Visual Cues Used to Evaluate Grasps from Images*

Matthew Sundberg[1], Walter Litwinczyk[2], Cindy Grimm[3] and Ravi Balasubramanian[3]

Abstract— **We analyze visual cues people used to evaluate a robot grasp. Participants were presented with two (front and side) orthogonal views of a robot hand grasping an object and asked how successful the grasp would be on a scale of 1-5; they were eye-tracked while completing this survey. Ground truth of the success of the grasps is known. Our primary observations were that 1) Most of the failed grasp predictions were false positives, and this was exacerbated for grasps that were ranked as human-like. 2) Two visual cues from human-grasp research (object center-line and top) were used, but not contact points. Instead, participants gazed at robot finger, wrist, and arm locations. 3) There was a difference in the visual patterns between the left and right images, indicating that the second image was primarily used to verify the locations of fingers and wrist while the first was used to establish the object's location and shape. Finally, we generate transition matrices to model the temporal aspect of the gaze patterns.**

I. INTRODUCTION

Humans are adept at grasping, whether using their own hands or tele-operating a robot hand. This skill arises from a (largely unconscious) set of complex visual and proprioceptive evaluations humans perform while doing physical interaction tasks. Specifically, when performing a grasp, humans scan the object and the environment to evaluate where and how to grasp the object for that task, then use proprioceptive and visual feedback to fine tune the grasp. Teasing out which visual and proprioceptive cues are used is challenging; however, an understanding of these cues can help both with designing interfaces for humans and learning which cues to use in automatic algorithms. In this paper we use eye gaze data from participants evaluating images of grasps to determine which features are critical for grasp evaluation.

Our long-term goal is to use information gained from crowd sourcing (so-called "human intelligence tasks") to train robots to perform grasping and manipulation tasks robustly. This information is also pertinent to designing better human-robot interaction interfaces. Specifically, understanding the visual cues that people use is a necessary first step to presenting them with visual information that both increases accuracy and reduces evaluation time. This is particularly important when using just images or video.

Allowing a person to physically manipulate a robot hand and arm to perform a grasp produces highly reliable data in terms of robust example grasps [1]. Unfortunately, this method of data collection does not scale because it requires the human to be physically present and because it takes a non-trivial amount of time to move the robot hand. On the other hand, crowd-sourcing — eg humans voting on images or videos of grasps — produces a lot of data quickly. The flip side is, the data may be incorrect. These incorrect evaluations arise both because of the participant's lack of familiarity with the robot hand's tactile properties (for instance, how "grippy" the hand is in terms of fingertip friction and compliance) and because of incorrect or missing visual cues. To overcome these limitations we need to understand where and when on-line evaluations may produce incorrect evaluations. We take a first step in this direction by analyzing which visual cues people focus on when using two orthogonal views of a robot hand to evaluate a grasp and where the evaluations go wrong.

Specifically, we use photos of grasps of objects for which we have ground truth information on whether or not the grasp succeeded on a physical robot [1]. The objects were chosen to represent an array of household objects, such as a cracker box, water pitcher, and soda can. The grasps were characterized by a variation in the number of fingers, palm contact, and overall orientation used in the grasps. The robot arm was photographed from two orthogonal viewpoints for each grasp-object combination, and the photos were presented side-by-side. Participants were asked to rank the grasps on the likelihood of success (1-5) and how human-like (yes/no) the grasp was. The images were presented as a survey, and the participants's eye gaze was tracked while they took the survey.

With the goal of identifying which object features humans focus on during grasping, we divided the images into regions (see Figure 1) and analyzed fixation patterns within those regions. In addition, we measured the visual saliency [2] of the images (essentially which areas are likely to draw someone's gaze). We compared the fixation patterns to the underlying saliency to determine if participants simply focused on regions of high saliency (they didn't). We analyzed how well the participants' grasp scores correlate with ground truth and where humans go wrong. We identified a pattern of false-positives (that is, humans said the grasp would succeed when it didn't). This was particularly true for grasps the participants labeled as human-like. We also analyzed fixation patterns to determine both what visual cues were important and when. This has potential use in analyzing camera views for new grasp-object pairs.

