Investigating the Effects of Visual Saliency on Deictic Gesture Production by a Humanoid Robot

Aaron St. Clair, Ross Mead, and Maja J Matarić, *Fellow, IEEE*

***Abstract*—In many collocated human-robot interaction scenarios, robots are required to accurately and unambiguously indicate an object or point of interest in the environment. Realistic, cluttered environments containing many visually salient targets can present a challenge for the observer of such pointing behavior. In this paper, we describe an experiment and results detailing the effects of visual saliency and pointing modality on human perceptual accuracy of a robot's deictic gestures (head and arm pointing) and compare the results to the perception of human pointing.**

I. Introduction

To carry on sustained interactions with people, autonomous robots must be able to effectively and naturally communicate in many different interaction contexts. Besides natural language, people employ both coverbal modalities (e.g., beat gestures) and nonverbal modalities (facial expression, proxemics, eye gaze, head orientation, and arm gestures, among others) to signal their intentions and to attribute intentions to the actions of others. Prior work has demonstrated that robots can successfully employ the same communication channels [1], [2], [3], [4]. Our aim is to develop a general, empirical understanding of design factors involved in multimodal communication with a robot. In this paper, we limit our focus to a study of deictic gestures since: (1) their use in human communication has been widely studied [6], [7], [8]; (2) they are relatively simple to map to intentional constructs in context [10]; and (3) they are generally useful to robots interacting in shared environments since they serve to focus attention and refer to objects. To achieve robust deixis via gesture in a human-robot context, it is necessary to validate the perceived referent. This paper presents results from an experimental study of human perception of a robot's deictic gestures under a set of different environmental visual saliency conditions and pointing modalities using our upper-torso humanoid robot, Bandit.

II. Embodied Deictic Gesture

Multi-disciplinary research from neuroscience and psychology has demonstrated that human gesture production is tightly coupled with language processing and production [11], [12]. There is also evidence that gestures are adapted by a speaker to account for the relative position of a listener and can, in some instances, substitute for speech functions [8], [10]. Bangerter [8] and Louwerse & Bangerter [10] demonstrated that deictic speech combined with deictic gesture offered no additional performance gain compared to one or the other used separately.

These findings have important implications for the field of human-robot interaction. Robots interacting with humans in a shared physical environment should be able take advantage of other social channels to both monitor and communicate intent during the course of an interaction, without complete reliance on speech. To make this possible, it is necessary to gain an empirical understanding of how to map well-studied human gestures to robots of varying capabilities and embodiments. Specifically, we are interested in identifying variables for proper production of robot gestures to best realize some fixed interpretation by a human observer. In general, this is difficult, for the same reasons that processing natural language is difficult: many gestures are context-dependent and rely on accurately estimating a mental model of the scope of attention and possible intentions for people's actions given only low-level perceptual input.

Deictic gestures, however, are largely consistent in their mapping to linguistic constructs, such as "that" and "there", and serve to focus the attention of observers to a specific object or location in the environment, or perhaps to indicate an intended effect involving such an object (e.g., "I will pick up that one"). These characteristics, while simplifying their interpretation and production, also make the gestures useful for referring to objects and grounding attention. Intentional analysis and timing are still nontrivial, except in the context of performing a specific pre-determined task.

Both recognition [13], [14], [15], [16], [17] and production [18], [19], [20], [21] of deictic gestures have been studied in human-human, human-computer, and human-robot interaction (HRI) settings. Our work adds to this field a step toward obtaining an empirically grounded HRI model of deictic gestural accuracy between people and robots, with implications for the design of robot embodiments and control systems that perform situated distal pointing. A study of the literature on human deictic pointing behavior suggests a number of possible variables that could potentially affect the robustness of referent

Manuscript received February 1, 2011. This work was supported in part by National Science Foundation (NSF) grants CNS- 0709296, IIS-0803565, and IIS-0713697 and ONR MURI grant N00014-09-1-1031. R. Mead was supported by an NSF Graduate Research Fellowship.

A. St. Clair (corresponding author), R. Mead, and M. J. Matarić are with the Computer Science Department at the University of Southern California, Los Angeles, CA 90089-0781 USA (e-mail: astclair@usc.edu; rossmead@usc.edu mataric@usc.edu).

selection when having a robot employ deictic gestures including physical appearance [25], timing, relative position, and orientation [6], [22], [23]. Assuming mobility, a robot could relocate or reorient itself relative to the viewer or to the referent target to improve the viewer's interpretation accuracy [24]. Most robots, however, point without accounting for distance to the target, by using a more or less constant arm extension or head gaze that is reoriented appropriately [18], [5]. Finally, since a gesture is grounded with respect to a specific referent in the environment, the robot must be able to correctly segment and localize visually salient objects in the environment at a similar granularity to the people with whom it is interacting. Studies of human reading of robot gaze [31], as well as biologically inspired methods to assess and map visually salient features and objects in an environment, exist [26], [27], as do models of human visual attention selection [28]; however, the role of visual saliency during deictic reference by a robot is largely uninvestigated.

