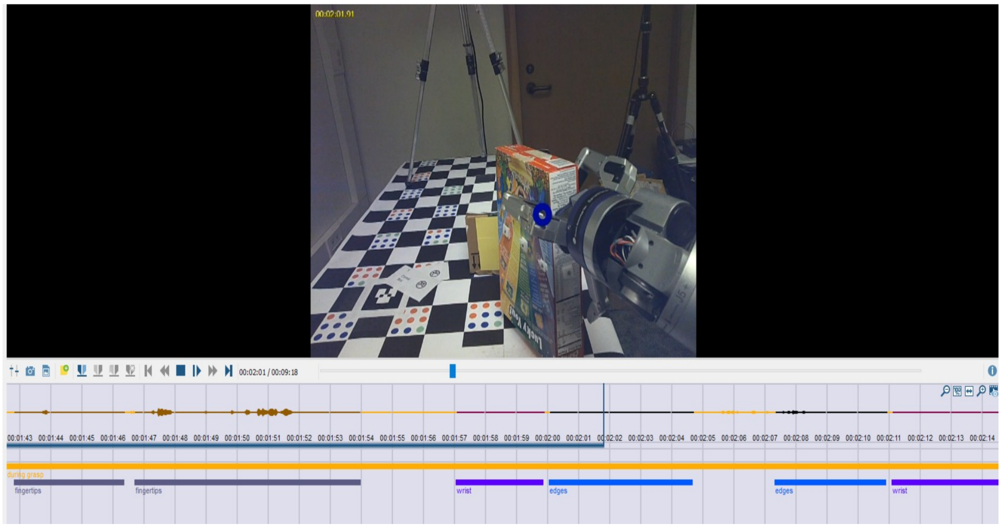


Figure 3.6 Annotations for the videos using MaxQDA

3.7.4 Think-aloud evaluations

We did not formally analyze the participant's comments; we summarize here overall statements. Around half of the participants said at least one grasp was not quite what they wanted, particularly for more complex objects such as the plane. The major refrain was that the participants didn't like that the joints in the fingers couldn't be controlled individually (the Barrett hand only supports bending the finger not controlling each joint independently). This was most noticeable in cases where the finger locks up due to collision with the object — one part of the finger comes in contact and stops, while the remaining part of the finger stops where it is and doesn't close all the way around the object. Other issues were the fingers being too thick, the hand too big, or the controls being too fidgety to achieve some of the more precise grasps the participants had intended to perform.

Figure 3.7 Images showing fixation points on different locations on objects and hand while performing the task

Chapter 4 Results from User Study

Here we summarize the eye-gaze difference between the two conditions (human hand versus robot hand), and the differences in fixations between objects. T-tests are performed on eye-gaze data at the significance level of $\alpha = 0.05$ between the different conditions.

4.1 Gaze difference

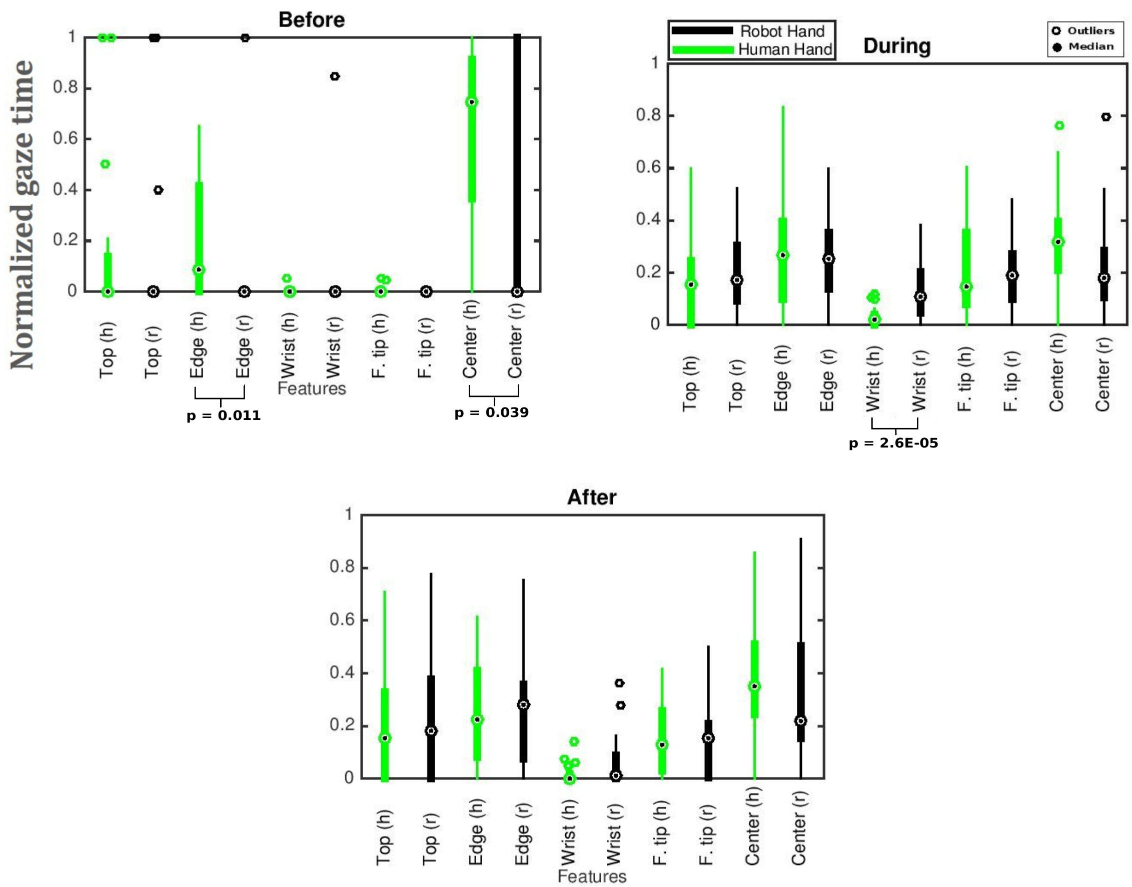


Figure 4.1 Gaze time spent on each feature while performing task with the human hand and the robot hand in three phases. From left to right: Before, during and after phases. Gaze time is normalized across all three phases.

Figure 4.1 shows the normalized gaze times for both using the robot hand and the human hand in the three phases of grasping (before, during, and after). Several patterns are clear. In the "before" phase when using the robot hand, the human subject almost never looks at the object's top and edges or the robot's wrist or fingertips. They only focus on the object's centerline. This is in contrast to using their own hand, where the focus is primarily on the top and edges of the object, as found in previous studies[35].

In the “during” and “evaluation” phases, the two gaze patterns were more similar. Primary differences are that the participants spent more time observing the robot's wrist (versus looking at their own) and less time looking at the object's centerline.

Overall, during the robotic grasp task participants spent significantly less time looking at the edges and top of the object before beginning the manipulation (some participants barely glance at the object before starting — see Figure 4.2). In human grasping studies, gaze on the edges and top corresponds to participants determining potential contact points for their fingers. We hypothesize that the lack of these pre-grasp glances for placing the robotic hand implies that visualizing contact points is part of the *control* strategy for guiding the robot hand to the desired grasp, but not for *planning* the grasp in the first place.

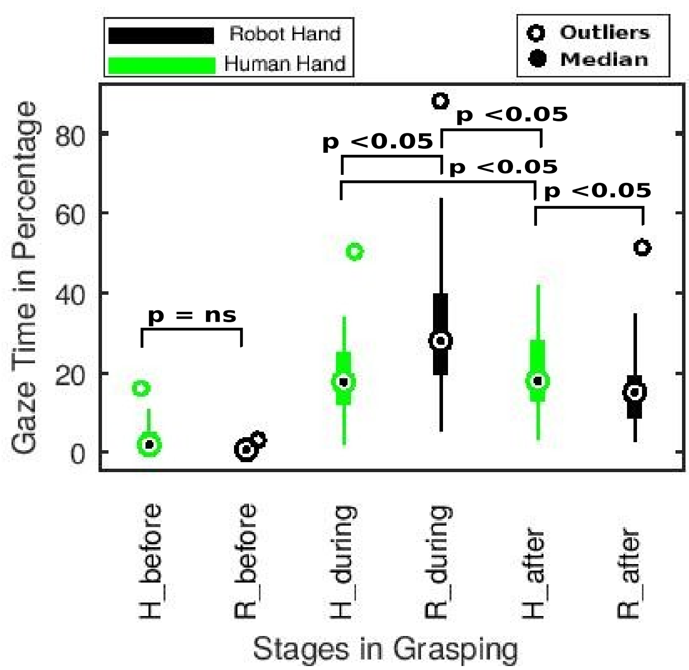
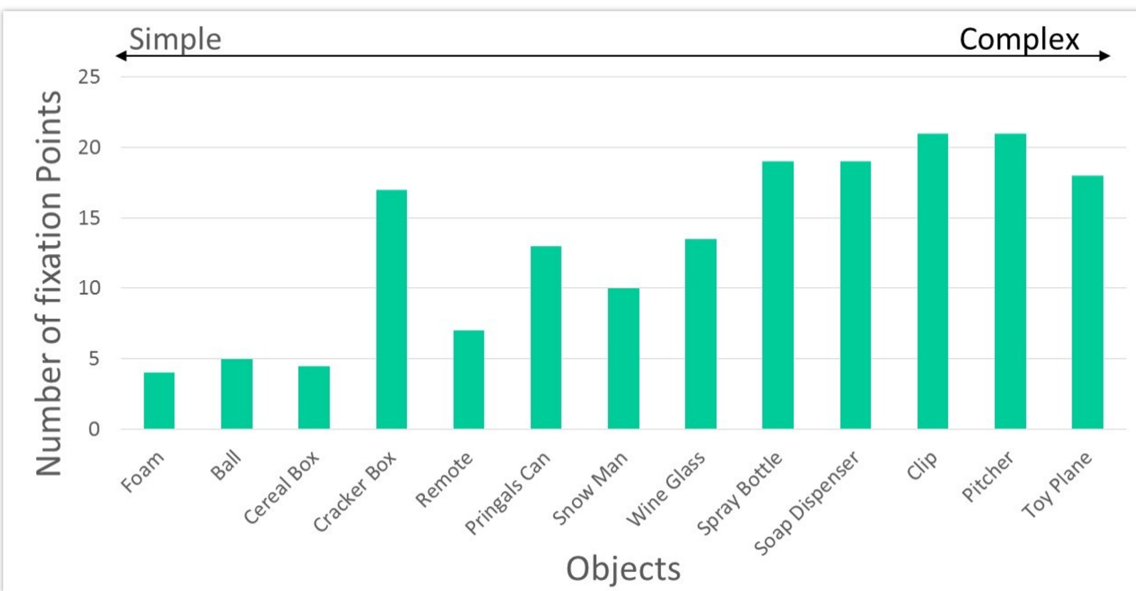


Figure 4.2 Percentage time spent in each stage during robot driven and human driven grasping.