Contributions: Our primary contribution is a systematic evaluation of visual cues used when evaluating images of grasps. We also identify a bias in human-based scoring that can be adjusted for by asking the participants if the gaze is

*Supported in part by NSF Grant CNS 1359480, REU Site: Robots in the Real World, and CAREER award IIS-0952631

[1]Corvallis, Oregon
[2]Rochester Institute of Technology
[3]Oregon State University
Cindy.Grimm, Ravi.Balasubramanian@OregonState.edu

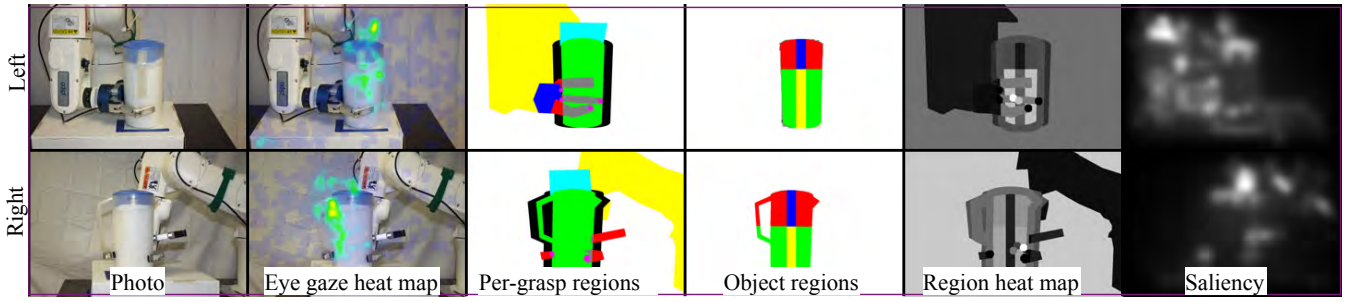

Fig. 1. From left to right: The original images, heat map of eye gaze (all participants), image regions used for analysis, regions colored by number of fixations, and visual saliency maps. Top row is left image, bottom row is right image in the survey.

human-like.

II. RELATED WORK

There is a strong need to improve the ability of robots to physically interact with the environment. While there has been significant progress in the domain of robotic grasping and manipulation both in terms of hardware [3], [4], [5] and software development [6], [7], [8], [9], more work is required. 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; that is, one in four grasps fail [10]. The primary reason for this poor performance is that robot grasps are not robust enough; that is, small differences in object shape or object position cause the object to, say, slip out during the grasping process. There has been significant effort to address these issues using physics-based heuristics and brute-force search algorithms to find more robust grasps with mixed success [**?**].

Prior work has also explored "learning from demonstration", where humans teach robots [11], [12], to advance robot performance. However, most previous approaches for gathering data are time-intensive [13]. In prior work, we have used crowd-sourcing where we have employed images or video of the grasps to receive human input. That work showed that humans are likely to over-estimate how successful the grasp will be. Despite this over-estimate, for some grasps humans are still more accurate than learning approaches based on standard grasp metrics (for instance, center of grasp, center of mass). 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 [10]. However, no prior work has studied human eye gaze when controlling a robot arm in a physical interaction task.

There is a growing body of prior work on where humans look when performing grasps using their own hands [14], [15], [16]. The work showed that 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 [15] or asking the participants to do the grasp from memory [16] changed the ratios of which region was gazed at, and in what order, but did not substantially change the types of regions.

III. STUDY DESIGN AND METHODOLOGY

In this section we outline the basic setup, study design, and participant pool. We eye-tracked participants using a 60Hz SMI remote eye-tracker while sitting in front of a standard 24 inch monitor in a well-lit room. Participants were told that they would be taking a survey which would last approximately 15 minutes. We calibrated the eye-tracker then brought up the survey window at full monitor size.

