Toward Open-World Human-Robot Interaction: What Types of Gestures Are Used in Task-Based Open-World Referential Communication?

Mark Higger, Polina Rygina, Logan Daigler,
Lara Ferreira Bezerra, Zhao Han, Tom Williams
MIRRORLab, Department of Computer Science
Colorado School of Mines, Golden CO 80401

Abstract

Gestures play a critical role in human-human and human-robot interaction. In task-based contexts, deictic gestures like pointing are particularly important for directing attention to task-relevant entities. While most work on task-based human-human and human-robot dialogue focuses on closed-world domains, recent research has begun to consider open-world tasks, where task-relevant objects may not be known to interactants a priori. In open-world tasks, we argue that a more nuanced consideration of gesture is necessary, as interactants may use gestures that bridge traditional gesture categories, in order to navigate the open-world dimensions of their task environment. In this work, we explore the types of gestures used in open-world task contexts, and their frequencies of use. Our results suggest a need to rethink the way that gesture analysis is approached in the study of human-human and human-robot interaction.

1 Introduction

For task-based human-robot interaction (HRI), effective communication can greatly increase task effectiveness (Cantrell et al., 2011; Tellex et al., 2020). Critically, this includes both verbal and non-verbal communication (Mavridis, 2015). Effective non-verbal communication can include both gaze and gesture (De Angeli et al., 1998), and are fundamental in human-human communication and substantially contribute to fluent communication (Kendon, 1997; Kita, 2003; Goldin-Meadow, 1999; Ping and Goldin-Meadow, 2010). Accordingly, understanding and generating these non-verbal communication modalities is critical for effective human-robot communication.

Recently, work on natural language understanding in robotics (Han et al., 2022; Han and Williams, 2022; Culpepper et al., 2022; Williams and Scheutz, 2015b) has been expanding beyond traditional tabletop domains to consider open-world contexts. In contrast to a closed-world context where there is an assumption that all possible referents have an a-priori representation in the robot's system, open-world contexts have referents which may be new or unknown to robot ahead of time. For example, if a nurse asks an assistive robot "Go to the kitchen to find a glass of water, then bring it to the patient". When trying to figure out what "it" means in this sentence the robot needs to be aware of the glass of water. However if the robot is using a vision-based system to ground referring language it may only contain a representation of the objects currently visible in the room. In order to ground "it" to a particular object, the robot may need to create a new representation for the glass of water outside of the vision system. By allowing for new representations to be created outside of what the robot is already aware of (in this case through vision), it can then allow for understanding reference in an open-world context. Work in this area has led to a number of algorithms for *open-world reference resolution* (Williams and Scheutz, 2015b; Williams et al., 2016; Culpepper et al., 2022) to allow for this type of behavior.

While there is work on linguistic grounding in open-world contexts, work on robot gesture still largely assumes a closed world where gestural targets are visible, known, and close-by (Lücking et al., 2015; Sauppé and Mutlu, 2014). We argue that this has led to an overly narrow focus on understanding and generating specific, narrow classes of gestures in human-robot interaction; i.e., deictic gestures that focus an interactant's attention on a visible, nearby area of the task environment. Based on this argument, we analyze the gestures used in a recent corpus of human-human interactions, collected in a novel task environment designed by Han et al. (2022) to elicit a more ecologically valid range of referring forms. As we will show, our analysis of the gestures used in this task context backs up our argument, yielding a novel taxonomy

Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue, August, 16–17, 2023, Maribor.

of gestures used in open-world dialogue contexts, and suggesting a need for human-robot interaction researchers to fundamentally rethink the types of gestures they are attempting to model in task-based human-robot interaction.

2 Related Work

2.1 Open World Communication

To understand the ways in which gesture in current task-based HRI are overly limited, let us first consider the linguistic work that specifically targets open-world interactions. While task-based natural language understanding and generation in human-robot interaction has traditionally considered only closed-world environments. There has recently been an increase in research relaxing this closed-world assumption to consider *open worlds*, especially in the context of reference resolution (Williams and Scheutz, 2015b; Williams et al., 2016; Culpepper et al., 2022).