III. Experimental Design

Given the large number of possible variables involved in optimizing a pointing gesture with a particular robot embodiment, we conducted an initial pilot study with our upper-torso humanoid robot, Bandit (Figure 1) sitting face-to-face with participants and gesturing to locations on a transparent screen between them. We varied distance and angle to target, distance and angle to viewer, and pointing modality, but as no strong correlations emerged in early testing, so we narrowed the set of conditions and hypotheses.

A. *Hypotheses*

We conducted a factorized experiment over three robot pointing modalities: the head with 2 degrees-of-freedom (DOF), the arm with 7 DOF, and both together (i.e., head+arm) with two saliency conditions: a blank (or non-salient) environment and an environment with several highly and equally visually salient targets. Since our results, particularly the modality condition, may be specific to Bandit, we also conducted a similar, but smaller, test with a human performing the pointing gestures, for comparison.

1) *Modality*

The conditions tested include head (Figure 2a), away-from-body, straight arm (Figure 2b), cross-body, bent arm (Figure 2c), and combined head and arm (Figure 2d). We hypothesized that the arm modality would lead to more accurate perception since, when fully extended, it is most expressive and easily interpreted as a vector from the robot to the screen. Since Bandit's head does not have moveable eyes, the point of reference is somewhat ambiguous and could lead to pointing error. Our kinematic calculations solved for the midpoint of the eyes to align to the target; this information was not shared with experiment participants. Additionally, we expected to see an effect between away-from-body, straight arm (Figure 1b) points that occur on one side of the screen and the cross-body bent-arm (Figure 1c) gesture used for the other side, with the bent-arm being more difficult to interpret, since it is staged in front of the robot's body rather then laterally [22], [23], [24]. A similar effect was seen in human pointing [9], showing that people are capable of estimating vectors accurately from body pose. Finally, we hypothesized that using both modalities together would reduce error relative to a single modality, since participants would have two gestures on which to base the estimate.

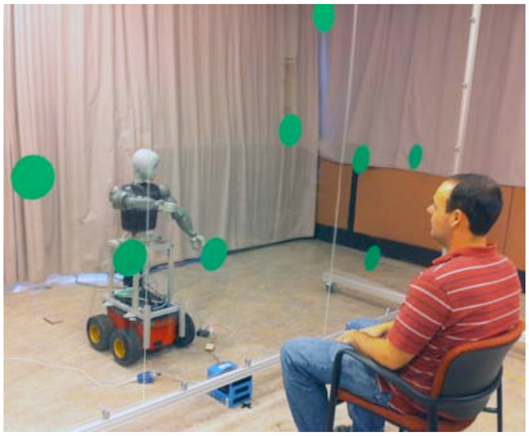

Figure 1. The experimental setup, with Bandit indicating a cross-body point to the study participant in the foreground.

2) *Saliency*

For the two saliency conditions, we hypothesized that the salient objects would affect people's interpretations of the points. Specifically, we anticipated that people would "snap to" the salient objects, thus reducing error for points whose targets were on or near markers, whereas in the non-salient condition there were no points of reference to bias estimates. We did not expect to see any difference in the performance of each pointing modality when comparing the salient and non-salient conditions.

3) *Implementation*

In the experiments, the participant is seated facing Bandit at a distance of 6 feet (1.8 meters). The robot and the

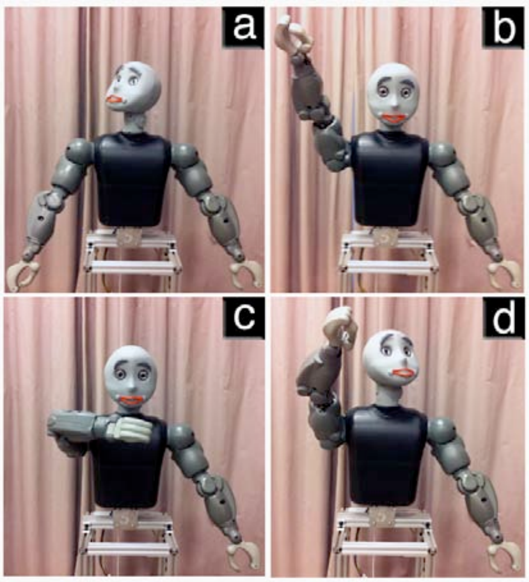

Figure 2. (a) Bandit pointing with its head; (b) straight-arm; (c), bent-arm; and (d) and head+arm.