A. Stimuli

To create the photos used in the survey the objects were placed at the same location on a box, oriented toward the camera (see blue tape in Figure 1). This placed the bottom of the object in the same place in all views. The robot arm and hand were automatically moved to the specified grasp configuration and photos taken from two orthogonal cameras in fixed locations. The arm, hand, and object were fully visible in all views (except for self-occlusions). Image resolution in the survey was 770 by 512 (original images were 3888 by 2592). The photos were placed side-by-side in the survey, with the main axis view always on the left. The question order ('Does this grasp look natural, similar to what a human would use?' [naturalness] and 'Will the robot be able to securely pick up the object?' [score]) was randomized, as was the presentation of the grasps. We used yes or no for the naturalness question and a 1–5 Likert scale [17] for the score question.

There were 9 objects (see Fig. 2) with 8 grasps each (characterized by two or three fingers, palm or no palm, and power versus precision — not all combinations are possible) for a total of 72 combinations. The object set included a cracker box, water pitcher, water bottle, sanitizer, CD case, remote control, wine glass, coil of wire, and soda can. Each participant saw 30 of the 72 combinations, presented in random order, to prevent fatigue. Each grasp had been evaluated previously by shaking it 10 times [1]. The grasp success rate (number between zero and one) is the number of times it remained in the hand, divided by 10. 25 grasps failed, 37 grasps succeeded, and the remaining were partially

Fig. 2. Objects used in the study.

successful. For analysis that required either successful or not labeling we labeled all grasps with a score bigger than 0.7 as successful.

B. Participants

We recruited 26 students, primarily from engineering, to take the study. All participants had normal, or corrected to normal, vision. Average time for the participants to take the survey was 14.5 minutes. This resulted in approximately 10 evaluations per grasp.

IV. ANALYSIS

In this section we describe how we analyzed the raw eye-tracking data and the image data using saliency measures. The SMI eye tracker yielded the coordinates of where the participant was looking on the monitor at a 60 Hz sampling rate. We analyzed this raw data both by reducing it to a sequence of fixations and determining the fixation location within regions identified (manually) in the images.

A. Fixations

We used the EyeMMV fixation detection algorithm [18], which filters the coordinate sequences by applying a threshold of dispersion to the points. We used standard settings [19] for the algorithm: a 100 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 $1/2$ DVA. The algorithm produces a sequence of fixations, with each fixation centered at the average of the coordinates and lasting a given duration. Using fixations compared to the raw coordinate data both reduces processing time and removes saccades, where the viewer is essentially blind.

We employed a two-step process to determine where (semantically) in each image the fixation occurred. In the first step we partitioned the full monitor image into the left and right images, the question area (below the images) and other (browser bars, clock, etc). We then further divided the individual grasp images into semantically meaningful regions (which may overlap). These regions were stored as color-coded "mask" images. Because the object was placed in the same location in the image for all grasps for that object, we were able to create a single object-only mask. The hand and arm, however, required manually creating a mask for each grasp for each object for each camera view (see Figure 1). The object and grasp masks were merged during analysis.

The areas defined by the grasp-specific mask were background, arm, wrist, object, and finger locations. The object masks defined the object's center line, silhouette, top, and bottom (anything not marked as one of those is labeled as 'object'). Additionally, we hand-placed markers denoting all contact points for each grasp (including inferred ones occluded by the object). These were used to create circular regions indicating the contact areas. Note that which item was occluded (hand or object) could always be determined from the grasp-specific mask.

We considered a fixation to be within a mask region if a circle with a radius of $1/2$ DVA centered at the fixation coordinates surpassed a threshold of 50% of the total number of pixels in the fixation. Each fixation was labeled with the one (or more) regions it overlapped.

B. Saliency

We used a graph-based saliency approach [2]. Saliency is, in general, problematic since it depends not only on low-level visual cues such as contrast but also on objects (such as faces) that have semantic meaning. In our case we are looking for low-level visual cues that might attract the viewer's gaze, so an image-based saliency measure is appropriate. When calculating the saliency measure for a fixation we average the saliency values within a circle of radius $1/2$ DVA centered at the fixation.

C. Glance counts and time

We additionally compress fixations into glances, where all consecutive fixations that lie within one of the three areas (left image, right image, questions, ignoring other) are gathered into a single glance. On average there are 32 fixations for each evaluated grasp, with an average of 12 glances. We also measure time from the gaze data. Glance count, number of fixations, and time spent are all strongly correlated ($r = 0.63$ to $r = 0.86$), but can vary substantially between participants. For this reason, when we are comparing participant-specific data we normalize by dividing by the participant's average for that data (time or number of fixations or glance count).