Reference resolution is the process of identifying what knowledge associated with particular entities in a robot's memory is being referred to by a speaker's referring language. While traditional approaches to reference resolution (and the related process of language grounding) have only attempted to associate incoming referring expressions with pre-existing knowledge representations, work on open-world reference resolution has additionally modeled how listeners might assess when an incoming referring expression is likely to refer to a previously unknown entity, and how a new knowledge representation might be created in such cases through the process of hypothesization (Williams and Scheutz, 2015a,b). More recently, Culpepper et al. (2022) presented a novel algorithm that allows for *incremental* reference resolution, which allows for a real-time word-by-word processing within this type of open-world framework.

While there has been less work on open-world language generation, this too is starting to change. Han et al. (2022), for example, recently presented a novel experimental setup designed to capture natural references to both visible, previously visible, and not yet visible referents, in order to develop computational referring form selection models that can handle these sorts of open world references. Like prior work on open-world reference resolution, however, this work has been solely focused on natural language generation, and has not attempted to account for the role that nonverbal communication, especially gesture, plays in open-world dialogue.

2.2 Gesture in Human-Robot Interaction

Because gestures are known to be a fundamental part of human communication (Kita, 2003; Goldin-Meadow, 1999), the use of gesture to enhance human-robot interaction has attracted significant attention across the history of human-robot interaction (Waldherr et al., 2000). Effective gesture has been shown both to promote sociability and interactions, making robots more natural and enjoyable to work with (Kim et al., 2013; Salem et al., 2012); and to enhance the effectiveness and productivity of task-based interactions (Gleeson et al., 2013; Gross et al., 2017). Because of the important role gestures play in human-robot interaction, HRI researchers have devised a number of taxonomies for categorizing the different types of gesture that can be used – and understood – by robots. Many of the existing taxonomies used for gesture generation and understanding for HRI research primarily focuses on physical arm motions (Allwood et al., 2007; Dael et al., 2012). These taxonomies are often very complex: the BAP taxonomy (Dael et al., 2012), for example, has nearly 40 non-mutually exclusive codes for gesture, with complex and specific codes such as "Left arm action curved repetition" or "Asymmetrical arms action". For understanding what physical movements are needed to create with robots these taxonomies can be very helpful, however these taxonomies are often not grounded with the intent of the gesture in mind. Additionally, research in the semiotics community (Goodwin, 2003) shows that physical manifestation is not always a good indicator of the intent of the gesture, as context and subtle indicators also play a large role in gestural meaning. This means that despite their comprehensive nature in terms of physical motion, these large-scale categorization systems fail to account for the ways that different types of gestures are typically used to achieve different types of communicative purposes in human-robot interactions.

In contrast, the taxonomy devised by McNeill and Levy (1982) for use in human-human interaction categorizes gestures according to a small number of conversational roles: deictic, iconic, metaphoric, beat, and emblematic. Deictic gestures, like pointing, direct an interactants's visual attention to a particular object or location. Iconic ges-

tures mime the physicality of an object or action to direct an interactant's internal attention to a particular concept. Metaphoric gestures are used to convey more abstract concepts such as time. Beat gestures are used for pacing or timing of linguistic structure. Emblematic gestures (e.g. a wave or thumbs-up) have distinct meanings derived through social and cultural context. While, this taxonomy for categorizing gestures (McNeill and Levy, 1982) and their incorporation into language-based communication (McNeill, 1985) has been widely adopted across the fields of psycholinguistics and human-robot interaction (de Wit et al., 2022) it lacks the specificity, and grounding to physical movement for interpretability that is often required, especially when it comes to the wide variety of gestures that fit under the broader umbrella of "deictic gestures".

Within the domain of human-robot interaction, there is wide recognition that understanding and generation of non-verbal communication is critical for situated interaction (Cantrell et al., 2011; Breazeal et al., 2005; Mavridis, 2015). While there has been work on all of the gestural categories described by McNeill and Levy (1982), deictic gesture in particular has attracted significant attention, due to its highly task-oriented and more easily interpretable nature. Indeed, in task-based human-robot interaction, there is often an exclusive focus on understanding and generating *deictic* gestures. In the foundational work of Sauppé and Mutlu (2014), a wide variety of subtypes of deictic gestures are studied, including pointing, presenting, touching, exhibiting, grouping, and sweeping, with each category defined according to the physical motion of the gesture. For example, a 'touching' gesture requires direct physical contact with the referent, while a 'sweeping' gesture used wide arm movements to direct attention to larger regions. Work on computational understanding and generation of deixis typically models deictic gestures by projecting a saliency cone from the origin of a deictic pointing gesture (Kranstedt et al., 2005) outwards in a particular direction. These cones can then be used in a multi-modal estimator that combines grounded language and non-verbal communication to identify which potential referent is most likely given the particular language and gestures used (Schauerte et al., 2010; Schauerte and Fink, 2010; Lücking et al., 2015).