participant are separated by a transparent, acrylic screen measuring 12 feet by 8 feet (2.4 by 3.6 meters) (see Figure 1). The screen covers a horizontal field of view from approximately -60 to 60 degrees and a vertical field of view -45 to 60 degrees. The robot performs a series of deictic gestures and the participant is asked to estimate their referent location on the screen. The robot is posed using a closed-form inverse kinematic solution; however, a small firmware "dead-band" in each joint sometimes introduces error in reaching a desired pose. To monitor where the robot actually pointed, we computed forward kinematics using angles from encoder feedback, which we verified were accurate in separate controlled testing. All gestures were static and held indefinitely until the participant estimated a location, after which the robot returned to a home location (looking straight forward with its hands at its sides) before performing the next gesture. Participants were given a laser pointer to mark their estimated location for each gesture. These locations were recorded using a laser rangefinder placed facing upwards at the base of the screen. For each gesture, an experimenter placed a fiducial marker over the indicated location, which was subsequently localized within approximately 1 cm using the rangefinder data. The entire experiment was controlled via a single Nintendo Wiimote™, with which the experimenter could record marked locations and advance the robot to point to the next referent target.

The face-to-face nature of the experiment was chosen intentionally although other work in gesture perception [8] has tested human deictic pointing accuracy when the pointer and the observer were situated more or less side-by-side, observing a scene. In our work, and in most HRI settings, the robot is facing the participant; side-by-side interaction, in terms of proxemics, is more likely to occur when coordinated motion and interaction are concurrent (e.g., the robot and human walking together while an interaction is taking place [29]). Our design tests the face-to-face scenario that is more applicable to the types of proxemic HRI configurations we have encountered.

3) *Modality*

The robot gestured to locations by moving its head, arm, or both together. A single arm was used, resulting in cross-body, bent-arm gestures for points on one side of the screen and away-from-body straight-arm gestures for points on the other side of the screen. The arm was not modified for pointing and, thus, the end-effector was simply the 1-DOF gripper in a closed position. We presume that using a pointer-like object for an end effector would increase accuracy since Bandit's hand has several sloped surfaces that make estimation challenging, but our goal was to establish baseline measures for unmodified hardware. All the gestures were static, meaning the robot left the home position, reached the gesture position, and held it indefinitely until the participant chose a point, after which it returned to the home position for the next gesture. This was intended to minimize any possible timing effects by giving participants as long as they needed to estimate a given gesture. Participants were also asked to only turn on the laser pointer after they had visually selected an estimated (perceived) target location, to prevent them from using the laser pointer to line up the robot arm and head with the actual target location.

4) *Saliency*

The screen itself was presented with two visual saliency conditions: one in which it was completely empty (i.e., non-salient) and one in which it was affixed with eight round markers distributed at random (i.e., salient). In the salient case, the markers were all 6 inches (15 cm) in diameter and were identical in shape and color. Experiments were conducted in two phases, reflecting the two saliency conditions. In the salient condition, the robot's gestures included 60 points toward salient targets and 60 points chosen to be on the screen, but not necessarily at a salient target. In the non-salient condition, 74 of the points within the bounds of the screen were chosen pseudo-randomly and the remaining 36 were chosen to sample a set of 4 calibration locations with each pointing modality. Randomization was performed such that each participant in a given condition saw the same set of points. The calibration points were used to assess the consistency and normality of the error in the robot's actual pointing and the participant's perception to determine whether a between-subjects comparison was possible. All three pointing modalities (head-only, arm-only, and head+arm) were used in both the salient and non-salient cases.

5) *Human Pointing*

The human-human pointing condition was conducted by replacing Bandit with an experimenter with the intention of anecdotally comparing robot pointing with typical human pointing in the same scenario. Since people point by aligning a chosen end-effector with the referent target using their dominant eye [8], conducting the experiment with a human pointer introduces the confound in that people cannot point with the arm-only modality. We also found the head-only modality difficult to measure accurately and, for this reason, only the head+arm modality was tested, which consequently conveys eye gaze. These experiments were conducted by replacing the robot with an experimenter who held a Nintendo Wiimote™ in his or her non-pointing hand. Two different vibration patterns signaled whether to point to a target location or a location selected arbitrarily. The experimenter pointed with a clenched fist grip similar to Bandit's, while holding a laser pointer concealed in the palm of the hand, and held the pose. A second experimenter then marked both the participant's and experimenter's points, as before. As with the robot, the human pointer only pointed with the right hand.

