A Framework for Robot-Generated Mixed-Reality Deixis

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

ABSTRACT

Language-capable robots interacting with human teammates may need to make frequent reference to nearby objects, locations, or people. In human-robot interaction, such references are often accompanied by deictic gestures such as pointing, using human-like arm motions. However, with advancements in augmented reality technology, new options become available for deictic gesture, which may be more precise in picking out the target referent, and require less energy on the part of the robot. In this paper, we present a conceptual framework for categorizing different types of mixed-reality deictic gestures that may be generated by robots in human-robot interaction scenarios, and presented an analysis of how these categories differ along a variety of dimensions.

KEYWORDS

Augmented Reality, Deixis, Nonverbal Interaction

ACM Reference Format:
Tom Williams. 2018. A Framework for Robot-Generated Mixed-Reality Deixis. In *Proceedings of the 1st International Workshop on Virtual, Augmented, and Mixed Reality for Human-Robot Interaction (VAM-HRI'18).* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

Language-capable robots interacting with human teammates may need to make frequent reference to nearby objects, locations, or people [11, 28, 29]. Consider the following example scenarios:

- A wheelchair user instructs his wheelchair: "I need my medication." The wheelchair can see pill bottles on two nearby tables and wishes to know which medication its user means.
- A mission commander in an alpine search and rescue scenario instructs a UAV "Search for survivors behind that fallen pylon." The UAV can see two fallen pylons and wishes to know which its user means.
- A robot working on the ISS observes an astronaut teammate struggling with a task, and spots a misplaced tool which might make her task easier. The robot thus wishes to point out the tool to its teammate.

In each of these scenarios, it will be advantageous if the robot is able to generate a deictic (e.g., pointing) gesture [19] toward its target referent simultaneous with its verbal description of that referent [8]. In human-robot interaction, this type of deictic gesture has typically been executed as a human-like arm motion [1, 15–17, 22, 23]. However, with advancements in augmented reality technology, new options become available for deictic gesture, which may be more precise in picking out the target referent, and require less energy on the part of the robot. Specifically, if a human teammate is wearing an AR headset, a robot teammate may be able to pick out an object it wishes to refer to in their teammate's field of view by circling it or drawing an arrow towards it. As such, augmented reality also presents the opportunity for deixis within the mixed-reality environment shared by robots and their human teammates.

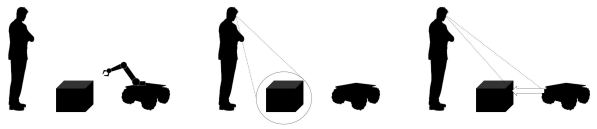
While there has been some previous work on using visualizations as "gestures" within virtual or augmented environments [27], this metaphor of visualization-as-gesture has not yet been fully explored. This is doubly true for human-robot interaction scenarios, in which the use of augmented reality is surprisingly underexplored, yet also in which the connection between visualization and gesture is strongest, as visualization becomes an alternative or complement to "true" (i.e., physical) deictic gesture.

In order to fully explore this metaphor and the possibilities it may afford to human-robot interactions, what is first needed is a framework for differentiating between the different types of gestures that may be used in mixed-reality contexts; as we will argue, human-robot interactions in mixed-reality afford not only the traditional deictic gestures available in "pure reality" and the traditional spatially grounded annotations available in augmented reality, but also entirely new types of gestures that bridge the perspectives inherently encoded in those traditional gestural forms. In this paper, we present an initial framework that allows for these distinctions to be made clear, and present a set of example dimensions along which differences between these categories can be observed, leaving for future work a full exploration of the space of possible dimensions that could be included in such a framework.

2 PREVIOUS WORK

While there has been work on using augmented reality to provide visualizations and annotations to accompany language for several decades [12, 13, 25, 26], there has been little research into using augmented reality for human-robot communication, and surprisingly, little research into using augmented reality for human-robot interaction at all. And in fact, in their recent survey of augmented reality, Billinghurst et al. [4] cite intelligent systems, hybrid user interfaces, and collaborative systems as areas that have been underattended-to in the AR community.

Most relevant to the current paper, Sibirtseva et al. [24] use augmented reality annotations to indicate different candidates referential hypotheses after receiving ambiguous natural language commands, and Green et al. [14] present a system that uses augmented reality to facilitate human-robot discussion of a plan prior to execution. Also related, [20] use augmented reality to provide a first-person view for robot teleoperators, integrating into this framework a joystick which the teleoperator can use to control the robot's dialogue behavior.