Yet critically, these methods assume that the target of a deictic gesture is visible and known to the robot so it can find the most likely target that falls within the deictic cone. While this seems reasonable at first glance, it is clear that humans frequently use gestures to refer to objects that cannot be seen or may not be known to the other person they are gesturing to. A simple example of this is the way that iconic gestures help to draw users' *internal* attention to a target referent representation, rather than drawing users' *visual* attention to a target referent stimulus. Moreover, some researchers (Stogsdill et al., 2021; Enfield et al., 2007) have recently begun to explore types of gestures that arise in large-scale and open-world environments, which seem to further trouble the boundaries between these categories of gesture.

2.3 Gesture in Large-Scale and Open Worlds

One example of how the boundaries between traditional gesture categories are being troubled and contested within psycholinguistics is the work by Enfield et al. (2007). They examined how general "pointing" gestures could have a more complex meaning than what is traditionally associated with deictic gestures. Specifically, Enfield et al. looked at how pointing was used during conversation in small Laotian villages and classified pointing gestures into two types; primary Big (B) and secondary Small (S) points. The B-points were composed of large arm movements, while S-points were smaller, single-armed gestures that had more complex hand movements. While B-points seemed to be used in the way typically expected of deictic gestures (i.e., to point to a physical location in space), S-points were instead found to have more complicated usage. Sometimes the S-points were used similarly to deictic gestures specifying a physical object or person which would be ambiguous through language alone. But other times, the S-Points were also used to refer to locations that were not in the current view while not necessarily pointing in the exact location of that object.

These S-points thus trouble McNeill and Levy (1982)'s conception of deictic gestures, in which the objective of a pointing gesture is to direct attention to a physical location via spatial information. The notion of "abstract deictic" gestures are explored in McNeill et al. (1993), which looks at how deictic gestures can be used to point to objects or people which are not physically there. While McNeill et al. (1993)'s "abstract deictic" gestures are used to point to entities without a physical pres-

ence, they still retain spatial information within a narrative context, being used to denote relative position in a non-present space. However, some of the behaviors present in Enfield et al. (2007)'s S-point's includes pointing gestures which do not necessarily contain direct spatial information, but instead are used to create a gestural representation of an entity, present or not. This suggests there is non-spatial information that can also be conveyed by pointing. While this use of non-spatial pointing is explored within the semiotics community (Enfield et al., 2007; McNeill, 2003), these gestures are still often classified as "Deictic gestures", despite their lack of spatial information making them inherently non-deictic. Perhaps this is due in part to the grounding of McNeill's categories in studies of conversational rather than task-based dialog. We argue that because of the lack of spatial information, a non-spatial pointing gesture does not fit cleanly into McNeill's "Deictic gesture" category. As such, this suggests a need to our work that studies task-based gestures in order to better understand the use of gestures in complex, large-scale, open-world environments.

These troubled category boundaries have also been recently noted in work in the HRI community. Specifically, Stogsdill et al. (2021) explored the use of vague, non-deictic pointing gestures that are very similar to the examples of S-Points found in Enfield et al. (2007). For example, if someone is trying to refer to the room next door, they may vaguely wave toward the room in question without pointing precisely or directing their gaze. Or, if someone is referring to another faraway city, they may point in a completely random direction. This use of an abstract pointing gesture removes the spatial information that is central to McNeill et al. (1993)'s deictic gestures. Instead, Stogsdill et al. (2021) argues that this gesture attempts to merely convey the concept of "away", which may be closer to a metaphoric gesture than a deictic gesture.

This spectrum of category-spanning gestures remains understudied, yet critically points to a need to reconsider the categories of gesture used in task-based interaction. More specifically, in order to enable open-world Human-Robot Interaction, we argue that there is a fundamental need to question how and where existing approaches to gesture understanding and generation might fall short, and how these shortcomings might be grounded in the general taxonomies used by roboticists and psycholinguists to make sense of gesture.