D. Accuracy

The physical shake test score is the number of successful shake tests divided by the number of tests (zero to one in intervals of 0.1). The survey score was a range from 1 to 5, with 1 being unsuccessful. We normalized this score to the range zero to one. Accuracy is defined as the absolute difference between the two, and ranges from zero to one.

E. Visualizations

The eye tracker produces several visualizations of the eye gaze over time; we primarily used SMI's default heat map visualization (see accompanying video). We created two types of heat maps for each grasp (see Figure 1). The first is a traditional gaze heat map (combining gaze data for

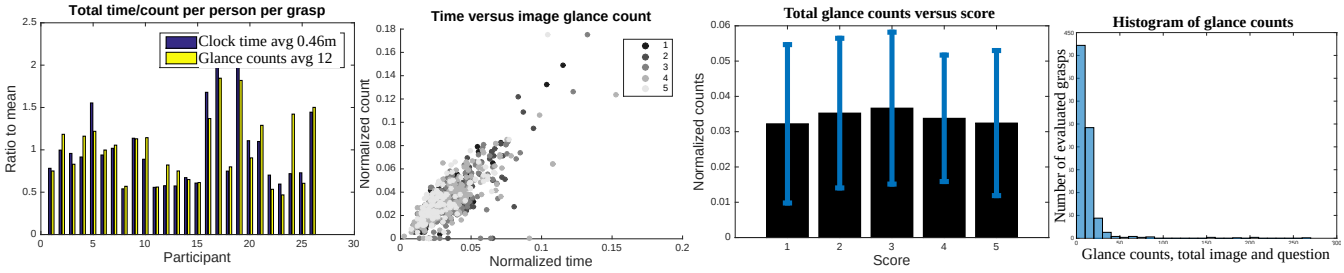

Fig. 3. From left to right: Time and glance count variation per participant, correlation of time and glance count, time versus uncertainty (with 95% confidence intervals), and histogram of glance counts per image.

each participant), the second is a *region*-based heat map. Both of these were used to qualitatively assess the data. The traditional heat map was also used to quantify the correlation between the saliency and gaze location.

V. RESULTS

We present results in four categories: Analysis of failed grasp predictions, visual cues used, differences between left and right image gaze patterns, and gaze patterns over time.

A. Failed grasp predictions

In this section we discuss overall relationships between how the participants ranked the grasps, how long it took to rank them, and comparison to the ground-truth physical test scores. There was no correlation between individual participant's accuracy and time, and average accuracy for the participants was fairly uniformly distributed between 0.27 and 0.43.

Time spent, Figure 3: Our first analysis focuses on how long the participants focused on the left and right images, and how many times they glanced back and forth between them and to the questions. On average participants spent 30 seconds looking at each survey page, of which 20 seconds was spent actually looking at the images (and not the questions or outside the survey). Gaze time is strongly correlated with fixations ($r = 0.77$). and fixations with glances ($r = 0.86$). On average participants made 12 unique glances (minimum 1, maximum 266), and 32 unique fixations (minimum 3, maximum 714) per image. The distribution is not equal, but highly skewed by five people who took substantially longer. Over 50% of the images were viewed using fewer than 9 glances. This corresponds approximately to looking at one image, the other, the questions, then repeating that pattern once (6 glances) or twice (9 glances).

There was a general pattern of high-confidence scores (1 and 5) taking a shorter amount of time than more uncertain scores (2-4) (see Figure 3, right), however this was a very weak correlation ($r = 0.16$). There was a stronger correlation between the grasp score and the standard deviation in score across the participants ($r = -0.6$). Grasps with low scores showed more agreement than grasps with high scores.

Incorrect scoring, Figure 4: The average score for the grasps was 3.2, and approximately half (46%) were labeled as human-like or natural (see Figure 4, left). There was a general pattern of false-positives, particularly for grasps labeled as human. If we set the threshold score difference as 0.4, there are 15 false positives and 11 false negatives; however, the biggest difference for false negatives is 0.61 versus 0.8 for false positives, and 7 of the 8 worst ones are false positives.