Robotic hand control strategy: Participants varied on exactly how they moved the robot hand, but in general they positioned the hand roughly where they wanted it and with the desired finger spread, and then iterated a few times between adjusting the fingers and repositioning the hand.

4.2 Fixation differences for objects

Objects with more complex geometry saw more fixations than simple objects in the "before" phase (see Fig.4.3). We define complexity by the number of components produced by an automated shape analysis approach such as [60]. From an informal observation of the gaze patterns, participants appeared to be moving between the center line and top of each convex regions of the object (e.g. the wings to the plane body). A more formal evaluation of what regions they were looking at would require tracking the object in the video.



4.3 Fixation counts on the object, organized from simpler shapes to more complex ones.

In fig 4.3 you can see that even though the geometry of cracker box is not complex it has more number of fixation points which is an anomaly. We have gone through the videos and checked why this has happened and we found out that subjects were reading the details which were on the cracker box. We will take care of this by covering the objects in our future studies. Apart from gaze differences and fixation points we haven't found any patterns in eye-gaze data.

4.3 Discussion

Results from eye-gaze data analysis shows that people mainly look at features like object's centerline, wrist and fingertips while performing the task, which is in-line with other research. We also found that people look very less time at their wrist when performing a task with their own hand when compared with performing task with a robot hand. We think this is because humans have been grasping and manipulating object's for a long time and we have complete control over our hands when compared with a robotic hand.

We also presented how much percentage of total time human subjects spent at looking each feature while performing the task in both condition. This will help us in finding what features are important during each stage of grasping which can be used in pruning decision space for grasping. We also found a distinct pattern in fixation points when compared between simple and complex objects. Complex objects saw more fixation points when compared with simple objects. We think more rigorous and formal analysis is required in this direction to make a strong hypothesis. Also we think that more analysis in eye-gaze data might lead to finding new patterns.

There is a huge scope and interest in finding what people are looking at while performing the task. Our research takes a major step in this direction. We are also interested in seeing how eye-gaze patterns differ when human subjects perform task with a different robotic hand and how they change depending upon the flexibility of robotic hand.

4.4 Conclusion

This work presents a major step in the direction of finding what features people think are important for grasping. By conducting a user study and analyzing eye-gaze data we found what features people look at while performing the task and for how much time they are looking at each of these features. We also presented how fixation points vary between complex and simple objects. This information is very useful for understanding how humans process scene environment and how they process data shown to them on a screen. These results are useful in designing better algorithms for grasping and also in designing better visual interfaces for grasping.

Chapter 5 View-point Selection Study

Analysis of eye gaze data has shown some promising results. But how can we leverage these results to improve performance of robotic grasping. In this chapter we describe how the results from previous analysis (eye-gaze data analysis) helped us to design an on-line survey to validate what features people think are important for grasping. The long-term goal is that this information on what features humans consider important for evaluating grasp quality may be used in huma-robot interfaces for teleoperation. We also present the results from the on-line survey in this section.

5.1 Follow-up on-line study on viewpoint selection

Data collected from physical studies are of high quality, but it is very tedious to collect and does not scale. Previous work [26] shows that we can leverage on-line surveys to quickly label and classify grasps by asking participants to evaluate images of them; however, on-going work also shows that view point selection plays a key role in how effective participants are in labeling grasps and how confident they are. Our goal is to use the eye-tracking data to guide an automatic view-point selection algorithm for this use case.

We use the relative percentage viewing time of the features during the robotic grasping hand stage to create a quality/scoring metric to choose the best view-point. We follow this with an on-line survey to evaluate if the algorithm selected views are both effective and useful.

5.1.1 Viewpoint optimization algorithm

Various researchers have used different techniques (presented in background section) for viewpoint selection based upon the domain they are working in. We have designed a viewpoint selection algorithm which depends upon the features people looked at when performing grasping task.

As gravity and object orientation are important components in grasp evaluation, we limit our viewpoint search to azimuth and elevation (see Figure 5.1) (essentially searching a hemisphere of viewpoints, see Figure 5.2). The camera is pointed at the center of the object and the **up** vector is aligned with gravity. The algorithm selects a sample of points on a hemisphere with fixed radius from the center of the object to determine a set of candidate viewpoints. Then we use our scoring metric to choose the 'best' viewpoint. The best viewpoint is the one with highest scoring metric value.

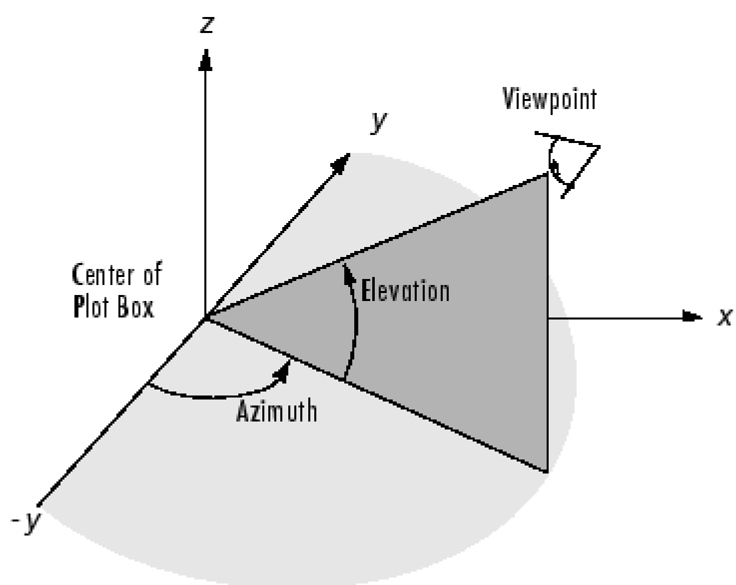


Figure 5.1 Azimuth and Elevation of a view-point

Pseudo code for camera view-point generation

---

view ← load_view()

constant_vec ← get_vec_up(view)

constant_radius ← get_radius(view_coordinate,model_coordinate)

initial_rot ← get_rot(view)

views =[initial_rot]

```
for i in range(# of view points needed)
        R ← get_new_rot()
        if R not in view:
                views.add(R)
end for
```

where,

load_view() = this function loads the view-point

get_vec_up() = given the view-point this function gets the UP vector of the view-point

get_radius() = given the required coordinates this function gets the radius

get_rot = given the initial view-point this function outputs the rotation matrix of the view-point

get_new_rot() = this function (hammersley sampling function) outputs a new view-point

Pseudo Code for hammsersley sampling function

```
def hammersley_points(n):
        points=[ ]
        for k in range(n):
                t = 0
                p = 0.5
                kk = k
                while kk > 0:
                        if (kk & 1):
                                t += p
                        p *=0.5
                        k >>=1
                t = 2.0 * t - 1.0
                theta = (k + 0.5) / n
                thetarad = theta * 2.0 * pi
```

```
                st = sqrt(1.0 - t*t)
                point = (st * cos(thetarad), st * sin(thetarad), t)
                points.append(point)
        return points;
```

where,

n = number of sample points needed

thetarad = value of theta in range [0,2pi]

For every grasp we sampled 250 different view-points. These view-points are generated using a hammersley sampling function which yields sample points on the view-point hemisphere. Our quality/scoring metric for a given viewpoint is simply the sum of the percentage of visible pixels for each feature (normalized by the maximum number of pixels seen from any viewpoint). Each feature is identified by a different color. Each feature is weighted by the percentage of time participants spent viewing that feature, averaged across all participants(top=0.17,edges=0.24,fingertips=0.27,wrist=0.024, center line=0.29). For the contact point feature we added a sphere roughly half the size of the fingertip, centered on the point. The best view is the one with highest score; the second best is the one with next highest score that is at least 30 degrees from the best view. Figure 5.3 shows the plot for range of values of the scoring metric for a single grasp.

$$\text{Scoring metric} = \frac{(no.of\ pixels\,on\,object * w1) + (no.of\ pixles\,on\,wrist * w2) + (no.of\ pixels\,on\,fingertips * w3)}{total\,no.of\ pixels}$$