(a) Egocentric (b) Allocentric (c) Multi-Perspective

Figure 1: Categories of Mixed-Reality Deictic Gestures

There have also been approaches which use augmented reality to non-verbally communicate the robot's intentions: Andersen et al. [2] and Chadalavada et al. [5] both use used projector-based augmented reality to project robots' intentions into the environment they share with their human teammates. Relatedly, Rosen et al. [21] visualize robots' intended arm trajectories using a mixed-reality head-mounted display, and Frank et al. [9] visualize robot-relevant spatial information, such as the robot's current field of reach, through AR visualizations on a hand-held tablet.

However, to the best of our knowledge, there has been no previous work looking at augmented reality or mixed reality as a means to supplement, replace, or stand in for physical deictic gestures generated by robots during the course of human-robot interaction. In this paper, we provide a conceptual framework for categorizing different kinds of mixed-reality gestures that might be used in human-robot interaction, and analyze how the different categories of gestures within this framework differ along dimensions such as perspective, legibility, cost, and privacy.

3 CONCEPTUAL FRAMEWORK

In this section, we present a conceptual framework for describing mixed-reality deictic gestures. Let R be a robot and H be that robot's human teammate. An object appearing to both R and H can be said to be located at point P_R, in the robot's coordinate frame, and at point P_H in the human's coordinate frame. Suppose R wishes to issue a deictic gesture to refer to this object. Note that unlike in standard reality, a robot operating within a mixed-reality environment may have access to perspectives other than its own, both with regards to perception (i.e., it may perceive perspectives other than its own) and with regards to praxis (i.e., it may act within perspectives other than its own). We argue that while standard reality allows for but a single class of deictic gesture (with regards to perspective), mixed-reality environments enable three classes of deictic gestures: *egocentric* deictic gestures, *allocentric* deictic gestures, and *multi-perspective* deictic gestures (as shown in Fig. 1).

3.1 Egocentric Deictic Gestures

We categorize the typical deictic gestures (such as pointing) available to a robot in standard reality as *egocentric* deictic gestures (as shown in Fig. 1a), as they are executed, and must be interpreted by others, from their own perspective. When viewed within the larger framework of mixed-reality deictic gestures, we argue that these gestures can be viewed as implicitly encoding the perspective of their generator.

This is reflected in the manner in which a robot must execute such actions. That is, to gesture towards object O located at point $P_O^R = (X_O^R, Y_O^R, Z_O^R)$ within its own coordinate frame, a(n armed) robot R may simply orient one of its arms along the vector $\begin{bmatrix} X_O^R & Y_O^R & Z_O^R \end{bmatrix}^\top$ (with possible transformations based on orientation of camera, e.g., relative to body relative to arm).

3.2 Allocentric Deictic Gestures

Unlike robots operating in standard reality, robots operating in mixed-reality environments may – like any mixed-reality technology – also generate *allocentric* deictic gestures (as shown in Fig. 1b) from and through the perspectives of their human teammates. A robot, may, for example, "gesture" to an object by circling it within its teammate's display. In contrast to egocentric gestures, we argue that these gestures can be viewed as implicitly encoding the perspective of the generator's intended interlocutor.

The allocentric nature of this type of action is similarly reflected in the manner in which it must be executed. That is, to gesture towards an object O located at point $P_O^H = (X_H, Y_H, Z_H)$ within its human teammate H's coordinate frame, a robot need not consider its own coordinate frame at all, and may simply draw a circle centered at pixel coordinates $p_O^H = (p_{O_x}^H, p_{O_y}^H)$ after transformation from camera coordinates to pixel coordinates. This circle must of course then be redrawn as the human teammate's field of view shifts.

3.3 Multi-Perspective Deictic Gestures

Finally, unlike robots operating in standard reality and unlike other mixed-reality technologies, robots operating in mixed-reality environments may generate gestures that connect their own perspective to the perspective of their human teammates. A robot may, for example, "gesture" to an object by drawing an arrow in its teammate's display from itself to its target object. Unlike egocentric and allocentric gestures, *multi-perspective* gestures (as shown in Fig. 1c) do not encode a single canonical perspective, but are rather defined entirely by their connection between multiple perspectives.

