Investigating Confidence-Based Category Transition of Spatial Gestures

Adam Stogsdill
astogsdill@mymail.mines.edu
Colorado School of Mines
Golden, Colorado, US

Thao Phung
thaophung@mymail.mines.edu
Colorado School of Mines
Golden, Colorado, US

Tom Williams
twilliams@.mines.edu
Colorado School of Mines
Golden, Colorado, US

ABSTRACT

Situated human-human communication typically involves a combination of both natural language and gesture, especially *deictic* gestures intended to draw the listener's attention to target referents. To engage in natural communication, robots must thus be similarly enabled not only to generate natural language, but to generate the appropriate gestures to accompany that language. In this work, we examine the gestures humans use to accompany spatial language, specifically the way that these gestures continuously degrade in specificity and then discretely transition into non-deictic gestural forms along with decreasing confidence in referent location. We then outline a research plan in which we propose to use data collected through our study of this transition to design more human-like gestures for language-capable robots.

1 BACKGROUND AND MOTIVATION

Research has demonstrated that nonverbal communication channels are critical for both human-human and human-robot interactions. Previous works show that children develop their verbal language skills much later than their non-verbal language skills, and that truly natural language-based communication fluidly combines these two communication styles [4]. If the hope of robotics is to make robots integrate further into human society, then robots will need to be able to communicate just as humans do, including mastery of humanlike gesture [9].

HRI researchers have previously enabled robots to use a variety of forms of human gestures, including deictic [10], beat [1], iconic [3], and metaphoric gestures [3]. However there has been little research on how robot gestures may need to naturally transition between these categories. As an illustrative example, we consider the case of gestures accompanying spatial language. When describing an object nearby to themselves, speakers will typically use *deictic gestures* that pick out the object in their user's field of view; they may even use deictic gestures to refer to objects nearby but not strictly visible, such as an adjacent room. But at a certain distance (either literal or cognitive) speakers abruptly transition to other forms of gesture, such as emblematic gestures that allow them to simply indicate that they are referring to some distal entity, or metaphoric and iconic gestures, which allow them to describe the position of that distal entity or to communicate how to travel to it.

For HRI researchers to enable robots to generate appropriate gestures to accompany spatial language, they must thus answer three key questions: (1) What are the factors that determine the form of gesture used to accompany spatial language? (2) At what parametrization of these factors do speakers switch *between* these categories? (3) How does the parametrization of these factors inform the performance of gestures *within* these categories? In this study we present preliminary results from a human-subject study designed to provide the data needed to answer these questions.

2 A THEORY OF CATEGORY TRANSITION FOR SPATIAL GESTURES

We hypothesize a number of factors that impact the gestures used to accompany spatial language:

Physical Distance: As the distance to an entity increases, we expect the probability of deictic gesture use to decrease.

Visibility: We expect deictic gesture use to be more probable for visible than not-currently-visible entities.

Confidence of Speaker in Location: As a speaker's confidence in the location of an entity increases, we expect the probability of deictic gesture use to increase.

Expected Confidence of Listener in Location: As a listener's expected confidence in the location of an entity increases, we expect the probability of deictic gesture use to decrease.

Certainty of Joint Attention: As a speaker's certainty that they and their interlocutor are already jointly attending to the intended entity, we expect the probability of deictic gesture use to decrease.

Furthermore, we hypothesize that the precision of a speaker's gaze and gesture will decrease as the speakers' context approaches the decision boundary at which they would switch to non-deictic forms of gesture: for an entity far from this decision point, we expect the speaker to point and gaze precisely and sustainedly, whereas for an entity close to this decision point, we expect the speaker merely to briefly point and glance in the general direction of the target.

In this paper we present preliminary results from a paper designed to assess a subset of these hypotheses, with particular emphasis placed on physical distance, visibility, and confidence of a speaker in their target referent's location.

3 EXPERIMENTAL PROCEDURE

11 Participants were recruited from a college campus through web postings and fliers. Upon arriving at the laboratory and providing informed consent, participants were given a list of 21 objects and locations that experimenters could assume to be familiar to participants (such as visible objects in the laboratory environment, campus landmarks, and nearby and distant towns and cities), with these objects and locations presented in ascending order of distance to the participant. For each object or location on the list, participants were asked to indicate their confidence in the location of that item or location using a ten-point Likert item from Not Confident At All

to Very Confident. After completing this confidence survey, participants were seated in a chair across from an experimenter, behind whom was positioned a Microsoft Kinect configured to track and log their skeletal data. The participant was then asked to proceed through the list of objects and locations again, and to communicate the position of each entity in the list. In total, this produced a set of 21 audio recordings and time-series of joint positions for each of the 11 participants.

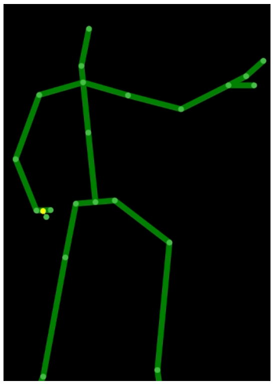

Figure 1: Kinect Skeletal Tracking data collected from a participant

4 PRELIMINARY RESULTS

4.1 Gesture Category Transition Boundary

While we are currently in the process of completing our experiment and have yet to quantitatively analyze our data, we are able to provide some qualitative summarizations of our data collected to date. Specifically, our data seem to suggest that the gestures used by participants were nearly entirely deictic in nature for objects within the visible reaches of the experimental environment; for objects outside this range, deictic gestures continued to be used but decreased in frequency in proportion to distance, being replaced by other types of gestures, such as metaphoric gestures. One notable exception to this is when participants chose to describe a location with respect to its cardinal direction, in which case speakers typically reverted to deictic gestures. However, this exception may not generalize to speakers in other locations due to the unique geographic landscape of our university campus, which allows for easy identification of cardinal directions at all times.