6) *Surveys*

In addition to the pointing task, we also administered a survey asking participants to estimate their average error with respect to modality and location on the screen and to rate each modality on a Likert scale in terms of preference. The surveys also collected background information such as handedness and level of prior experience with robots.

IV. Results

A. Participants and Data

A total of 40 runs of the experiment were conducted as described, with 20 (12 female, 8 male) participating in the non-salient condition and 20 (11 female, 9 male) participants in the salient condition. In total, around 4500 points were estimated and recorded. The conditions were close to equally weighted with the exception of the non-salient arm-only and head+arm conditions, which was done initially to allow for comparison of the cross-body versus away-from-body arm gestures. The number of points collected for each condition is presented in Table 1. Participants were recruited from on-campus sources and all were undergraduate or graduate students at USC from various majors. The participants were roughly age- and gender-matched with an average age of 20. The data collected for each run included the desired target on the screen (i.e., the desired location the robot should have pointed to), the actual target on the screen for each modality (i.e., the location the robot actually pointed), and the perceived point as indicated by the participant and recorded by the laser rangefinder. We also captured timing data for each point and video of the sessions taken from a camera mounted behind and to the side of the robot.

B. Perceived Error Analysis

We conducted a two-way analysis of variance (Type III SS ANOVA) of various angular error measures with modality and saliency as the independent factors. Both were found to have significant effects on the angular error between perceived and desired target points as well as on perceived and actual target points (Table 2). Additionally, the interaction effects between the modality and saliency factors were found not to be significant. Mean angular error computed from the perspective of the person and confidence intervals are shown in Figures 3-5. We used angular error as a metric to effectively normalize for different distances to target. For comparison purposes, human perceptual error when estimating human pointing gestures (arm or eye gaze) has been measured to be approximately 2-3 degrees for people up to 2.7 meters apart [8], [30]. We conducted post-hoc analysis using the Tukey's honestly significant differences (HSD), which revealed that mean error tends to be on average 1 degree larger for arm points with $p<0.01$, and that using both modalities tends to outperform the arm modality in most cases with the means differing by 1.1 degrees in the salient case and 0.8 degrees in the non-salient case with significance of $p<0.01$. The arm alone, however, was equally poor in both saliency conditions.

To compare cross-body bent-arm versus away-from-body straight-arm gesture, we looked at arm points in the non-salient case. Partitioning them into two sets, depending on the side of the screen they were on, resulted in 667 cross-body points and 649 straight-arm points. Conducting a one-way ANOVA with arm gesture type as the dependent variable, we found a significant difference between the straight-arm case (M=5.85 degrees, SD=3.9) and the bent-arm case (M=10.2 degrees, SD=7.1) with $p<0.0001$ (see Figure 6). We also obtained similar results when conducting a 3-way ANOVA between modality, saliency, and binned screen location, although there was a significant interaction with screen location likely due to differences in the arm appearance and accuracy between the straight arm and bent-arm case as screen location varied..

TABLE I
DATA COUNTS PER EXPERIMENTAL CONDITION

	Head	Arm	Both
Non-salient	565	1175	298
Salient	800	801	811

Arm is over-represented in the non-salient condition to compare away-from-body and cross-body

TABLE 2
F AND P VALUES FOR 2-WAY ANOVA

Condition	Perceived-Desired	Perceived-Actual
Saliency	F=3.2, p<0.07	F=4.7, p<0.02
Modality	F=8.4, p<0.0002	F=5.0, p<0.006

To assess whether the accuracy of the other modalities varied with angle to target, we first fit a linear regression model with angular error as the dependent variable and the desired target as the independent variable. The resulting model did not perform well upon cross validation, suggesting that the error was nonlinear in nature. To cope with the nonlinearity, we binned points by angle to target in 9 uniform intervals covering the extent of the screen. We then performed an n-way ANOVA with target x and y coordinates as a conditional factor, and found that the head-only and head+arm conditions were significantly better ($p<0.001$) in the center of the screen with error increasing about halfway to the edge before leveling out. These effects were largely symmetric for the head-only and head+arm conditions. The arm-only condition, as described above, was asymmetric and was significantly worse in almost all cases except for the middle of the left side of the screen corresponding to away-from-body straight-arm points. All modalities tended to result in more erroneous estimates in the lower extreme of the screen. Finally, the average time taken to estimate each point was nearly a second faster in the non-salient case (M=5.6, SD=2.3) than in the salient case (M=6.6, SD=2.1). This effect was found to be significant with $p<0.0001$, while there was no significant difference for the modality conditions (Figure 7).