where,

w1 = normalized percentage of time spent looking at object

w2 = normalized percentage of time spent looking at wrist

w3 = normalized percentage of time spent looking at contact points/fingertips

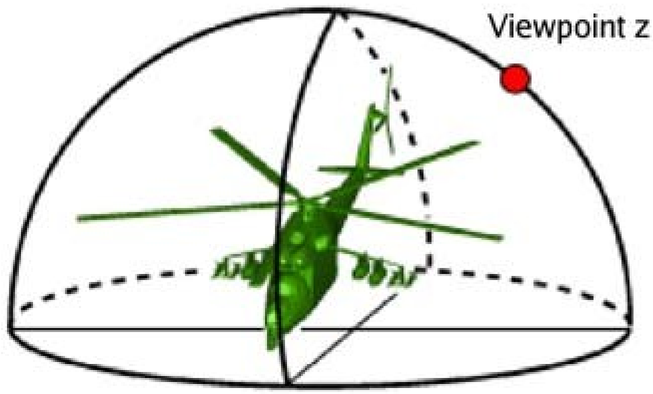


Figure 5.2 A hemisphere from which viewpoints are selected

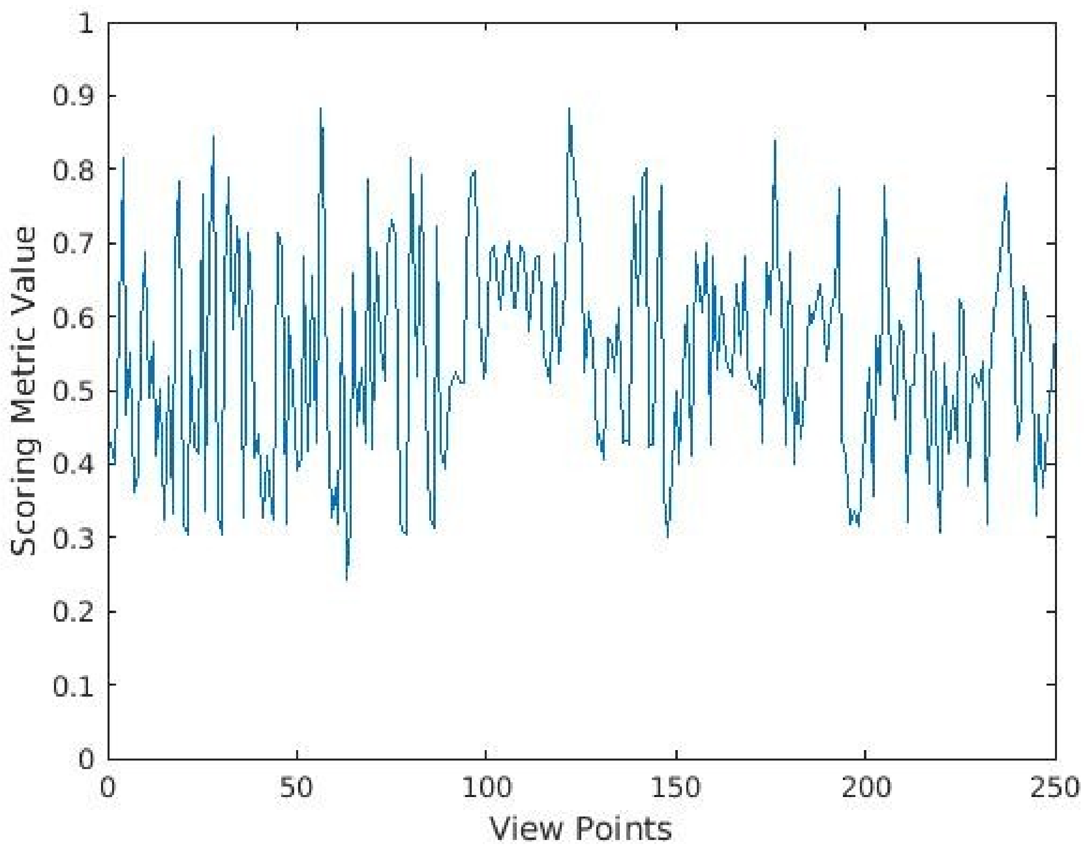


Figure 5.3 Graph showing range of values of scoring metric for a single grasp.

Pseudo code to find quality of view point

```
max ← 0
viewPoint_index ← 0
for i in all viewPoints
    temp_quality ← compute quality metric for ith viewPoint
    if temp_quality ← max then
        max ← temp_quality
        viewpoint_index ← I
    end if
end for
save max and viewPoint_index
```

5.1.2 Survey 1

Our survey was designed to measure both how effective the views were for evaluation (did this grasp work, yes or no?) and perceived usefulness. We used four viewpoints: best and second best (good pair) (see Fig.5.4), and worst and second worst views (bad pair). The survey has four questions (5-point Likert scale) (see Appendix A.1):

1. (All good and bad pairs): Would the grasp work, yes or no, and how confident are you in your answer.
2. (Good pair and bad pair): Rank the usefulness of the first pair with respect to the second pair.
3. (Best and second best/Worst and second worst): Rank the usefulness of the *second* view.
4. (Good pair and bad view): Rank the usefulness of the three views (see Fig.5.5).

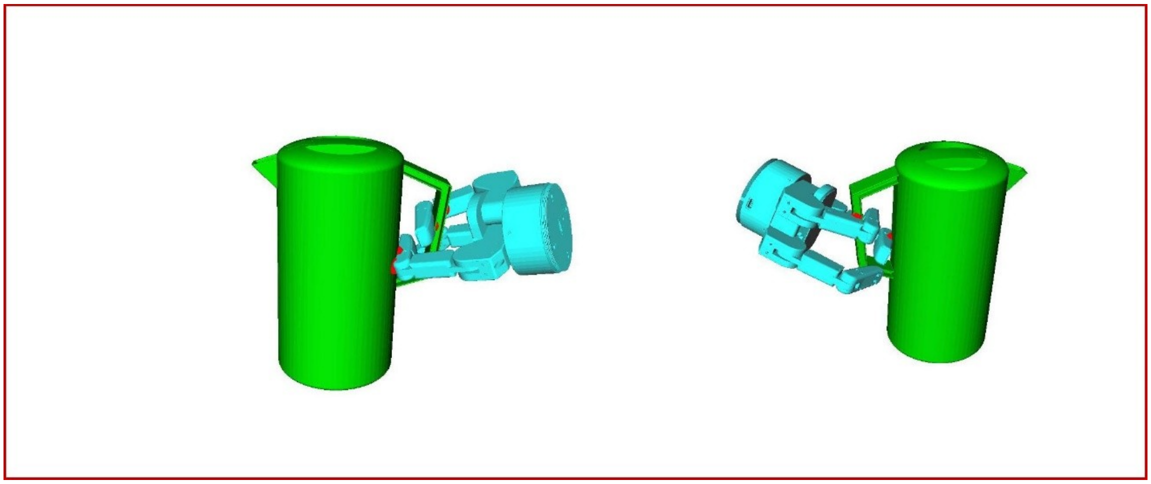


Figure 5.4 Image showing two different views (good view) of same grasp

We evaluated our algorithm with three objects (spray bottle, ball, glass) and two grasps each. These grasps were given by participants in the physical study and subsequently verified as being effective using a shake test. Image order was randomized. Each of the 30 participants saw 20 of the 24 total questions, again randomized. Participants were recruited from Mechanical Turk; we verified that the participants spent enough time on each question to have seen the images and didn't click same thing for every question. This is the preprocessing step on data to eliminate erroneous data.

We conducted a two sample T-test at 5% significance level to determine statistical significance in the data. We determined that there was an order bias; Images on the left tended to be preferred over images shown on the middle or right (Question 1 mean 0.68 versus 0.35, Question 4 mean 3.45 versus 3.07 and 2.31, $p < 0.0007, 0.0005$ respectively). In a previous study [35] we also saw a distinct pattern of looking at the left image then the right, with only far more salient views on the right drawing the gaze first.

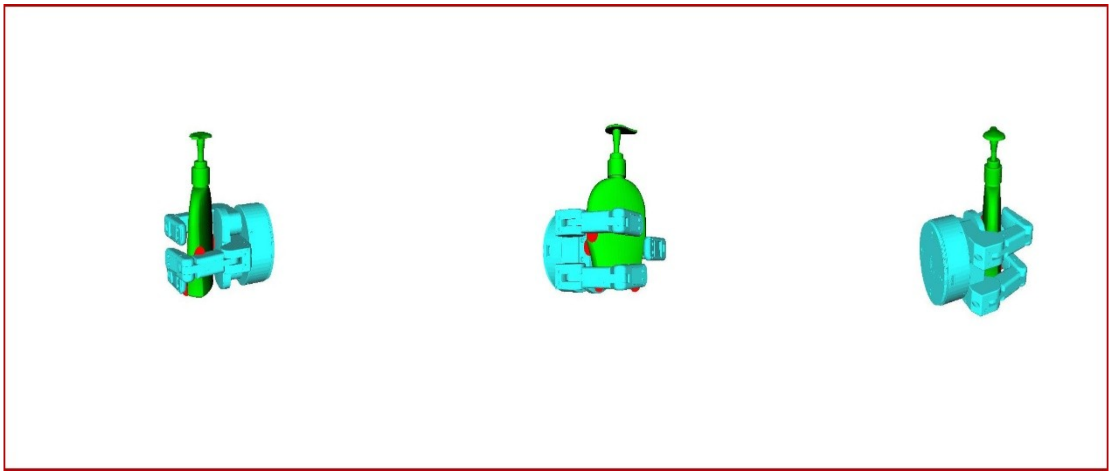


Figure 5.5 Image showing three different views of same grasp. From left to right: first best, second best and first worst view-points

5.2 Results from On-line Survey 1

Refer to Figure 5.6. For Q1 (effectiveness) participants were not only more likely to rank the grasps as effective (means 0.68, 0.35, $p < 0.0007$) but were also more confident in their answer (means 2.3, 1.4, $p < 0.05$). Grasp views were ranked as expected for usefulness both as pairs and individual images (Q2 and Q4, See Fig.5.7, Fig.5.8).

Interestingly, the second best (and first worst) views were both rated approximately as useful (mean 3.07,2.31, $p<0.00039$, 0.00019) relative to their corresponding first view (Q4). The answers to Q3 did not yield data with statistical significance.

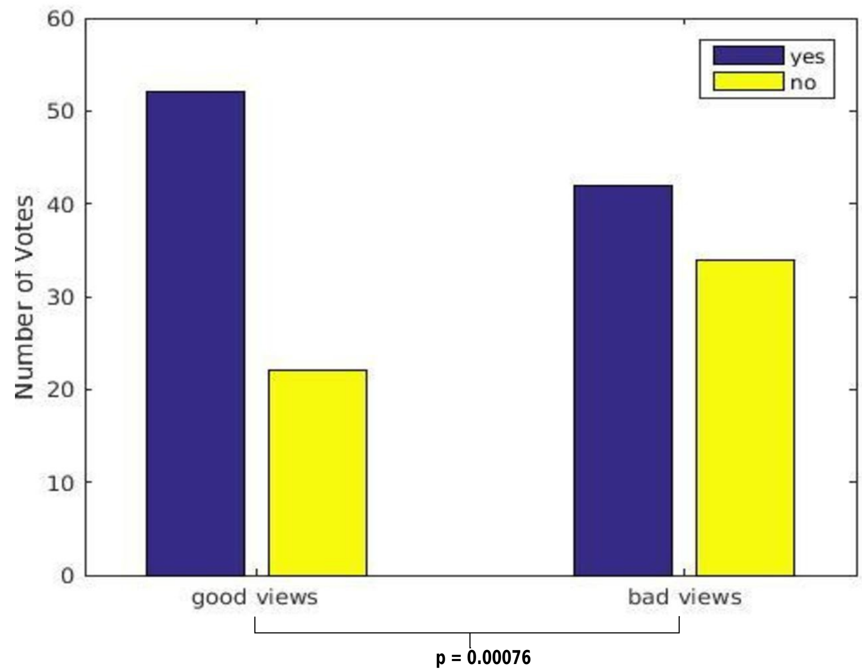


Figure 5.6 Results of on-line survey 1. Grasp works y/n (Q1, good grasps)