The multi-perspective nature of this type of action is, as before, reflected in the manner in which it must be executed. That is, to gesture towards an object located at pose $\mathcal{P}_R$ within its own perspective and at pose $\mathcal{P}_H$ within its teammate's perspective requires the robot to first calculate its own pose from the perspective of its interlocutor. The robot must then convert the poses of both the object and itself from its human teammate's coordinate frame to pixel coordinates $p_{OH} = (p_{O_x}^H, p_{O_y}^H)$ (i.e., the pixel coordinates of the object from the human's perspective) and $p_R^H = (p_{R_x}^H, p_{R_y}^H)$ (i.e., the pixel coordinates of the robot from the human's perspective) in its human teammate's display, and then draw an arrow from p_R^H to p_O^H in that display. This arrow must of course then be redrawn as the human teammate's field of view shifts.

4 ANALYSIS OF MIXED-REALITY DEICTIC GESTURES

Each of these three gesture categories comes with its own unique properties. Here, we specifically examine six: perspective, embodiment, capability, legibility, cost, and privacy. These dimensions are summarized in Table 1.

Category	Perspective	Embodiment	Capability	Legibility (D)	Legibility (S)	Cost (G)	Cost (M)	Privacy
Egocentric	Robot	Yes	Yes	Low	Low	High	Low	Low
Allocentric	Human	No	No	High	High	Low	High	High
Multi-Perspective	Human	Yes	No	Low	High	Low	High	High

Table 1: Analysis of Mixed-Reality Deictic Gestures

Category	Perspective	Embodiment	Capability	Legibility (D)	Legibility (S)	Cost (G)	Cost (M)	Privacy
$\{E\}$	Robot	Yes	Yes	Low	Low	High	Low	Low
$\{A\}$	Human	No	No	High	High	Low	High	High
$\{M\}$	Human	Yes	No	Low	High	Low	High	High
$\{E, A\}$	Both	Yes	Yes	High	High	High	High	Low
$\{E, M\}$	Both	Yes	Yes	Low	High	High	High	Low
$\{A, M\}$	Human	Yes	No	High	High	Low	High	High
$\{E, A, M\}$	Both	Yes	Yes	High	High	High	High	Low

Table 2: Analysis of Combinations of Mixed-Reality Deictic Gestures

The most salient dimensions that differentiate these three categories of mixed-reality deictic gestures are the perspectives and embodiment that they require. The perspectives required for these three categories are clearly defined: egocentric gestures require access to the robot's perspective, while allocentric and multi-perspective gestures require access to the human teammate's perspective. These gestures are categorized differently, however, when viewed with respect to embodiment; egocentric and multi-perspective gestures require the gesturer to be embodied (and co-present), while allocentric gestures do not require an embodied form. Finally, egocentric gestures require a particular form of embodiment, i.e., they require the gesturer to not only be embodied and co-present, but to be able to take physical action in the world; whereas multi-perspective (and, clearly, egocentric) gestures do not require this capability for physical action.

4.1 Legibility

In previous work, Dragan et al. [7] defined the notion of the legibility of an action, which describes the ease at which a human is able to determine the goal or purpose of an action as it is being carried out. In later work with Holladay et al. [16], Dragan then applies this notion to deictic gestures as well, analyzing the ability of the final gestural position to enable humans to pick out the target object. We believe, however, that this is really a distinct sense of legibility from Dragan's original formulation, and as such, must first refine this notion of legibility as applied to deictic gestures into two categories. Specifically, we will use *dynamic legibility* to refer to the degree to which a deictic gesture enables a human teammate to pick out the target object *as the action is unfolding* (in line with Dragan's original formulation) and *static legibility* to refer to the degree to which the final pose of a deictic gesture enables a human teammate to pick out the target object after the action is completed (in line with Holladay's formulation).

The three categories of mixed-reality deictic gestures we describe in this paper differ with respect to both dynamic and static legibility. Allocentric gestures have high dynamic legibility (given that there is no dynamic dimension) and high static legibility (given that the target is uniquely picked out). Egocentric gestures have low dynamic legibility (relative to allocentric gestures) given that their target may not be clear at all as the action unfolds, and low static legibility, as the target may not be clear after the action is performed either, depending on distance to the target and density of distractors. Multi-perspective gestures have high static legibility (given that the target is uniquely picked out) but may have low static legibility if they are portrayed as an animation unfolding over time rather than instantaneously appearing.