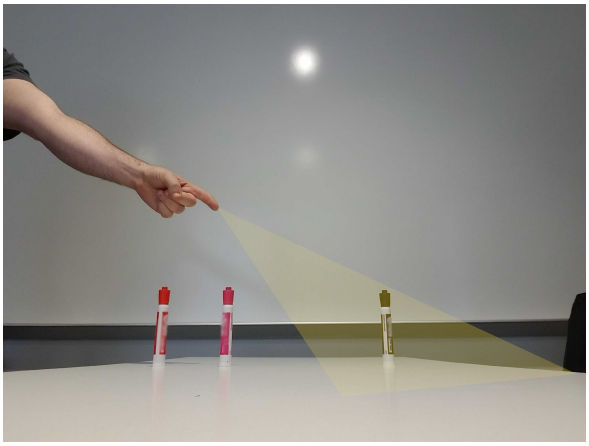

Figure 1: Example of the conical deictic projection often found in computer-vision-based deictic gesture representation. As can be seen, the gesture becomes less precise as the target becomes farther away from the source of the gesture.

In this work, we thus propose to investigate the following research questions:

R1: How might we better categorize the types of referring gestures used in open-world task-based environments?

R2: How prevalent are these different categories of gestures?

R3: Is the overwhelming focus of the HRI research community on precise, deictic, *B-point* gestures justified by this distribution of observed gestures?

3 Methodology

3.1 Dataset

To evaluate our research questions, we analyzed the experimental data from the experiments conducted by Han et al. (2022). We will briefly describe the context in which Han et al. collected that data, as reported in Han et al. (2022), to explain why this was an ideal dataset for answering our research questions. The task environment used in Han et al. (2022)'s work (shown in Figure 2) was partitioned into four quadrants, each containing a variety of colored blocks. In their experiment, pairs of participants (an *instructor* and a *learner*) participated in a sequence of four building tasks, one in each quadrant, in which the *instructor* taught the *learner* how to construct a different building from those blocks. Specifically, Han et al. (2022) designed

their experiment so that each of these four building tasks required blocks not available within the task's quadrant, including blocks seen in previous quadrants or located in quadrants that were to be visited in the future. This task structure was thus inherently open-world in nature; the instructor was required not only to refer to blocks that were immediately visible, but also to refer to blocks that had been seen in previous quadrants, as well as blocks whose locations were as-yet-unknown. By having discreet sections with items that participants were aware of and other sections they not aware of, they established quadrants that belonged to a current, closed working context and quadrants that belonged to an open-world. While Han et al. (2022) report designing this experiment to analyze open-world *language* production, we realized that their dataset could also be a rich source of open-world *gesture* production.

We analyzed twelve videos from Han et al. (2022)'s video dataset, a total of 337 minutes with an average of 28 minutes per video. From these videos, we identified 1067 gestures in total and an average of 89 gestures per video. In the next section, we will thus describe how we analyzed the gestures found in these videos.

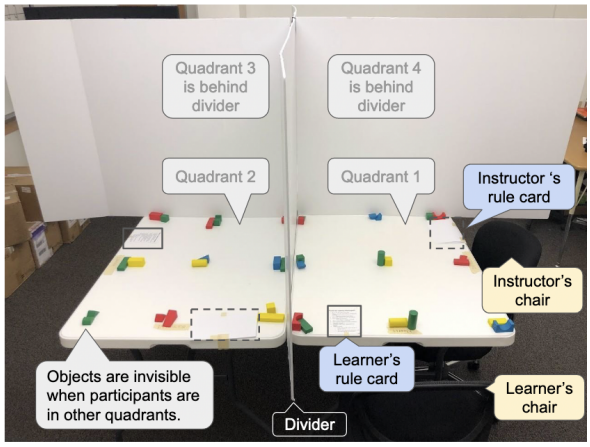

Figure 2: Setup for testing referring form selection from "Evaluating Referring Form Selection Models in Partially-Known Environments" (Han et al., 2022)

3.2 Qualitative analysis and gesture coding

To analyze those videos, we performed an iterative qualitative coding procedure. First, we began by breaking each video into a set of *communicative action segments* in which discrete gestures were used. Next, we performed open coding to identify, for each gesture, the (1) physical manifestation of the gesture and (2) information conveyed by the gesture. Then, we removed from consideration any gestures that did not appear to be related to the task itself, or which did not appear to provide the interactant with meaningful referential information to help them accomplish the task, such as beat gestures or emblematic gestures. Finally, we analyzed the literature on gesture classification to cluster gestures into discrete categories that were informed by prior literature whenever possible, while also striving to ensure (1) clear boundaries between gesture clusters, and (2) clear criteria for assignment of gestures to clusters.