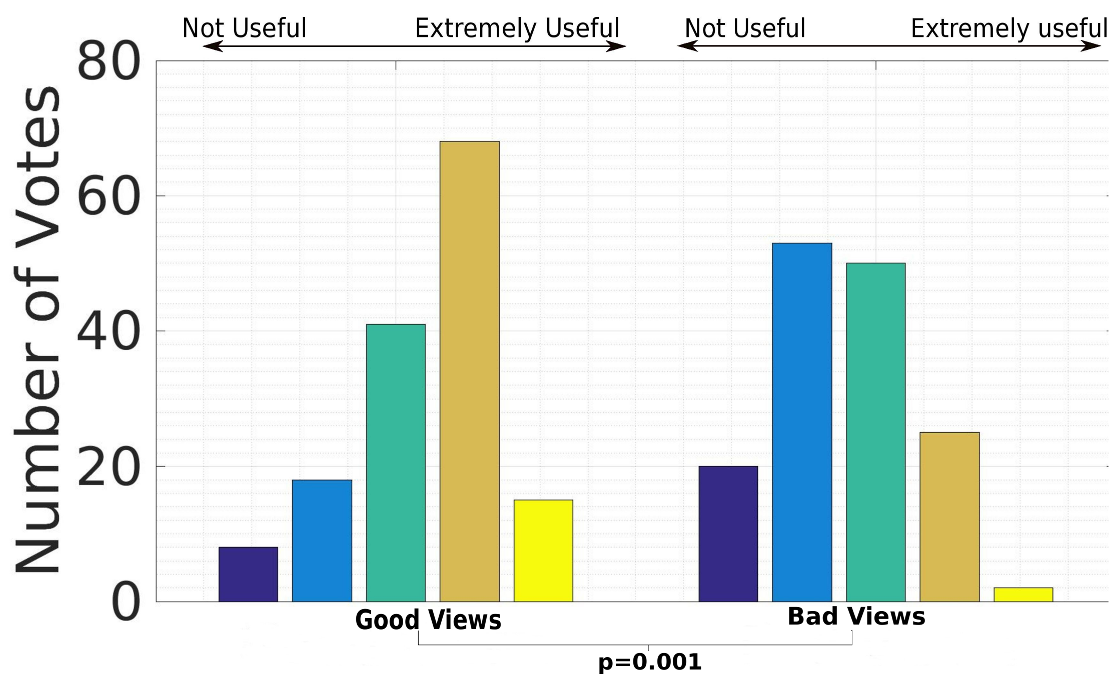


Figure 5.7 Results of on-line survey 1. Usefulness of view pairs (Q2, good grasps).

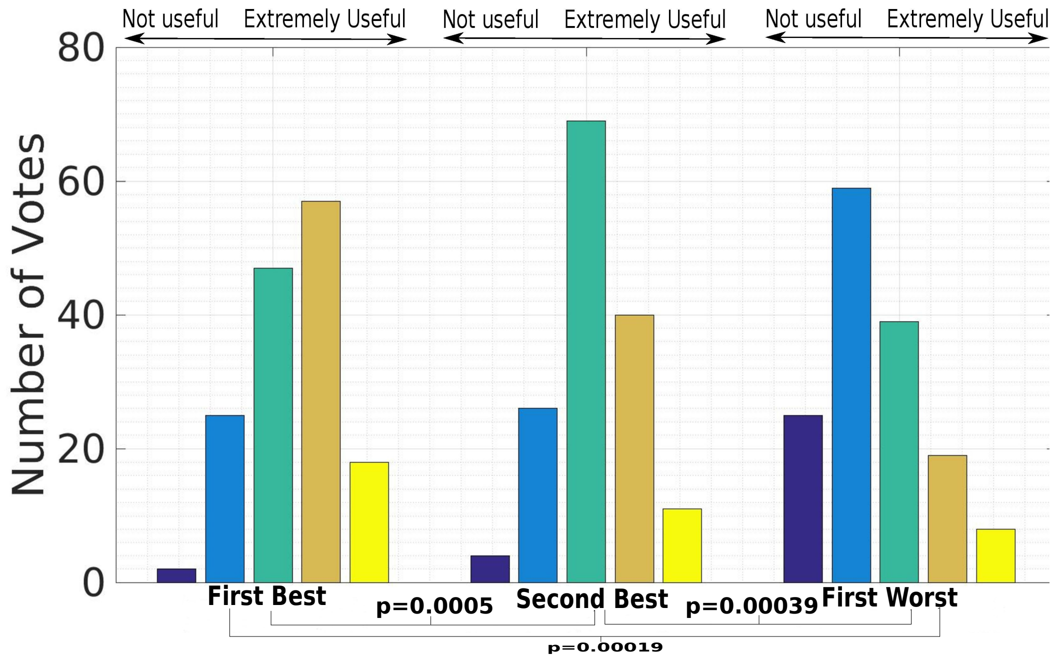


Figure 5.8 Results of on-line survey 1. Usefulness of first and second best view to worst (Q4, good grasps). (useful is to the right.)

5.3 Survey 2

In previous survey we shown the participants images of good grasps, i.e.,a good grasp is the one which has passed shake test and asked them to answer a set of questions. But in survey 2 we have shown the participants images of bad grasps (see Fig 5.9), i.e., a bad grasp is the one which has failed shake test and asked them a set of questions about quality of grasp. These views are produced by the same algorithm used for generating views in survey 1.

Survey 2 consists of 2 set of questions. In set 1, we showed them images of bad grasps. And in set 2, we showed them those images which got most number of yes and no votes for good and bad viewpoints respectively. The main idea of set 2 questions is to cross verify consistency of answers from the participants. In this survey we did not visualize contact points. These are the questions in the survey (5-point Likert scale) (see Appendix A.2):

1. (All good and bad pairs): Would the grasp work, yes or no, and how confident are you in your answer.
2. Why do you think the grasp will fail (this option only shows up if they answered 'no' for question 1)?
    - Object will slip from hand
    - Object will rotate when grasping
    - Fingers are not closed around object
    - Other reasons – please say why (descriptive answer)

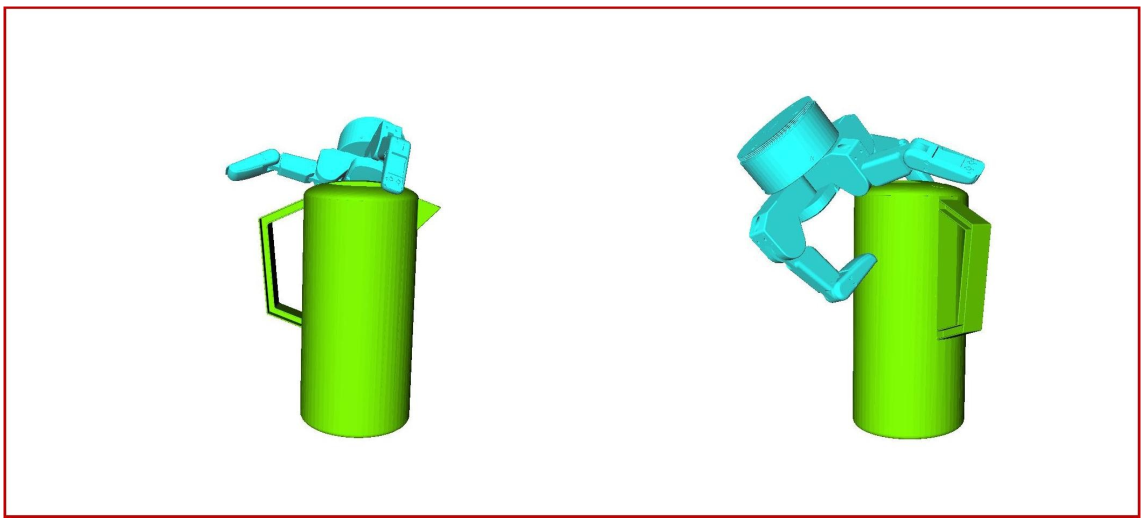


Figure 5.9 Image showing two different views of same grasp (bad grasp)

5.4 Results from On-line Survey 2

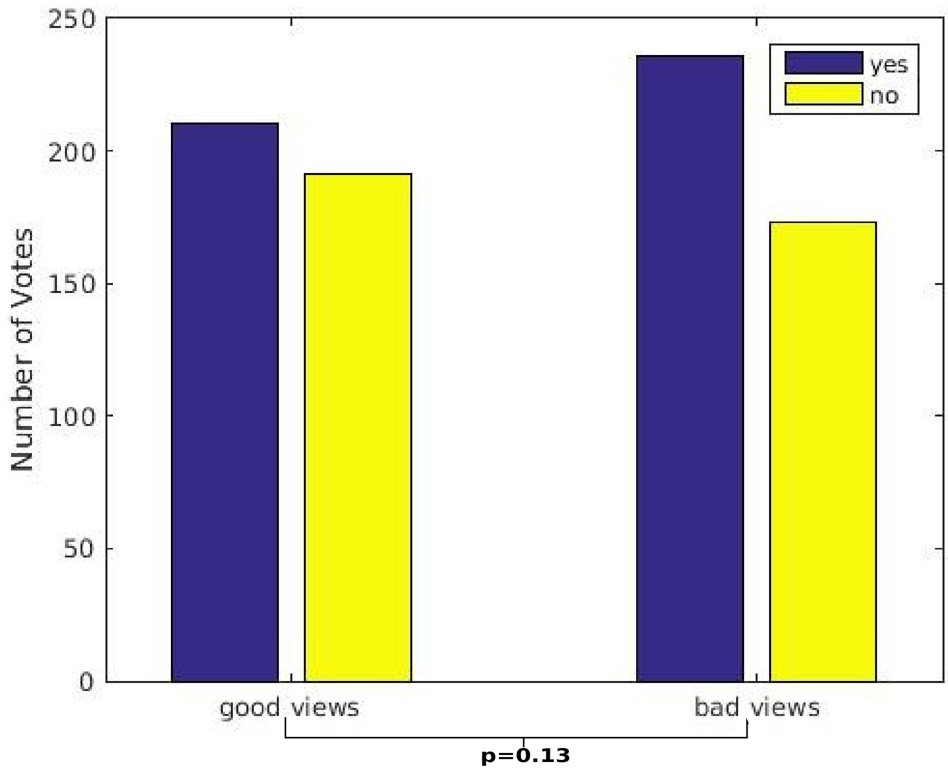


Figure 5.10 Results of on-line survey 2. Grasp works y/n (Q1, bad grasps)

We conducted a two sample T-test at 5% significance level to show that the data is statistically significant. Figure 5.10 shows the results from survey 2 question 1, where we asked the participants whether the grasp works or not based upon the image shown to them. For both good and bad views we almost got equal yes and no votes, which means participants were unable to say whether the grasp works or not based upon the image shown to them ($p > 0.1$). We think this is because grasps shown in the survey are really bad that the quality of view-point did not make a difference in their decision if the grasp would fail or not. Most of the participants answered that the object might slip from the hand while grasping or fingers are not completely closed around the object. We think this is because of not visualizing the contact points as all fingers didn't make any contact with objects in most grasps.