The grasps that were labeled as false-positives lacked the caging property; for these the object slipped out of the grasp. We hypothesize that the participants were more likely to label these grasps as successful *because* they looked human. For unfamiliar grasps they were more parsimonious. Although further research is required to verify this, in practice discounting or lowering scores for grasps which are both labeled as human and rely on friction is advisable.

B. Visual cues

In this section we discuss the visual cues used to evaluate grasps, and how they relate to saliency values. Perception, eye-gaze, visual saliency, and cognition are all inter-related in complex ways; broadly speaking, though, fixations are caused by a mix of low-level visual cues such as high-contrast edges (saliency) and top-down task requirements (eg, count the number of dots in the image). Complicating this is the fact that, while the fixation *location* can be reasonably well-determined, *what* the user was attending to could be ambiguous. For example, in our case the fingers and the object often occlude each other; in these cases it is not clear if the user is paying attention to the contact area, the location of a hidden finger, the object, or some combination thereof. For fixation analysis we conservatively assume that the user could be gazing at any of the overlapped regions.

Refer to Figure 5. Participants spent approximately $10-15\%$ of their time looking at the background, $30-35\%$ looking at the robot hand and arm, and $30-40\%$ at the object [1]. Nearly all the background fixations were on at least mildly salient areas (> 0.1). Although a cloth was draped as a backdrop there were still high-salient edges on the box, the blue tape, and the cord (in some images). The robot fixations were primarily on the fingers, but participants also spent about $1/4$ of the robot-viewing time looking at the wrist and arm. The wrist position has been found to be a useful grasp

[1]Numbers do not add up to one because some fixations are double-counted.

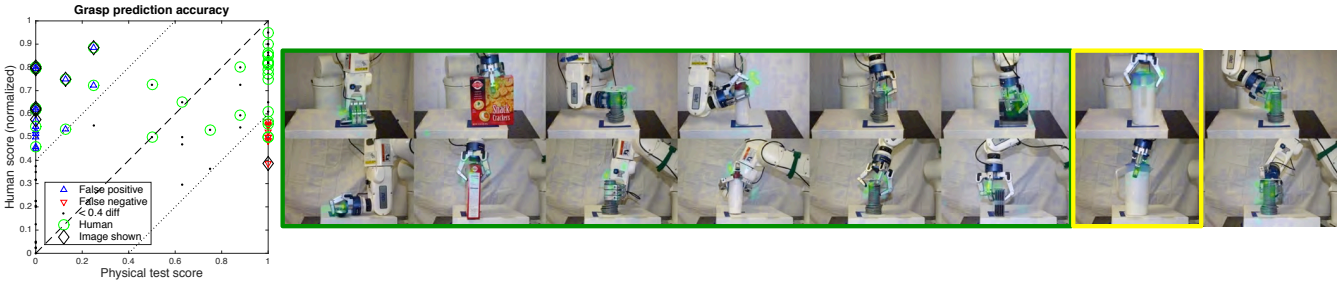

Fig. 4. Left: Difference between physical grasp score and human grasp score (normalized to 0-1). Right: Heat maps for grasps with the biggest difference in scores. The 6 worst (outlined in green) are marked as human, the seventh image (outlined in yellow) is the only false-negative.

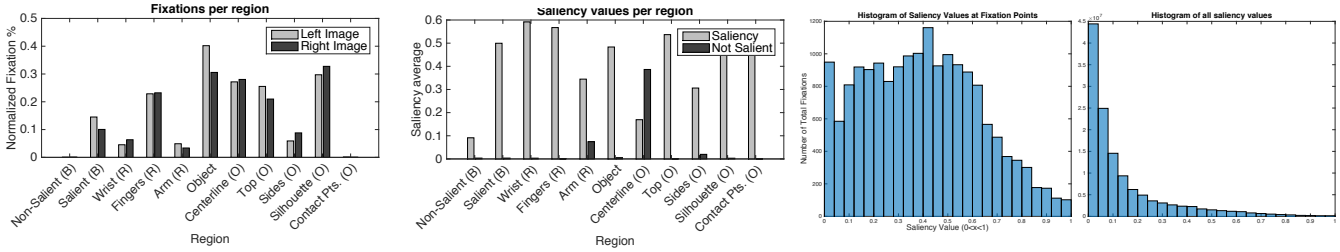

Fig. 5. From left to right: Fixation counts per region (normalized by number of fixations). Regions may overlap (eg, object and finger). Saliency values of fixations per region and percentage of which are not salient. Histogram of saliency values over the fixation regions and the entire image set.