C. Human Pointing

For the human pointing phase of the experiment, we collected a total of 70 data points from 2 participants. While this is not a considerable amount of data and is not meant for drawing definitive conclusions, we did find a significant ($p<0.09$) effect between the two saliency conditions. For the salient condition (M=2.23, SD=2.46), the error was small enough that both the pointer and the observer were able to "hit" all the salient targets, while the non-salient condition (M=3.74, SD=2.74) resulted in a 60% increase in error. We also observed that points directed at the center of the screen appeared to result in lower perception error than points directed between the center and the periphery. Overall, the error in estimating human-produced points appears to have a similar profile to that of the robot-produced points; however, more investigation is necessary.

D. Survey Responses

In the responses to the survey, participants in the non-salient condition estimated that their points were within an average of 28 centimeters (11 inches); this is very close to the mean error of 27 centimeters we found in practice. There was no significant difference between participants' estimated error when comparing across the two conditions. Pointing with the head-only and with the head+arm were preferred by the majority of the participants, with only 6 (or 15%) stating a preference for the arm modality. When asked if there was a noticeable difference in straight-arm points versus the bent-arm, 68% said there was, with the remainder not seeing a difference. Fourteen out of 20 (or 67%) of the participants in the salient condition said that the markers would have an effect on their estimate of the referent target.

V. Discussion

A. Visual Saliency

The mean error, as computed (using the perceived point and the desired target points), tells us how close to a desired target (either a randomly chosen one in the non-salient case or one of the markers in the salient case) the robot was actually able to indicate. The performance of the head-only, arm-only and head+arm in the salient condition is improved by approximately 1 degree. This suggests that the snap-to-target effect that we expected to see when salient objects were introduced is modest, resulting in a best-case improvement of approximately 1 degree. This is also seen when we consider the mean perceived-actual error, which is slightly lower for nearly every condition. This suggests that participants estimate the referent to be closer to the actual point the robot is physically indicating than the nearest salient object. This could be a useful property because it allows us to consider pointing without having to assess scene saliency beforehand. That is, if the referential target of a point has not been specified a priori, through some other means such as verbal communication or previous activity, people tend to evaluate the point in an ad hoc manner by taking a guess. When disambiguating referents, if there are unknown salient objects in the environment, their effects on the perception of a given gesture can be expected to be small enough in most cases that a precise point to the actual target should suffice to communicate the referent.

B. Modality

As we hypothesized, the modalities did result in different pointing accuracy profiles. When considering modalities, pointing with the head+arm does appear to perform appreciably better than either the arm-only or the head-only, in most cases. One possible explanation is that head+arm more closely emulates typical human pointing, in which

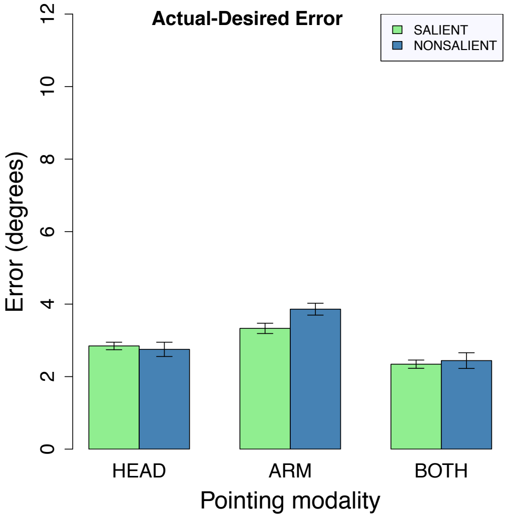

Figure 3. Mean angular pointing error.

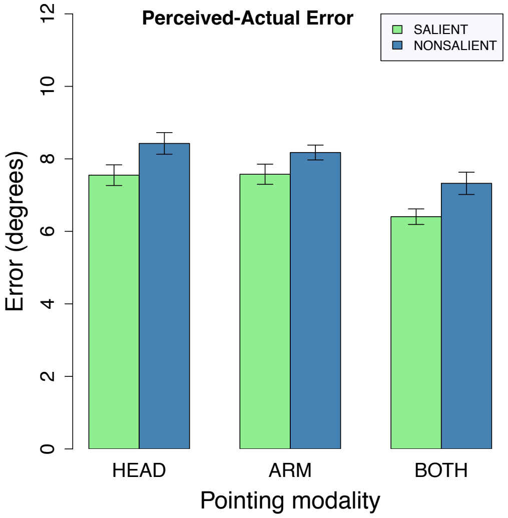

Figure 4. Mean angular error between perceived and actual targets.