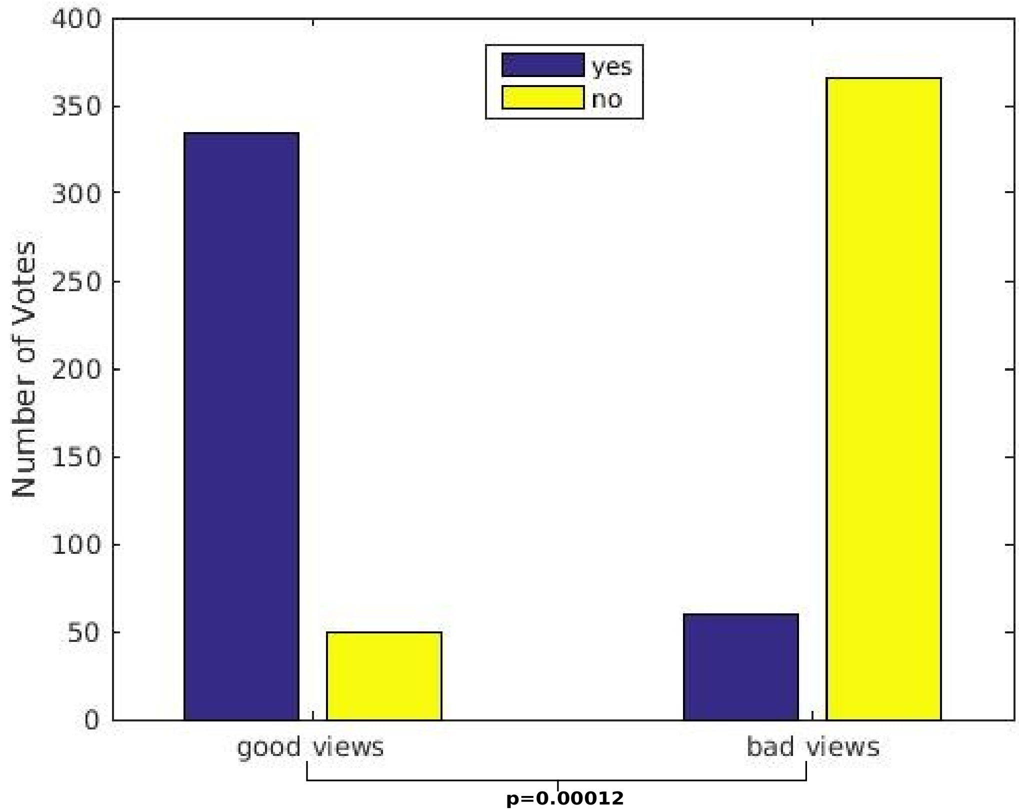


Figure 5.11 Results of on-line survey 2. Grasp works y/n (Q2, good grasps)

Figure 5.11 shows the results from survey 2 question 2, where we showed images of good grasps and asked participants whether the grasp works or not. These are the images which got significant number of yes and no votes for good and bad views respectively($p < 0.01$) . As you can see, when we shown them good views of a grasp they were able to say that the grasp works and when we shown them bad views of the grasp they voted that the grasp wouldn't work. These results are in line with results from survey 1.

From survey 2 results we hypothesized that visualizing contact points plays a vital role in deciding whether the grasp works or not. Also people assume grasps which are not human like tends to fail more often when compared to human-like grasps. More research has to to be done in order to make a strong hypothesis.

5.5 Discussion

Our results from question 1 in on-line survey 1 suggests that people were able to distinguish whether a grasp works or not based upon the views we provided. When we showed them a bad view of a good grasp most people answered that the grasp won't work, but when we showed them good view of the same grasp most people answered that the grasp will work and confident about their answer. We think this is because good views captured most useful information about the scene and they showed those features what people think are important for grasping. But in case of bad views, they didn't capture useful information about the scene. Results from question 2 and 4 from on-line survey 1 suggests that people were also able to rank the relative usefulness of different views (best, second best and worst) correctly.

Most researchers has pointed out that operators had a hard time in understanding the visual information, i.e., 2-D views, presented to them and in using interfaces while manipulating an objects. We think these results helps researchers to better understand how humans perceive visual information presented to them which indeed helps them to design better interfaces and they can use the visual cues we found in our research to present the visual information in the best possible way to a teleoperator so that they can understand the environment better.

Results from question 1 survey 2 did not yield any important information. At present we don't know the exact reason why this has happened. We think this is because grasps shown in the survey are really bad that the quality of view-point did not make a difference in their decision if the grasp would fail or not. Most participants answered that grasps might fail because the object might slip from the hand. It is interesting to know how visualizing contact points might affect the judgment of the participant. More research should be done in order to find whether visualizing contact points has any affect or not. Results from question 2 are in agreement with results from survey 1.

5.6 Conclusion

In this work we have explored how efficient the features we found from the previous study are in finding whether a grasp works or not. We have explored this with both good and bad grasps with three different objects and two grasps per object. We also identified that providing good views helps the users to understand the scene better.

Our main contribution from this study is that by showing useful features like contact points, have tremendous effects on how people judge the effectiveness of a grasp. This is a major step towards understanding how people perceive the information presented to them. More work needs to be done to have a better understanding about the effects of these features and we think that there is a lot of scope for future research in this area. Our future work will include making the selection of best views automated and integrating it with visual interfaces like Rviz and seeing the effects of it.

Chapter 6 Conclusion

Understanding human gaze can tell us a lot about how humans perceive visual information while performing a task and this knowledge can be used to design better interfaces for teleoperation and also improve robotic grasping performance using those same heuristics. In chapter 3, 4 and 5, I presented what features people think are important for grasping and evaluated their efficiency using an on-line survey. This begins to provide an understanding of how humans perceive visual information and which visual cues they substitute for missing tactile cues when performing a physical interaction task. This research is just the beginning step in understanding human gaze patterns.

An analysis of eye gaze data has shown us that people seldom looked at the object's top and edges or the robot's wrist or fingertips in the “before phase” when performing the task with robot hand (see Fig.4.1). But when they performed the task using their own hand they mainly focus on the top and edge of the object, as found in previous studies. The main difference between two eye gaze patterns in “during” and “evaluation” phases is that participants spent more time observing the robot's wrist and less time looking at the object's centerline. We think this is because participants are trying to adjust the robot's wrist to lower the skewness of the grasp (that is, the angle between the object’s principal axis and the wrist orientation). We made an informal observation that for objects with complex geometry has more fixations than simple objects (see Fig.4.3). A more formal evaluation of this result is required to understand what regions they were looking at during grasping. Along with these interesting results, Chapter 4 raised interesting questions regarding changes in gaze patterns when participants uses a different robotic hand and how does the flexibility of robotic hand changes gaze patterns.

Chapter 5 explored in detail the effectiveness of features to decide whether a grasp works or not based upon the views we provided. Participants rated that the best views — which shows useful features — are more effective and useful in deciding whether the grasp works or not when compared to other views (see Fig.5.6, Fig.5.7,Fig.5.8). The results from on-line study suggests that participants were able to judge whether a grasp works or not when we provided them good views. This suggests that showing good views which captures useful information about the environment helps teleoperator tremendously in successfully grasping an object. We have also formulated a scoring metric to evaluate the quality of views. Along with these interesting results, on-line survey raised interesting questions regarding effects of visualization of features on deciding grasp quality. In this thesis, We have presented an analysis of the difference in eye gaze when participants used their own hand versus manipulating a robotic one.

Eye-Gaze analysis is totally different from the existing methods used by grasping community like Human in The Loop, Learning from Demonstration and we think that this domain needs our attention and efforts due to it's uniqueness and potential impact on grasping community. Even though there is an overlap with other research areas, most of them have not addressed questions regarding human perception of visual information and we think eye-gaze analysis should be classified as a new domain in robotics community. We just scratched the surface of how we can use eye-gaze in improving robotic grasping. Researchers has analyzed eye-gaze to improve performance of human-robot interaction [61]. And there is a lot of research going on in understanding human perception. There is a multitude of possibilities on how we can use eye-gaze information in robotics.

This research will push the boundaries in understanding human gaze patterns. Findings from this research reinforces the idea of providing good views improves the performance of teleoperators and questions raised in Chapter 4 and Chapter 5 will encourage more researchers to study about eye-gaze patterns and how it will be helpful in improving performance of grasping. This research provides initial results regarding how eye-gaze results can be used to improve decision making of teleoperators and in presenting visual data in a best possible way.

Grasping community will need to focus more on understanding human gaze patterns in various scenarios. The long-term goal of this work is to help in designing better interfaces for tele-operation and help in understanding how people process visual information when performing grasping tasks in various environments.

BIBLIOGRAPHY

[1] R. R. Murphy, "Human – Robot Interaction in Rescue Robotics," vol. 34, no. 2, pp. 138–153, 2004.

[2] J. L. Casper and R. R. Murphy, "Workflow Study on Human-Robot Interaction in USAR," pp. 368–370, 1995.

[3] D. M. Taylor, S. I. H. Tillery, and A. B. Schwartz, "Direct cortical control of 3D neuroprosthetic devices.," *Science*, vol. 296, no. 5574, pp. 1829–32, 2002.

[4] A. Edsinger and C. C. Kemp, "Human-Robot Interaction for Cooperative Manipulation : Handing Objects to One Another," pp. 1167–1172, 2007.

[5] J. D. Drury, J. Scholtz, and H. A. Yanco, "Awareness in human-robot interactions," *Syst. Man Cybern.*, vol. 1, pp. 912–918, 2003.

[6] J. L. Burke and R. R. Murphy, "Human-robot interaction in USAR technical search: two heads are better than one," *RO-MAN 2004. 13th IEEE Int. Work. Robot Hum. Interact. Commun. (IEEE Cat. No.04TH8759)*, pp. 307–312, 2004.

[7] A. Steinfeld, T. Fong, M. Field, M. Lewis, J. Scholtz, and A. Schultz, "Common Metrics for Human-Robot Interaction."

[8] S. Myung, "The Effect of Automated Compensation Axes on Teleoperator Performance for Incongruent," vol. 40, no. 4, pp. 541–553, 1998.

[9] D. B. Kaber, J. M. Riley, R. Zhou, and J. Draper, "Effects of Visual Interface Design, and Control Mode and Latency on Performance, Telepresence and Workload in a Teleoperation Task," *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 44, no. 5, pp. 503–506, 2000.

[10] B. Gooch, E. Reinhard, C. Moulding, and P. Shirley, "Artistic Composition for Image Creation," *Proc. 12th Eurographics Work. Render. Tech.*, pp. 83–88, 2001.

[11] a. M. Dollar and R. D. Howe, "The Highly Adaptive SDM Hand: Design and Performance Evaluation," *Int. J. Rob. Res.*, vol. 29, no. 5, pp. 585–597, 2010.