predictor [1]. The object fixations are fairly evenly split over the center line, the top, and the silhouette. The former two correspond to visual cues used in human grasping; we did not find much evidence for fixations at potential contact points. We hypothesize that the silhouette fixations are instead used to determine how the fingers are wrapped around the object.

For the most part participants looked at visually salient portions of the image (average 0.4). The correlation between the salient portions of the image and the heat maps for eye gaze (limited to non-background regions) is reasonably high (mean $r = 0.46$), so in this case low-level cues and task-based fixations align. Two regions (arm 23% and object center line 50%) accounted for nearly all of the non-salient fixations. As in the human grasping literature, the object center line appears to be important enough in analyzing a grasp to overcome any lack of visual saliency.

C. Left-right gaze patterns

In this section we look at the differences in eye gaze between the left and right images. We assume the natural gaze pattern is the reading one — from left to right, top to bottom (none of our participants spoke languages where this was not true). We did not vary the presentation order of the images; our subject pool was not large enough to support any statistically significant findings from doing so.

When first viewing the images the participants mostly looked at the left image (68%), then the right image (21%) or the questions (11%). The most common initial viewing pattern was left image, right image (55%). The next most common viewing pattern was right image, left image (20%). This follows the general pattern of reading left to right, top to bottom. The per-grasp patterns were not consistent across participants, though, so the variation is likely due to differences in the participants and not inherent in the visual data.

There were substantially more fixations on the object and salient background features in the left image (see Figure 5 - note that the fixation counts have been normalized for total number of fixations on the left or right). The first fixation was on the object (44%), background (29%), or fingers (27%). The right image saw more fixations on the sides and silhouette of the object. We hypothesize that participants first gained an understanding of the scene and object from the left image, and then used the right image to verify finger and wrist placement relative to the object.

D. Temporal analysis

In this section we analyze gaze patterns over time (see Figure 6). The most noticeable pattern, as mentioned before, is the left-right-question one, but there are subtler patterns. We define five temporal groups: the first, second, and third glance, shifting of glances after the second glance, and within a glance. We additionally examined the fixations by three larger categories (hand, object, background).

We created transition matrices from *all* of the fixation and glance data. For the left-right-question transitions we used the glance data; the other ones used the fixation data. Recall that the regions overlap; we assume a transition if the fixation overlaps that region. Transition tables are included in the appendix.

Figure 6, left, shows the left-right-question pattern. The left image bias is also present in the transitions (far right