4.2 Cost of Execution

These three categories of mixed-reality deictic gestures come with different technical challenges. From the perspective of energy usage, egocentric gestures are expensive due to their physical component, while allocentric and multi-perspective gestures are cheap. On the other hand, allocentric and multi-perspective gestures are difficult to maintain due to registration challenges, whereas egocentric gestures have no additional cost once executed. Accordingly, we categorize egocentric gestures as having a high generation cost but low maintenance cost, and allocentric and multi-perspective gestures as having low generation costs but high maintenance costs.

4.3 Privacy

In addition, these three categories afford different levels of privacy. Allocentric and multi-perspective gestures, are only visible to the human teammate with whom the robot is communicating, and as such we regard them as high-privacy. Egocentric gestures, on the other hand, are also visible to observers, and as such, we describe them as low-privacy. This dimension is particularly important for human-robot interaction scenarios involving both sensitive user populations (e.g., elder care or education) or in adversarial scenarios (e.g., competitive [6], police [3], campus safety [10], or military domains (as in DARPA's "Silent Talk" program) [18]).

5 COMBINATION OF MIXED-REALITY DEICTIC GESTURES

Given these three classes of mixed-reality deictic gestures, we can now also reason about combinations of these gestures, as summarized in Table 2. Specifically, we define seven categories of gesture combinations: $\{E\}$: an egocentric gesture alone; $\{A\}$: an allocentric gesture alone; $\{M\}$: a multi-perspective gesture alone; $\{E, A\}$: simultaneous egocentric and allocentric gestures; $\{E, M\}$: simultaneous egocentric and multi-perspective gestures; $\{A, M\}$: simultaneous allocentric and multi-perspective gestures; $\{E, A, M\}$: simultaneous egocentric, allocentric, and multi-perspective gestures; (we ignore the empty combination, $\{\}$, i.e., no gesture).

For example, simultaneously pointing to and circling an object would fall into class $\{E, A\}$. These gesture combinations are particularly interesting because they combine properties of their constituent gestures in various ways. Simultaneous generation of gestures requiring different perspectives results in both perspectives being needed. The embodiment and capability requirements of simultaneous gestures combine disjunctively. The legibilities and costs of simultaneous gestures combine using a *max* operator, as the legibility of one gesture will excuse the illegibility of another, but the low cost of one gesture will not excuse the high cost of another. And the privacies of simultaneous gestures combine using a *min* operator, as the high privacy of one gesture does not excuse the low privacy of another.

6 CONCLUSION

In this paper, we have presented a framework for categorizing different types of mixed-reality deictic gestures that may be generated by robots in human-robot interaction scenarios, and presented an analysis of how these categories differ along a variety of dimensions. In future work, we first plan to expand this framework to incorporate additional gestures, and to evaluate this framework on a set of case studies. We then plan to apply this framework to the design of mixed-reality deictic gestures, which we will use to supplement the referring expressions generated by robots through natural language.

REFERENCES

[1] Henny Admoni, Thomas Weng, and Brian Scassellati. 2016. Modeling communicative behaviors for object references in human-robot interaction. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 3352–3359.

[2] Rasmus S Andersen, Ole Madsen, Thomas B Moeslund, and Heni Ben Amor. 2016. Projecting robot intentions into human environments. In *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*. IEEE, 294–301.

[3] Cindy L Bethel, Daniel Carruth, and Teena Garrison. 2012. Discoveries from integrating robots into SWAT team training exercises. In *Safety, Security, and Rescue Robotics (SSRR), 2012 IEEE International Symposium on*. IEEE, 1–8.

[4] Mark Billinghurst, Adrian Clark, and Gun Lee. 2015. A survey of augmented reality. *Foundations and Trends in Human-Computer Interaction* 8, 2-3 (2015), 73–272.

[5] Ravi Teja Chadalavada, Henrik Andreasson, Robert Krug, and Achim J Lilienthal. 2015. That's on my mind! robot to human intention communication through on-board projection on shared floor space. In *Mobile Robots (ECMR), 2015 European Conference on*. IEEE, 1–6.

[6] Filipa Correia, Patrícia Alves-Oliveira, Nuno Maia, Tiago Ribeiro, Sofia Petisca, Francisco S Melo, and Ana Paiva. 2016. Just follow the suit! trust in human-robot interactions during card game playing. In *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*. IEEE, 507–512.

[7] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. 2013. Legibility and predictability of robot motion. In *Human-Robot Interaction (HRI), 2013 8th ACM/IEEE International Conference on*. IEEE, 301–308.