[12] L. BIRGLEN, T. LALIBERTE, and C. GOSSELIN, *Under actuated Robotic Hands*. 2008.

[13] E. Brown, N. Rodenberg, J. Amend, a. Mozeika, E. Steltz, M. R. Zakin, H. Lipson, and H. M. Jaeger, "From the Cover: Universal robotic gripper based on the jamming

of granular material," *Proc. Natl. Acad. Sci.*, vol. 107, no. 44, pp. 18809–18814, 2010.

[14] S. Chitta, I. Sucan, and S. Cousins, "MoveIt! [ROS Topics]," *IEEE Robot. Autom. Mag.*, vol. 19, no. 1, pp. 18–19, 2012.

[15] A. Saxena, J. Driemeyer, and a. Y. A. Ng, "Robotic Grasping of Novel Objects using Vision," *IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, vol. 27, no. 2, pp. 157–173, 2008.

[16] E. Lopez-Damian, D. Sidobre, and R. Alami, "A grasp planner based on inertial properties," *Proc. - IEEE Int. Conf. Robot. Autom.*, vol. 2005, no. April, pp. 754–759, 2005.

[17] B. León, S. Ulbrich, R. Diankov, G. Puche, M. Przybylski, A. Morales, T. Asfour, S. Moisio, J. Bohg, J. Kuffner, and R. Dillmann, "OpenGRASP: A toolkit for robot grasping simulation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6472 LNAI, pp. 109–120, 2010.

[18] R. Balasubramanian, L. Xu, P. D. Brook, J. R. Smith, and Y. Matsuoka, "Physical human interactive guidance: Identifying grasping principles from human-planned grasps," *Springer Tracts Adv. Robot.*, vol. 95, no. 4, pp. 477–500, 2014.

[19] A. K. Goins, R. Carpenter, W.-K. Wong, and R. Balasubramanian, "Implementation of a Gaussian process-based machine learning grasp predictor," *Auton. Robots*, vol. 40, no. 4, pp. 687–699, 2015.

[20] a. T. Miller and P. K. Allen, "GraspIt!," *IEEE Robot. Autom. Mag.*, vol. 11, no. 4, pp. 110–122, 2004.

[21] J. Bohg, A. Morales, T. Asfour, and D. Kragic, "Data-Driven Grasp Synthesis---A Survey," *IEEE Trans. Robot.*, vol. 30, no. 2, pp. 289–309, 2013.

[22] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Rob. Auton. Syst.*, vol. 57, no. 5, pp. 469–483, 2009.

[23] S. Ekvall and D. Kragic, "Interactive grasp learning based on human demonstration," *Robot. Autom. 2004. Proceedings. ICRA '04. 2004 IEEE Int. Conf.*, vol. 4, pp. 3519–3524 Vol.4, 2004.

[24] B. Argall, B. Browning, and M. Veloso, "Learning by demonstration with critique from a human teacher," *Proceeding ACM/IEEE Int. Conf. Human-robot Interact. - HRI '07*, p. 57, 2007.

[25] R. Balasubramanian, L. Xu, P. D. Brook, J. R. Smith, and Y. Matsuoka, "Human-guided grasp measures improve grasp robustness on physical robot," *Proc. - IEEE Int. Conf. Robot. Autom.*, pp. 2294–2301, 2010.

[26] M. Unrath, Z. Zhang, A. Goins, R. Carpenter, W. Wong, and R. Balasubramanian, "Using Crowdsourcing to Generate Surrogate Training Data for Robotic Grasp Prediction," pp. 60–61, 2014.

[27] Y. Lin and Y. Sun, "Robot grasp planning based on demonstrated grasp strategies," *Int. J. Rob. Res.*, vol. 34, no. 1, pp. 26–42, 2014.

[28] J. D. Sweeney and R. Grupen, "A Model of Shared Grasp Affordances from Demonstration," *Proc. 2007 7th IEEE-RAS Int. Conf. Humanoid Robot. HUMANOIDS 2007*, pp. 27–35, 2008.

[29] J. J. Gibson, "The theory of affordances."

[30] H. O. Song, M. Fritz, C. Gu, and T. Darrell, "Visual grasp affordances from appearance-based cues," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 998–1005, 2011.

[31] M. Gridseth, K. Hertkorn, and M. Jagersand, "On Visual Servoing to Improve Performance of Robotic Grasping," *Proc. -2015 12th Conf. Comput. Robot Vision, CRV 2015*, pp. 245–252, 2015.

[32] T. B. Sheridan, "Space Teleoperation through Time-Delay - Review and Prognosis," *Ieee Trans. Robot. Autom.*, vol. 9, no. 5, pp. 592–606, 1993.

[33] A. E. Leeper, K. Hsiao, M. Ciocarlie, L. Takayama, and D. Gossow, "Strategies for human-in-the-loop robotic grasping," *ACM/IEEE Human-Robot Interact.*, pp. 1–8, 2012.

[34] P. Michelman and P. Allen, "Shared autonomy in a robot hand teleoperation system," vol. 1, pp. 253–259 vol.1, 1994.

[35] M. Sundberg, W. Litwinczyk, C. Grimm, and R. Balasubramanian, "Visual cues used to evaluate grasps from images *."

[36] L. Desanghere and J. J. Marotta, "'Graspability' of objects affects gaze patterns during perception and action tasks," *Exp. Brain Res.*, vol. 212, no. 2, pp. 177–187, 2011.

[37] S. L. Prime and J. J. Marotta, "Gaze strategies during visually-guided versus memory-guided grasping," *Exp. Brain Res.*, vol. 225, no. 2, pp. 291–305, 2013.

[38] J. M. Lawrence, K. Abhari, S. L. Prime, B. P. Meek, L. Desanghere, L. a Baugh, and J. J. Marotta, “A novel integrative method for analyzing eye and hand behaviour during reaching and grasping in an MRI environment.,” *Behav. Res. Methods*, vol. 43, no. 2, pp. 399–408, 2011.

[39] V. Blanz, M. J. Tarr, and H. H. B??lthoff, “What object attributes determine canonical views?,” *Perception*, vol. 28, no. 5, pp. 575–597, 1999.

[40] S. Edelman and H. H. Bülthoff, “Orientation dependence in the recognition of familiar and novel view of three-dimensional objects,” *Vision Res.*, vol. 32, no. 12, pp. 2385–2400, 1992.

[41] M. Christie and P. Olivier, “Camera Control in Computer Graphics: Models, Techniques and Applications,” *Proc. ACM SIGGRAPH ASIA 2009 Courses*, p. Article No.: 3, 2009.

[42] P. Burelli, L. Di Gaspero, A. Ermetici, and R. Ranon, “Virtual camera composition with particle swarm optimization,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5166 LNCS, pp. 130–141, 2008.

[43] E. Marchand and N. Courty, “Image-Based Virtual Camera Motion Strategies,” *Graph. Interface Conf., GI2000*, pp. 69–76, 2000.

[44] J. Assa, D. Cohen-Or, I.-C. Yeh, and T.-Y. Lee, “Motion overview of human actions,” *ACM Trans. Graph.*, vol. 27, no. 5, p. 1, 2008.

[45] E. Bourque and G. Dudek, “Automatic Creation of Image-Based Virtual Reality,” *Sens. Fusion Decentralized Control Auton. Robot. Syst.*, vol. 3209, pp. 292–303, 1997.

[46] K. Tomihisa and K. Satoru, “A simple method for computing general position in displaying three- dimensional objects.”

[47] C. C, “A system for exploring the universe of polyhedral shapes.”

[48] D. PLEMENOS and M. BENAYADA, “Intelligent display in scene modelling. new techniques to automatically compute good views.”

[49] M. Feixas, M. Sbert, and W. Heidrich, “viewpoint selection using viewpoint entropy,” 2001.

[50] E. Serin, C. Doger, and S. Balcisoy, “3D Object Exploration using Viewpoint Entropy,” *Hand, The*, vol. 2011, no. EuroVis, 2011.

[51] P. P. V??zquez, M. Feixas, M. Sbert, and A. Llobet, “Realtime automatic selection of good molecular views,” *Comput. Graph.*, vol. 30, no. 1, pp. 98–110, 2006.

[52] P.-P. Vazquez and M. Sbert, “fast adaptive selection of best views.”

[53] P. P. V??zquez, M. Feixas, M. Sbert, and W. Heidrich, “Automatic View Selection Using Viewpoint Entropy and its Application to Image-Based Modelling,” *Comput. Graph. Forum*, vol. 22, no. 4, pp. 689–700, 2003.

[54] T. Kato, T. Ikai, S. Aoki, M. Izumi, and K. Fukunaga, “Viewpoint Selection Based on Fechner Type Information Quantities for 3D Objects,” pp. 1–2.

[55] D. Sokolov and D. Plemenos, “Viewpoint Quality and Scene Understanding,” *Proc. 6th Int. Conf. Virtual Reality, Archaeol. Intell. Cult. Herit.*, pp. 67–73, 2005.

[56] F. González, M. Sbert, and M. Feixas, “Viewpoint-based ambient occlusion,” *IEEE Comput. Graph. Appl.*, vol. 28, no. 2, pp. 44–51, 2008.

[57] T. Vieira, A. Bordignon, A. Peixoto, G. Tavares, H. Lopes, L. Velho, and T. Lewiner, “Learning good views through intelligent galleries,” *Comput. Graph. Forum*, vol. 28, no. 2, pp. 717–726, 2009.

[58] V. Krassanakis, V. Filippakopoulou, and B. Nakos, “EyeMMV toolbox: An eye movement post-analysis tool based on a two-step spatial dispersion threshold for fixation identification,” *J. Eye Mov. Res.*, vol. 7(1), no. 1, pp. 1–10, 2014.

[59] D. D. Salvucci and J. H. Goldberg, “Identifying Fixations and Saccades in Eye-Tracking Protocols,” *Proc. Eye Track. Res. Appl. Symp.*, pp. 71–78, 2000.

[60] O. Van Kaick, N. Fish, Y. Kleiman, S. Asafi, and D. Cohen-or, “Shape Segmentation by Approximate Convexity Analysis,” *ACM Trans. Graph.*, vol. VV, no. 1, 2014.

[61] J. D. Boucher, U. Pattacini, A. Lelong, G. Bailly, F. Elisei, S. Fagel, P. F. Dominey, and J. Ventre-Dominey, “I reach faster when i see you look: Gaze effects in human-human and human-robot face-to-face cooperation,” *Front. Neurorobot.*, vol. 6, no. MAY, pp. 1–11, 2012.