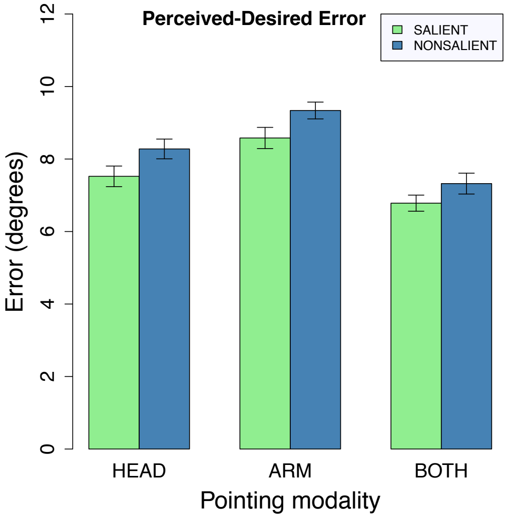

Figure 5. Mean angular error between perceived and desired targets.

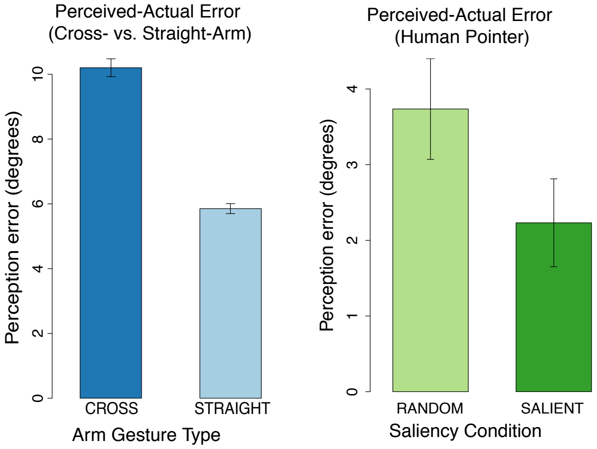

Figure 6. Mean angular error for straight vs. bent-arm and mean error by saliency condition with human pointer.

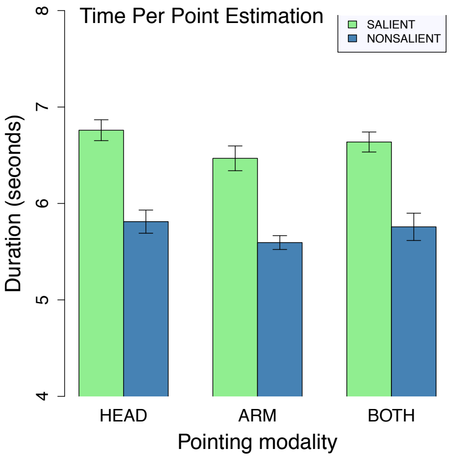

Figure 7. Mean time from start of pose to marking of estimated point.

people tend to align an end-effector with their dominant eye [8]; another is that multiple modalities provide more diverse cues that indicate the referential target resulting in better priming of the viewer to interpret the gesture. The poor performance of the arm in the salient condition was somewhat unexpected. This might be due to its higher actual error compared to the head. Another source of the error could be the use of the cross-body arm gesture, which, while equally weighted, resulted in nearly twice the perceptual error compared to the away-from-body arm. This might be a result of the reduced length of the arm, which forces people to estimate the vector based on only the forearm versus the entire arm as in the away-from-body case. Another explanation is that the gesture is staged against the body, that is, with minimal silhouette and is thus more difficult to see. In either case, roughly one-third of the participants did not notice a difference in the arm gestures while their performance was, in fact, affected. This illustrates the impact that gesture and embodiment design can have on interpretation, and underscores the need to validate gestural meaning with people.

When considering the horizontal and vertical target position analyses, we see that people are best at estimating points directly between themselves and the robot. Performance then drops off when the target is located laterally, above, or below. This effect could be due to a field-of-view restriction, preventing the viewer from seeing both the robot's gesture and the target at the same time in high acuity, foveal vision. Estimating these points then requires the viewer to saccade the head between the two points of interest. We believe the slight improvement at the far periphery for some of the modalities is due to the fact that we informed participants that the points would be on the screen, thus creating a bound for points near the screen edges.