[8] Charles J Fillmore. 1982. Towards a descriptive framework for spatial deixis. *Speech, place and action: Studies in deixis and related topics* (1982), 31–59.

[9] Jared A Frank, Matthew Moorhead, and Vikram Kapila. 2017. Mobile Mixed-reality interfaces That enhance human–robot interaction in shared spaces. *Frontiers in Robotics and AI* 4 (2017), 20.

[10] Scott Goldfine. 2017. Assessing the Prospects of Security Robots. (October 2017). https://www.campussafetymagazine.com/hospital/42476/

[11] Georgia M Green. 1996. *Pragmatics and natural language understanding*. Psychology Press.

[12] SA Green, Mark Billinghurst, X Chen, and GJ Chase. 2008. Human-robot collaboration: A literature review and augmented reality approach in design. *International Journal of Advanced Robotic Systems* (2008).

[13] Scott A Green, Mark Billinghurst, XiaoQi Chen, and J Geoffrey Chase. 2007. Human Robot Collaboration: An Augmented Reality ApproachâĂŤA Literature Review and Analysis. In *Proceedings of the ASME International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. American Society of Mechanical Engineers, 117–126.

[14] Scott A Green, J Geoffrey Chase, XiaoQi Chen, and Mark Billinghurst. 2009. Evaluating the augmented reality human-robot collaboration system. *International journal of intelligent systems technologies and applications* 8, 1-4 (2009), 130–143.

[15] Khurram Gulzar and Ville Kyrki. 2015. See what i mean-Probabilistic optimization of robot pointing gestures. In *Proceedings of the fifteenth IEEE/RAS International Conference on Humanoid Robots (Humanoids)*. IEEE, 953–958.

[16] Rachel M Holladay, Anca D Dragan, and Siddhartha S Srinivasa. 2014. Legible robot pointing. In *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on*. IEEE, 217–223.

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

[18] Ivan S Kotchetkov, Brian Y Hwang, Geoffrey Appelboom, Christopher P Kellner, and E Sander Connolly Jr. 2010. Brain-computer interfaces: military, neurosurgical, and ethical perspective. *Neurosurgical focus* 28, 5 (2010), E25.

[19] David McNeill. 1992. *Hand and mind: What gestures reveal about thought*. University of Chicago press.

[20] André Pereira, Elizabeth J Carter, Iolanda Leite, John Mars, and Jill Fain Lehman. 2017. Augmented Reality Dialog Interface for Multimodal Teleoperation. In *26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), 2017*.

[21] Eric Rosen, David Whitney, Elizabeth Phillips, Gary Chien, James Tompkin, George Konidaris, and Stefanie Tellex. 2017. Communicating robot arm motion intent through mixed reality head-mounted displays. *arXiv preprint arXiv:1708.03655* (2017).

[22] 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, 2 (2012), 201–217.

[23] Allison Sauppé and Bilge Mutlu. 2014. Robot deictics: How gesture and context shape referential communication. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM, 342–349.

[24] Elena Sibirtseva, Dimosthenis Kontogiorgos, Olov Nykvist, Hakan Karaoguz, Iolanda Leite, Joakim Gustafson, and Danica Kragic. 2018. A Comparison of Visualisation Methods for Disambiguating Verbal Requests in Human-Robot Interaction. *arXiv preprint arXiv:1801.08760* (2018).

[25] Wolfgang Wahlster, Elisabeth André, Winfried Graf, and Thomas Rist. 1991. Designing illustrated texts: how language production is influenced by graphics generation. In *Proceedings of the fifth conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 8–14.

[26] Peter Wazinski. 1992. Generating spatial descriptions for cross-modal references. In *Proceedings of the third conference on Applied natural language processing*. Association for Computational Linguistics, 56–63.

[27] Sean White, Levi Lister, and Steven Feiner. 2007. Visual hints for tangible gestures in augmented reality. In *Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on*. IEEE, 47–50.

[28] Tom Williams and Matthias Scheutz. 2017. Referring Expression Generation Under Uncertainty: Algorithm and Evaluation Framework. In *Proceedings of the 10th International Conference on Natural Language Generation*.

[29] Tom Williams and Matthias Scheutz. 2018 (in press). Reference in Robotics: A Givenness Hierarchy Theoretic Approach. In *The Oxford Handbook of Reference*, Jeanette Gundel and Barbara Abbott (Eds.).