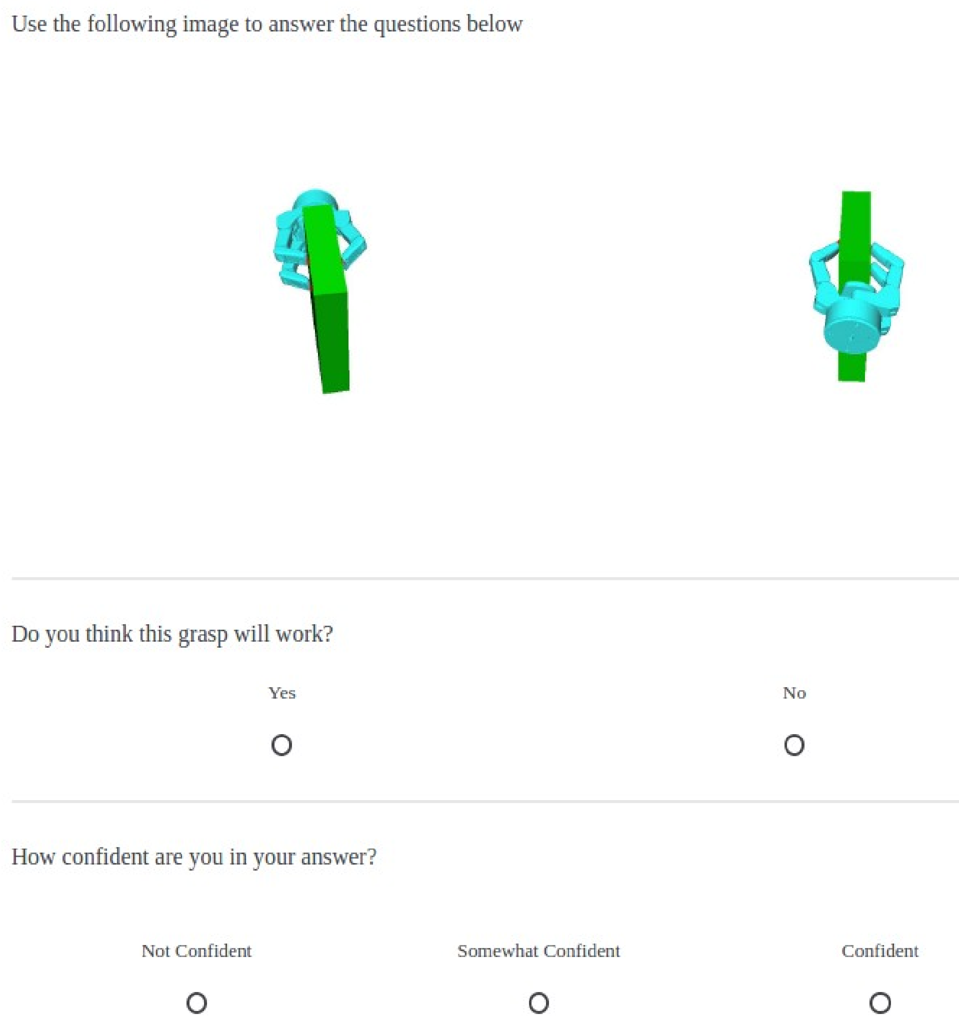
APPENDICES

Appendix A: On-line Survey

A.1 On-line Survey 1

This sections describes the questions we asked the participants in survey 1. Survey 1 consists of four questions which are described below:

Question 1:


Use the following image to answer the questions below

Do you think this grasp will work?

Yes No

How confident are you in your answer?

Not Confident Somewhat Confident Confident

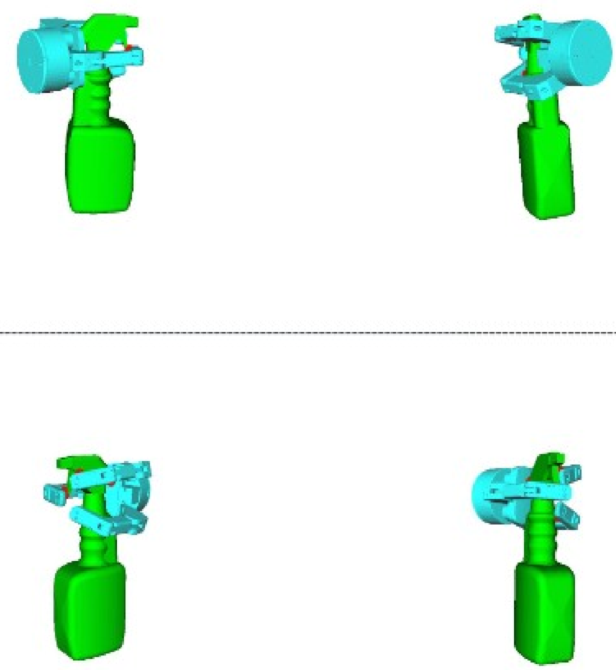
Figure A.1 Question 1 from survey 1

In question 1, we shown them an image (good view or bad view) of a grasp and asked questions shown in figure A.1.

Question 2:

In question 2, we shown them two images (good view and bad view) of same grasp and asked questions shown in figure A.2

Use the following images to answer the question below



These are two pairs of images of the same grasp. Rank how useful each pair of images are for determining hc well the grasp will work.

	Not at all useful	slightly useful	Moderately useful	very useful	Extremel useful
Pair 1(top)	○	○	○	○	○
Pair 2(bottom)	○	○	○	○	○

Figure A.2 Question 2 from survey 1

Question 3:

In question 3, we shown them an image (good view or bad view) of a grasp and asked questions shown in figure A.3

Use the following image to answer the question below

The above image consists of two views of same grasp. Rank how useful second view is in determining how well the grasp will work.

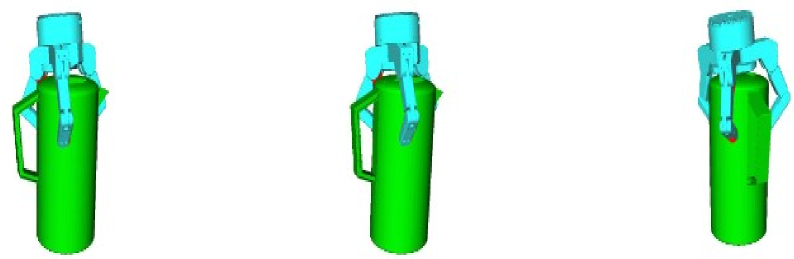
	Not at all useful	Slightly useful	Moderately useful	Very useful	Extremely useful
View 2(right)	O	O	O	O	O

Figure A.3 Question 3 from survey 1

Question 4:

In question4, we shown them three different views (best, second best and worst) of same grasp and asked questions shown in figure A.4.

Use the following image to answer the question below



The image show three different views of the *same* grasp. Rank how useful each view is for determining how well the grasp will work.

	Not at all useful	Slightly useful	Moderately useful	Very useful	Extremely useful
View 1(left)	○	○	○	○	○
View 2(middle)	○	○	○	○	○
View 3(right)	○	○	○	○	○

Figure A.4 Question 4 from survey 1

A.2 Survey 2

In this survey we have shown the participants images of both bad and good grasps and asked them set of questions which are described below:

Question1:

In question 1, we showed them image (good view or bad view) of a grasp and asked them questions shown in figure A.5

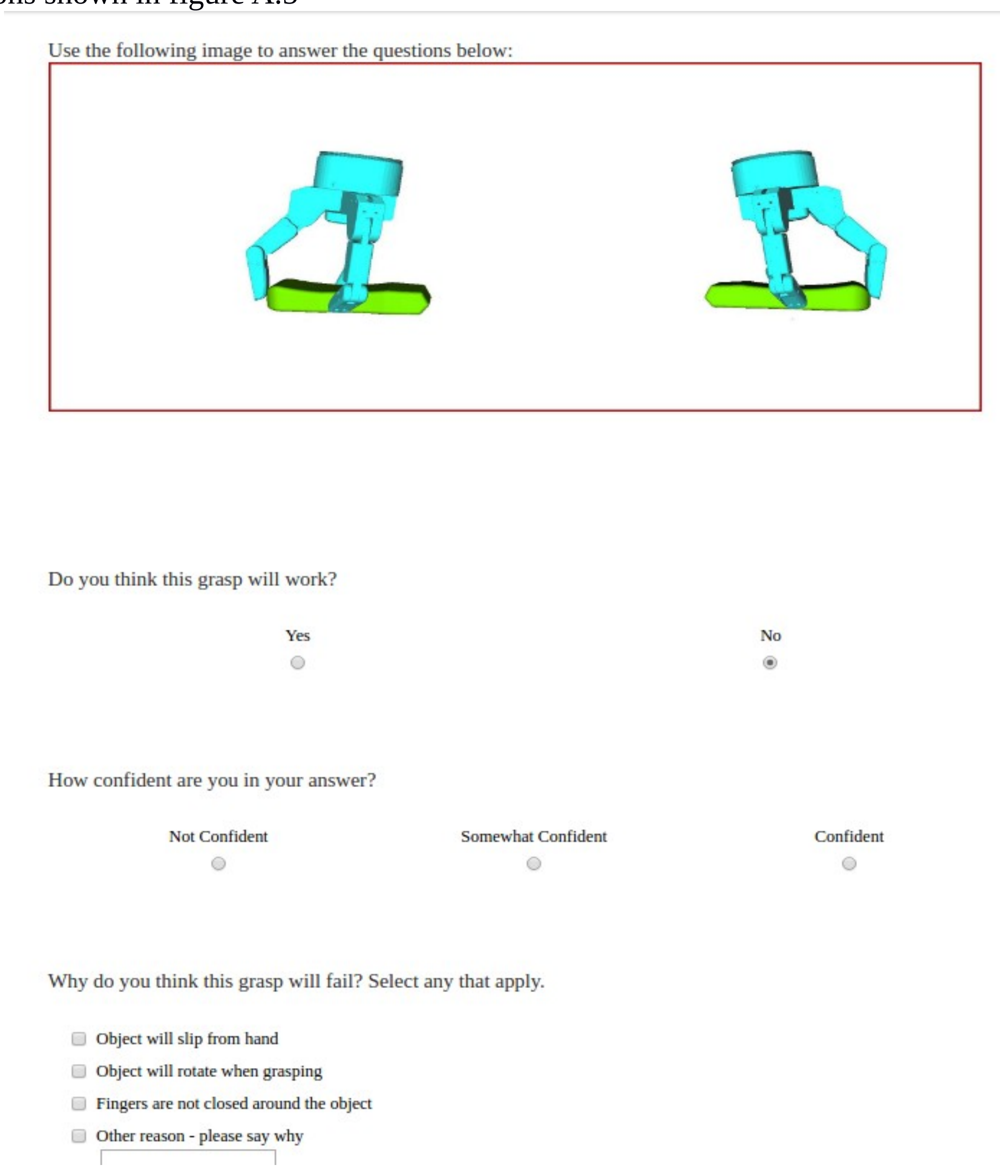
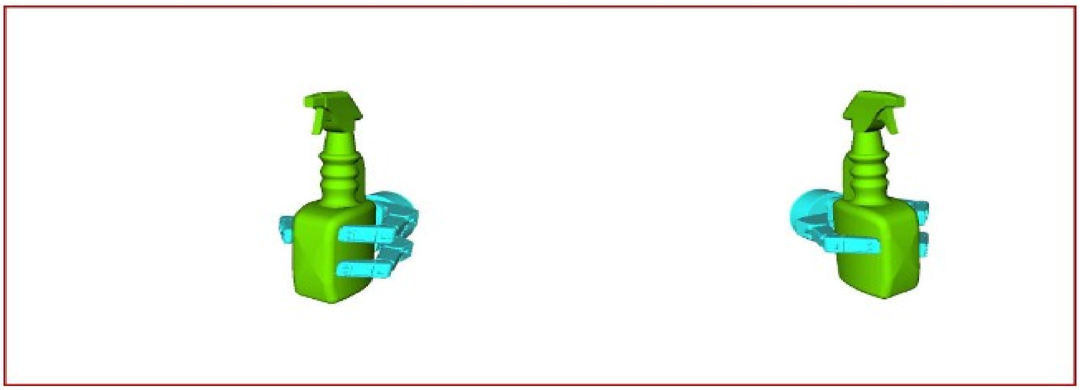