C. *Human Pointing*

The results of our smaller-scale investigation of human pointing did find that the salient condition resulted in approximately 1.5 degrees less error than in the non-salient condition, which is consistent with our finding using the robot pointer. Also, the 2-degree perceptual accuracy that we found when testing a human pointer seems to agree with prior studies of humans conducted in relevant literature. It is also worthy of note that, although the deictic pointing performance of the robot is several times worse than what we saw in the human experiment or would expect from the literature, we can use the estimate of our resolving power (i.e., the minimum angle between referents that we could hope to convey) to inform controller design and ensure that the robot repositions itself or gets close enough to prevent these effects. The salient condition also resulted in a 16% increase (or approximately 1 second) in time needed to estimate the gesture. This is intuitive, as the participants were presented with more stimuli in the form of the salient objects and, thus, took some extra time to ground the point, possibly checking to see if it is coincident with any objects first. This information could be useful in developing methods for effective timing control.

VI. Future Work

One obvious next step is to conduct a similar experiment with a robot of a different embodiment to evaluate whether the same general conclusions hold true or if they are tightly coupled to the specific appearance of Bandit. Since the project was developed using Willow Garage's Robot Operating System (ROS), substituting different robots into the experiment will involve minimal changes to the codebase. We are currently developing a deictic gesture (pointing) package for the PR2 robot and others that have an URDF specification, and plan on using it to run the experiment with the PR2. Formal studies of other relevant variables (such as angle to target and timing), as well as comparable studies with human pointing (our results were only anecdotal in nature), are also necessary to develop a better understanding of human perception of robot deictic gestures.

We also plan to analyze head movements as participants perform the task. We observed a saccade effect in which people transitioned back and forth between an estimated point and the robot before finally glancing at the experimenter to mark the point. This is similar in nature to the regressive eye movements used to measure gesture clarity in [8].

We are seeking ways to automatically measure estimated points, thereby allowing us to remove the potentially-distracting experimenter from the room. Finally, we plan on using the presented data to construct a parameterized error model to allow Bandit to perform effective deixis to objects in a mapped environment.

VII. Conclusion

In this paper, we presented the results of a study of human perception of robot gestures intended to test whether visual saliency and embodied pointing modality have an effect on the performance of human referent resolution. Our results suggest that environmental saliency, when employing deictic gesture alone to indicate a target, results in only a modest bias effect. We also demonstrated that pointing with two combined and synchronized modalities, such as head+arm, consistently outperforms one or the other individually. Additionally, we found that the physical instantiation of the gesture (i.e., how it is presented to the observer) can have drastic effects on perceptual accuracy, as noted in comparing bent-arm and straight-arm performance.

Acknowledgment

The authors would like to thank Mary Ermitanio, Hieu Minh Nguyen and Karie Lau for their help with data collection.

References

[1] B. Scassellati, “Investigating models of social development using a humanoid robot,” vol. 4, pp. 2704 – 2709 vol.4, Jul. 2003.

[2] A. Brooks and C. Breazeal, “Working with robots and objects: Revisiting deictic reference for achieving spatial common ground,” in *Proc of the 1st Conf on Human-robot Interaction*, p. 304, ACM, 2006.

[3] C. Breazeal, C. Kidd, A. Thormaz, G. Hoffman, and M. Merlin, “Effects of nonverbal communication on efficiency and robustness in human-robot teamwork, ieee/rsj int,” in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS2005)*, pp. 383–389, 2005.

[4] C. Sidner, C. Kidd, C. Lee, and N. Lesh, “Where to look: a study of human-robot engagement,” in *Proceedings of the 9th International Conference on Intelligent User Interfaces*, p. 84, ACM, 2004.

[5] B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, and N. Hagita, “Footing in human-robot conversations: how robots might shape participant roles using gaze cues,” in *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction, (HRI’09)*, 2009.

[6] A. Ozyurek, “Do speakers design their co-speech gestures for their addresees? the effects of addressee location on representational gestures,” *Journal of Memory and Language*, vol. 46, no. 4, pp. 688–704, 2002.

[7] N. Nishitani, M. Schurmann, K. Amunts, and R. Hari, “Broca’s region: From action to language,” *Physiology*, vol. 20, no. 1, p. 60, 2005.

[8] A. Bangerter, “Accuracy in detecting referents of pointing gestures unaccompanied by language,” Gesture, vol. 6, no. 1, pp. 85–102, 2006.