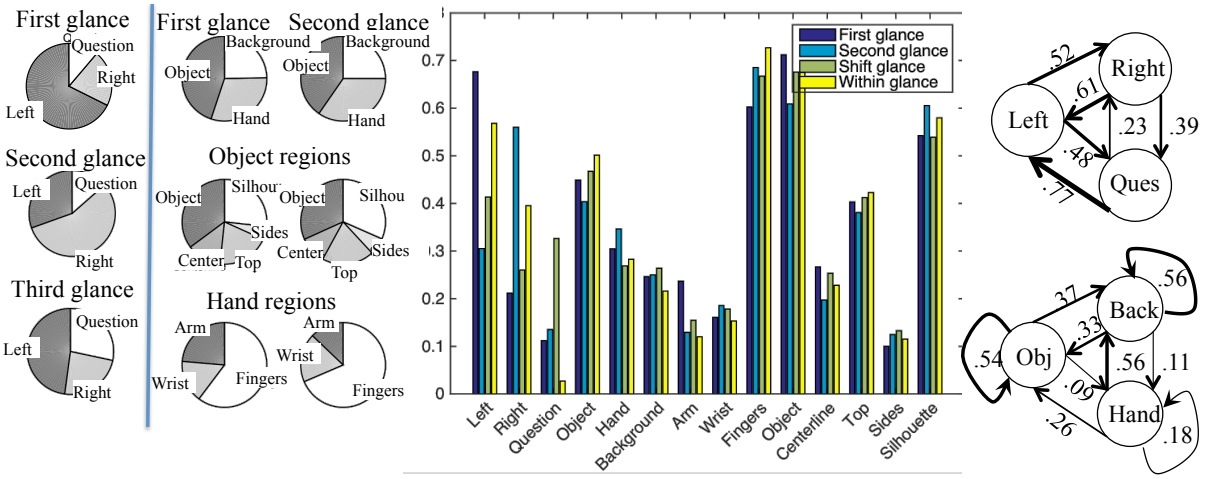

Fig. 6. From left to right: First three glances distribution (left versus right image or questions). First and second glance distribution for the object, hand, and background (subsequent fixations had similar distributions). Differences between fixation distributions between the first glance, the second, shifting glances, and within a glance. Transition diagrams.

Figure 6).

There was a tendency to glance first at the object, then at the hand. The most common transitions within this group were object to object, background to background, and hand to background. One possibility is simply that many of the background fixations are in areas of moderate saliency. An alternative is that the relatively large number of hand to background transitions arise because the participant is looking for the matching hand location in the other image (but mentally rotated the scene incorrectly). Supporting this is that approximately 18% of the hand or object transitions to the background also involved a shift between the two images.

Within the hand transitions there was a tendency to look at the arm in the first glance, and the fingers in the second and subsequent glances. This may be because the arm occupies a large portion of the image, so low-level cues draw the gaze to the arm, after which the participant focuses on the more relevant fingers.

Within the object transitions the object center and top fixations decreased after the first glance, while the side and silhouette ones increased.

VI. RECOMMENDATIONS

The following are recommendations for presenting images of grasps based on the results of this study. First, it is worth asking how human-like the grasp is in addition to if the grasp will succeed in order to account for the bias for human-like grasps succeeding (particularly when the grasp is not enclosing). Second, ensure that high-saliency regions are located only at relevant points — i.e., use a simple, neutral color background that is different than the robot's hand/arm color, and remove anything (such as marks and cords) on the arm or in the background that might draw their attention. Third, marking the fingers in some way (such as placing green dots on one of them) and using two camera angles that are *less* than ninety degrees would substantially reduce the need to mentally map from one viewpoint to the other. Forth, place the image with the most useful information (number of contact points, silhouettes, etc) on the left to reduce overall viewing time.

VII. CONCLUSIONS

We have analyzed eye-tracking data from a survey asking participants to rank robotic hand grasps using two orthogonal views. Our results suggest that participants used a mix of the visual cues that they would normally use in order to determine a grasp (except contact points) and additional cues to locate the robot hand. Visual saliency clearly plays a role in gaze dwell time; ensuring that only the hand, object, and shadows are salient might reduce evaluation times. Fixation transition information can be used to both pick views that are useful and to determine the number of views needed to disambiguate the grasp.

There was a pattern of false positives for human-like grasps; this implies participants may be better at determining if a grasp is one they would use themselves versus the quality of the grasp for a robotic hand. Regardless, asking if the grasp is human-like is useful for identifying this situation.

Future work: There are several questions this study does not answer: How well can humans rank *human* grasps? Would videos or different image views result in better evaluations?

APPENDIX (TRANSITION TABLES)

	Left	Right	Question
Left	0	0.52	0.48
Right	0.61	0	0.39
Question	0.77	0.23	0

	Object	Hand	Background
Object	**0.54**	0.09	0.37
Hand	0.26	0.18	**0.56**
Background	0.33	0.11	**0.56**

	Arm	Wrist	Fingers
Arm	**0.66**	0.25	0.09
Wrist	0.32	**0.62**	0.05
Fingers	**0.54**	0.22	0.25