4.2 Data Provided via Route Description

When participants were sufficiently uncertain of an object or location's position, they typically tried to describe that position through reference to a landmark in whose position they *were* more confident, or by giving step-by-step route instructions. These cases provide additional data we did not intend to collect, as we can examine the type of gesture used to accompany the description of each landmark or waypoint in this unsolicited description. Once the experiment is completed, we intend to conduct not only analysis of participants' gestures towards the locations they were asked to describe, but also analysis of these supplementary locations they they decided to describe and gesture towards through their own volition.

4.3 Confidence Survey

While the confidence survey was intended to measure participants' confidence in the precise location of target objects, participants' results suggest that they in fact interpreted confidence-in-location differently depending on the distance to the object or location. As an example, four of the 11 participants provided a confidence score of 10 for New York City (the most distant location on the survey) even though they could not give an accurate description as to the city's location. This suggests that while for nearby objects and locations participants may have assessed their confidence with respect to the metric location relative to their current position, for far-off locations participants may have assessed their confidence with respect to the geographical location relative to the United States.

5 FUTURE WORK

Once our experiment is completed, we plan to perform statistical analyses to address the key research questions proposed earlier in this paper. To do so, we will begin by quantifying the most probable parameter-space boundary at which users transfer from deictic to non-deictic forms of gesture.

We will then use the collected joint data to learn separate models for gestures that fall on each side of this decision boundary (i.e., for deictic and non-deictic gestures to accompany spatial language). Specifically, we plan to convert our collected joint data into images that can be fed into a Gestural Generative Adversarial Network (GAN) [2] to learn a gesture space that mimics (rather than directly represents) the gestures used by humans in our experiment.

These models should then allow a robot to generate an uniquely appropriate and human-like gesture using the appropriate GAN's generator module [7]; an approach that we expect to be particularly successful due to GAN models' previous success in generating realistic looking outputs under context shift. However, because the traditional GAN model architecture will not provide enough context for the model to always generate appropriate and meaningful gestures [5], we instead intend to provide this context based on the features of the target referent as defined in Section 2 (cf. [6]).We want to keep this strictly informed by just these variables that we observed in the experiment. Adding too many inputs can make it harder for the GAN model to generalize well and would require more understanding about the inputs.

After modifying our GAN model in this way, we plan to conduct additional human-robot interaction experiments to analyze the success of this model. Specifically, we intend to compare robot gestures generated using this GAN model to gestures hand-designed by HRI researchers, using both subjective measures such as human-likeness and naturalness, and objective measures such as ease of referent identification by the robot's interlocutors [8].

ACKNOWLEDGMENTS

This work was funded in part by grant IIS-1909864 from the National Science Foundation.

REFERENCES

[1] Paul Bremner, Anthony G Pipe, Mike Fraser, Sriram Subramanian, and Chris Melhuish. 2009. Beat gesture generation rules for human-robot interaction. In *Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 1029–1034.

[2] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *NeurIPS*.

[3] Chien-Ming Huang and Bilge Mutlu. 2013. Modeling and Evaluating Narrative Gestures for Humanlike Robots.. In *Robotics: Science and Systems*. 57–64.

[4] Polychronis Kondaxakis, Khurram Gulzar, Stefan Kinauer, Iasonas Kokkinos, and Ville Kyrki. 2018. Robot–Robot Gesturing for Anchoring Representations. *IEEE Transactions on Robotics* PP (10 2018), 1–15. https://doi.org/10.1109/TRO.2018.2875388

[5] Andrew Kusiak. 2019. Convolutional and generative adversarial neural networks in manufacturing. *International Journal of Production Research* (2019). https://doi.org/10.1080/00207543.2019.1662133 arXiv:https://doi.org/10.1080/00207543.2019.1662133

[6] Mehdi Mirza and Simon Osindero. 2014. Conditional Generative Adversarial Nets. *CoRR* abs/1411.1784 (2014). arXiv:1411.1784 http://arxiv.org/abs/1411.1784

[7] Igor Rodriguez, José María Martínez-Otzeta, Itziar Irigoien, and Elena Lazkano. 2019. Spontaneous talking gestures using Generative Adversarial Networks. *Robotics and Autonomous Systems* 114 (2019), 57 – 65. https://doi.org/10.1016/j.robot.2018.11.024

[8] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2013. To Err is Human(-like): Effects of Robot Gesture on Perceived Anthropomorphism and Likability. *International Journal of Social Robotics* 5 (08 2013). https://doi.org/10.1007/s12369-013-0196-9

[9] Maha Salem, Stefan Kopp, Ipke Wachsmuth, Katharina Rohlfing, and Frank Joublin. 2012. Generation and Evaluation of Communicative Robot Gesture. *International Journal of Social Robotics* 4 (04 2012), 201–217. https://doi.org/10.1007/s12369-011-0124-9

[10] Allison Sauppé and Bilge Mutlu. 2014. Robot deictics: How gesture and context shape referential communication. In *Proceedings of the 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 342–349.