[9] A. Bangerter, “Using pointing and describing to achieve joint focus of attention in dialogue,” *Psychological Sci.*, vol. 15, no. 6, p. 415, 2004.

[10] M. Louwerse and A. Bangerter, “Focusing attention with deictic gestures and linguistic expressions,” in *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*, 2005.

[11] R. Mayberry and J. Jaques, Gesture production during stuttered speech: Insights into the nature of gesture-speech integration, ch. 10, pp. 199–214. Cambridge University Press, 2000.

[12] S. Kelly, A. Ozurek, and E. Maris, “Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension,” *Psychological Science*, vol. 21, no. 2, pp. 260–267, 2009.

[13] R. Cipolla and N. Hollinghurst, “Human-robot interface by pointing with uncalibrated stereo vision,” *Image and Vision Computing*, vol. 14, no. 3, pp. 171–178, 1996.

[14] D. Kortenkamp, E. Huber, and R. Bonasso, “Recognizing and interpreting gestures on a mobile robot,” in *Proceedings of the National Conference on Artificial Intelligence*, pp. 915–921, 1996.

[15] K. Nickel and R. Stiefelhagen, “Visual recognition of pointing gestures for human-robot interaction,” *Image and Vision Computing*, vol. 25, no. 12, pp. 1875–1884, 2007.

[16] P. Pook and D. Ballard, “Deictic human/robot interaction,” *Robotics and Autonomous Systems*, vol. 18, no. 1-2, pp. 259–269, 1996.

[17] N. Wong and C. Gutwin, “Where are you pointing? : the accuracy of deictic pointing in cves,” in *Proc. of the 28th Intl. Conference on Human Factors in Computing Systems.*, pp. 1029–1038, ACM, 2010.

[18] M. Marjanovic, B. Scassellati, and M. Williamson, “Self-taught visually guided pointing for a humanoid robot,” in *From Animals to Animats 4: Proc. of the 4th Intl, Conf, Simulation of Adaptive Behavior*, pp. 35–44.

[19] J. Trafton, N. Cassimatis, M. Bugajska, D. Brock, F. Mintz, and A. Schultz, “Enabling effective human–robot interaction using perspective-taking in robots,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 35, no. 4, pp. 460–470, 2005.

[20] O. Sugiyama, T. Kanda, M. Imai, H. Ishiguro, N. Hagita, and Y. Anzai, “Humanlike conversation with gestures and verbal cues based on a three-layer attention-drawing model,” *Connection science*, vol. 18, no. 4, pp. 379–402, 2006.

[21] Y. Hato, S. Satake, T. Kanda, M. Imai, and N. Hagita, “Pointing to space: modeling of deictic interaction referring to regions,” in *Proc. of the 5th ACM/IEEE Intl. Conf. on Human-Robot Interaction*, pp. 301–308, ACM, 2010.

[22] F. Thomas and O. Johnston, *The Illusion of Life: Disney Animation*. Hyperion, 1981.

[23] J. Lasseter, “Principles of traditional animation applied to 3D computer animation,” in *ACM Computer Graphics*, vol. 21, no. 4, pp. 35-44, July 1987.

[24] R. Mead and M.J. Matarić, “Automated caricature of robot expressions in socially assistive human-robot interaction, “ in *The 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI2010) Workshop on What Do Collaborations with the Arts Have to Say about HRI?*, Osaka, Japan, March 2010.

[25] A. Bangerter, “Using pointing and describing to achieve joint focus of attention in dialogue,” *Psychological ScI.*, vol. 15, no. 6, p. 415, 2004.

[26] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Transactions on Pattern Analysis and Machine Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[27] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch, “Attentional selection for object recognition: a gentle way,” in *Biologically Motivated Comp. Vision*, pp. 251–267, Springer, 2010.

[28] R. Desimone and J. Duncan, “Neural mechanisms of selective visual attention,” *Annual review of neuroscience*, vol. 18, no. 1, pp. 193–222, 1995.

[29] G. Butterworth and S. Itakura, “How the eyes, head and hand serve definite reference,” *British Journal of Developmental Psychology*, vol. 18, no. 1, pp. 25–50, 2000.

[30] R. Mead, “Space: a social frontier,” poster presented at the *Workshop on Predictive Models of Human Communication Dynamics*, Los Angeles, California, August 2010.

[31] F. Delaunay, J. de Greeff, and T. Belpaeme, “A study of a retro-projected robotic face and its effectiveness for gaze reading by humans,” in *Proc. of the 5th ACM/IEEE Intl Conf. on Human-Robot Interaction*, pp. 39–44, ACM, 2010.