	Object	C-line*	Top	Sides	Silh*
Object	0.20	0.27	0.25	0.10	0.18
C-line*	0.05	0.29	**0.32**	0.11	0.23
Top	0.03	0.17	**0.41**	0.14	0.26
Sides	0.02	0.16	**0.40**	0.18	0.25
Silh*	0.03	0.17	**0.38**	0.13	0.29

*Centerline; Silhouette

ACKNOWLEDGMENT

We would like to acknowledge Dr. Reynold Bailey for his generosity in loaning both his eye tracker and his student Walter for this study. We thank the participants of our study for their time and patience. We thank the Saturday Academy's Apprenticeships in Science and Engineering (ASE) program (`http://www.saturdayacademy.org/ase`) for sponsoring Matthew to work in the lab for the summer.

REFERENCES

[1] A. Goins, R. Carpenter, W.-K. Wong, and R. Balasubramanian, "Evaluating the efficacy of grasp metrics for utilization in a gaussian process-based grasp predictor," in *Intelligent Robots and Systems (IROS 2014)*, Sept 2014, pp. 3353–3360.

[2] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. MIT Press, 2007, pp. 545–552.

[3] A. M. Dollar and R. D. Howe, "The highly adaptive SDM Hand: Design and performance evaluation,," *Internat. J. Robotics Res*, vol. 29, no. 5, pp. 585–597, 2010.

[4] L. Birglen, T. Laliberté, and C. Gosselin, *Underactuated Robotic Hands*. Springer, 2008.

[5] E. Brown, N. Rodenberg, J. Amend, A. Mozeika, E. Steltz, M. R. Zakin, H. Lipson, and H. M. Jaeger, "Universal robotic gripper based on the jamming of granular material," *Proceedings of the National Academy of Sciences*, vol. 107, no. 44, pp. 18 809–18 814, 2010.

[6] A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision," *Int. J. Robotics Res.*, vol. 27, no. 2, pp. 157–173, 2008.

[7] E. Lopez-Damian, D. Sidobre, and R. Alami, "A grasp planner based on inertial properties," in *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*. IEEE, 2005, pp. 754–759.

[8] B. León, S. Ulbrich, R. Diankov, G. Puche, M. Przybylski, A. Morales, T. Asfour, S. Moisio, J. Bohg, J. Kuffner *et al.*, "Opengrasp: a toolkit for robot grasping simulation," in *Simulation, Modeling, and Programming for Autonomous Robots*. Springer, 2010, pp. 109–120.

[9] S. Chitta, I. Sucan, and S. Cousins, "Moveit!" *IEEE Robotics Automation Magazine*, vol. 19, no. 1, pp. 18–19, 2012.

[10] R. Balasubramanian, L. Xu, P. Brook, J. R. Smith, and Y. Matsuoka., "Physical human interactive guidance: A simple method to study human grasping." *IEEE Transactions on Robotics*, 2012, dOI: 10.1109/TRO.2012.2189498. (In press).

[11] S. Ekvall and D. Kragic, "Interactive grasp learning based on human demonstration," in *Robotics and Automation (ICRA)*, vol. 4, April 2004, pp. 3519–3524 Vol.4.

[12] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and Autonomous Systems*, vol. 57, no. 5, pp. 469 – 483, 2009.

[13] R. Balasubramanian, L. Xu, P. Brook, J. Smith, and Y. Matsuoka, "Human-guided grasp measures improve grasp robustness on physical robot," in *Robotics and Automation (ICRA)*, May 2010, pp. 2294–2301.

[14] J. Lawrence, K. Abhari, S. Prime, B. Meek, L. Desanghere, L. Baugh, and J. Marotta, "A novel integrative method for analyzing eye and hand behaviour during reaching and grasping in an mri environment," *Behavior Research Methods*, vol. 43, no. 2, pp. 399–408, 2011.

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

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

[17] T. J. Maurer and H. R. Pierce, "A comparison of likert scale and traditional measures of self-efficacy." *Journal of applied psychology*, vol. 83, no. 2, p. 324, 1998.

[18] 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," *Journal of Eye Movement Research*, vol. 7, no. 1, pp. 1–10, 2014.

[19] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *ETRA 2000*. New York, NY, USA: ACM, 2000, pp. 71–78.