Figure A.5 Question 1 from survey 2 (bad grasp)

Question 2:

In question 2, we showed them image (good view or bad view) of a grasp and asked them questions shown in figure A.6.

Use the following image to answer the questions below:



Do you think this grasp will work?

Yes ○ | No ◉

How confident are you in your answer?

Not Confident ○ | Somewhat Confident ○ | Confident ○

Why do you think this grasp will fail? Select any that apply.

- ☐ Object will slip from hand
- ☐ Object will rotate when grasping
- ☐ Fingers are not closed around the object
- ☐ Other reason - please say why

Figure A.6 Question 2 from survey 2 (good grasp)

A.3 P-values for eye-gaze data

The following to tables shows p-values for eye-gaze data across all participants in three different stages of grasping.

Table A.1 P-values for eye-gaze data

	b_top	d_top	a_top	b_wrist	d_wrist	a_wrist	b_edge	d_edge	a_edge	b_fingertip	d_fingertip	a_fingertip	b_centerline	d_centerline	a_centerline
b_top	1	0.4	0.33	0.02	0.06	0.04	0.38	0.04	0.05	0.02	0.18	0.69	0	0	0
d_top	0.4	1	0.86	0	0	0	0.87	0.11	0.16	0	0.52	0.49	0	0	0
a_top	0.33	0.86	1	0	0	0	1	0.17	0.23	0	0.66	0.38	0	0	0
b_wrist	0.02	0	0	1	0	0.03	0	0	0	0.6	0	0	0	0	0
d_wrist	0.06	0	0	0	1	0.25	0	0	0	0	0	0	0	0	0
a_wrist	0.04	0	0	0.03	0.25	1	0	0	0	0.05	0	0	0	0	0
b_edge	0.38	0.87	1	0	0	0	1	0.22	0.3	0	0.7	0.46	0	0	0
d_edge	0.04	0.11	0.17	0	0	0	0.22	1	0.74	0	0.29	0.02	0	0.06	0.06
a_edge	0.05	0.16	0.23	0	0	0	0.3	0.74	1	0	0.42	0.02	0	0.02	0.02
b_fingertip	0.02	0	0	0.6	0	0.05	0	0	0	1	0	0	0	0	0
d_fingertip	0.18	0.52	0.66	0	0	0	0.7	0.29	0.42	0	1	0.15	0	0	0
a_fingertip	0.69	0.49	0.38	0	0	0	0.46	0.02	0.02	0	0.15	1	0	0	0
b_centerline	0	0	0	0	0	0	0	0	0	0	0	0	1	0.01	0.01
d_centerline	0	0	0	0	0	0	0	0.06	0.02	0	0	0	0.01	1	1
a_centerline	0	0	0	0	0	0	0	0.06	0.02	0	0	0	0.01	1	1
R_b_top	0.58	0.14	0.11	0.1	0.26	0.18	0.14	0.01	0.01	0.11	0.05	0.27	0	0	0
R_d_top	0.25	0.73	0.89	0	0	0	0.91	0.15	0.21	0	0.71	0.22	0	0	0
R_a_top	0.18	0.5	0.62	0	0	0	0.66	0.4	0.55	0	0.91	0.18	0	0.01	0.01
R_b_wrist	0.11	0	0	0.35	0.92	0.66	0.01	0	0	0.38	0	0	0	0	0
R_d_wrist	0.94	0.24	0.18	0	0	0	0.26	0.01	0	0	0.05	0.58	0	0	0
R_a_wrist	0.2	0	0	0	0.12	0.03	0.01	0	0	0	0	0	0	0	0
R_b_edge	0.15	0.01	0	0.35	0.81	0.6	0.01	0	0	0.37	0	0.01	0	0	0
R_d_edge	0.06	0.19	0.28	0	0	0	0.36	0.6	0.84	0	0.51	0.02	0	0.01	0.01
R_a_edge	0.09	0.25	0.35	0	0	0	0.41	0.6	0.83	0	0.58	0.05	0	0.01	0.01
R_b_fingertip	0.02	0	0	0.32	0	0.01	0	0	0	0.16	0	0	0	0	0
R_d_fingertip	0.3	0.86	0.97	0	0	0	0.97	0.1	0.14	0	0.56	0.29	0	0	0
R_a_fingertip	0.69	0.49	0.39	0	0	0	0.46	0.02	0.02	0	0.15	1	0	0	0
R_b_centerline	0.02	0.04	0.05	0	0	0	0.05	0.22	0.15	0	0.07	0.01	0.09	0.89	0.89

R_d_centerline	0.11	0.33	0.44	0	0	0	0.5	0.46	0.65	0	0.72	0.07	0	0.01	0.01
R_a_centerline	0.01	0.02	0.03	0	0	0	0.04	0.37	0.2	0	0.05	0	0	0.34	0.34

	R_b_top	R_d_top	R_a_top	R_b_wrist	R_d_wrist	R_a_wrist	R_b_edge	R_d_edge	R_a_edge	R_b_fingertip	R_d_fingertip	R_a_fingertip	R_b_centerline	R_d_centerline	R_a_centerline
b_top	0.58	0.25	0.18	0.11	0.94	0.2	0.15	0.06	0.09	0.02	0.3	0.69	0.02	0.11	0.01
d_top	0.14	0.73	0.5	0	0.24	0	0.01	0.19	0.25	0	0.86	0.49	0.04	0.33	0.02
a_top	0.11	0.89	0.62	0	0.18	0	0	0.28	0.35	0	0.97	0.39	0.05	0.44	0.03
b_wrist	0.1	0	0	0.35	0	0	0.35	0	0	0.32	0	0	0	0	0
d_wrist	0.26	0	0	0.92	0	0.12	0.81	0	0	0	0	0	0	0	0
a_wrist	0.18	0	0	0.66	0	0.03	0.6	0	0	0.01	0	0	0	0	0
b_edge	0.14	0.91	0.66	0.01	0.26	0.01	0.01	0.36	0.41	0	0.97	0.46	0.05	0.5	0.04
d_edge	0.01	0.15	0.4	0	0.01	0	0	0.6	0.6	0	0.1	0.02	0.22	0.46	0.37
a_edge	0.01	0.21	0.55	0	0	0	0	0.84	0.83	0	0.14	0.02	0.15	0.65	0.2
b_fingertip	0.11	0	0	0.38	0	0	0.37	0	0	0.16	0	0	0	0	0
d_fingertip	0.05	0.71	0.91	0	0.05	0	0	0.51	0.58	0	0.56	0.15	0.07	0.72	0.05
a_fingertip	0.27	0.22	0.18	0	0.58	0	0.01	0.02	0.05	0	0.29	1	0.01	0.07	0
b_centerline	0	0	0	0	0	0	0	0	0	0	0	0	0.04	0	0
d_centerline	0	0	0.01	0	0	0	0	0.01	0.01	0	0	0	0.89	0.01	0.34
a_centerline	0	0	0.01	0	0	0	0	0.01	0.01	0	0	0	0.89	0.01	0.34
R_b_top	1	0.07	0.06	0.35	0.43	0.58	0.42	0.01	0.02	0.1	0.08	0.27	0.01	0.03	0
R_d_top	0.07	1	0.66	0	0.07	0	0	0.26	0.34	0	0.83	0.23	0.04	0.45	0.02
R_a_top	0.06	0.66	1	0	0.07	0	0	0.65	0.7	0	0.54	0.18	0.09	0.84	0.09
R_b_wrist	0.35	0	0	1	0.01	0.47	0.91	0	0	0.32	0	0	0	0	0
R_d_wrist	0.43	0.07	0.07	0.01	1	0.01	0.03	0	0.01	0	0.09	0.58	0.01	0.02	0
R_a_wrist	0.58	0	0	0.47	0.01	1	0.62	0	0	0	0	0	0	0	0
R_b_edge	0.42	0	0	0.91	0.03	0.62	1	0	0	0.32	0	0.01	0	0	0
R_d_edge	0.01	0.26	0.65	0	0	0	0	1	0.97	0	0.16	0.02	0.12	0.78	0.13
R_a_edge	0.02	0.34	0.7	0	0.01	0	0	0.97	1	0	0.25	0.05	0.12	0.83	0.15
R_b_fingertip	0.1	0	0	0.32	0	0	0.32	0	0		0	0	0	0	0
R_d_fingertip	0.08	0.83	0.54	0	0.09	0	0	0.16	0.25	0	1	0.29	0.04	0.33	0.01
R_a_fingertip	0.27	0.23	0.18	0	0.58	0	0.01	0.02	0.05	0	0.29	1	0.01	0.07	0
R_b_centerline	0.01	0.04	0.09	0	0.01	0	0	0.12	0.12	0	0.04	0.01	1	0.1	0.49
R_d_centerline	0.03	0.45	0.84	0	0.02	0	0	0.78	0.83	0	0.33	0.07	0.1	1	0.1
R_a_centerline	0	0.02	0.09	0	0	0	0	0.13	0.15	0	0.01	0	0.49	0.1	1