Once a gesture taxonomy was identified, coders were used to identify when in the videos particular gestures were used. Then 17% of communicative action segments were used to establish an inter-coder reliability rating (IRR) using Cohen's Kappa Coefficient (Cohen, 1960) of $\kappa = 0.35$, denoting fair agreement. After an IRR was established, a single coder was used for the remaining segments. In the next section, we will describe the taxonomy of gestures that resulted from this qualitative coding procedure.

4 Gesture Taxonomy

Based on our qualitative analysis, we formulated the following taxonomy of open-world task-based gestures which strike a balance between the intent of the gesture while allowing for the visually separable physical interpretability needed for robotics. This taxonomy is comprised of five key categories: (1) Precise Deictic Gestures, (2) Small Region Deictic Gestures, (3) Large Region Deictic Gestures, (4) Abstract Pointing Gestures, and (5) Iconic Reference Gestures. Summary statistics for the use of these gestures are shown in Table 1.

Gesture Type	Count	pct of total
Precise Deictic	391	36.6%
Small Region Deictic	434	40.6%
Large Region Deictic	13	1.2%
Abstract Point	103	9.7%
Iconic Reference	126	11.8%

Table 1: Number of gestures observed in the analyzed dataset

4.1 Precise Deictic Gesture

We identified three distinct types of deictic gestures that manifest in open-world task-based interactions.

The first category was what we term "Precise Deictic Gestures". These gestures were those most closely related to the traditional conceptualization of Deictic Gesture, and were relatively common (36.6% of all analyzed gestures) due to their important role in specifying which block will be needed next for the experiment. The apparent purpose of this gesture was to physically direct attention to a single target object with a high level of specificity. This type of gesture physically manifested either as touching that object, or as pointing to that object in a way where it was the only task-relevant object entirely within the saliency cone extending from the gesturer's hand (cf. Lücking et al., 2015; Schauerte et al., 2010). Thus, while the physical motions of precise deictic gestures are dependent on the environement, for example the gesturer may need to put their finger closer to an object if it it is near another object, the purpose of precise decictic gestures is maintained.

This category of gesture captures gestures that Sauppé and Mutlu (2014) would have categorized as *exhibiting, touching*, and *presenting*, as well as the single-target subset of those gestures Sauppé and Mutlu would have categorized as *pointing*. The difference in our categorization schemes is thus grounded in a difference in focus on *motion* vs *purpose*. That is, while Sauppé and Mutlu primarily focuses on observable differences in how gestures are physically executed, we instead primarily focus on differences in what gestures are intended to achieve, while still taking into account the contextual information that the physical motion brings.

By analyzing these gestures in terms of purpose rather than motion alone, we can understand both when and why these gestures are used. In cases where the gesturer's purpose is to direct attention to a single object, they generate a precise gesture in which only the target appears in the cone, because (cf. Schauerte and Fink (2010)) if the gesture were less precise, and multiple objects fell into the cone, then additional effort would be needed to further pick out the object through other channels such as language.

Figure 3: "Precise Deictic Gesture": These are direct and unambiguous spatial gestures to a target referent. This figure shows the participant directly pointing at a block, demonstrating a "Precise Deictic Gesture" to directly specify a red rectangle as the referent of the gesture.

4.2 Small Region Deictic Gesture

The second category of Deictic gesture we observed was what we term "Small Region Deictic Gestures" which accounted for a plurality (40.6%) of all gestures observed. The apparent purpose of this gesture was to direct attention toward a small group of objects, either due to an intention to pick out the blocks as a group, or due to an intention to pick out a single object within the group, without certainty as to which object should be attended to. This type of gesture physically manifested as pointing toward the general area containing those objects, so that all objects in the group were entirely within the saliency cone extending from the gesturer's hand (cf. Lücking et al., 2015; Schauerte et al., 2010).

This category of gesture captures gestures that Sauppé and Mutlu (2014) would have categorized as *grouping gestures*, as well as the multi-target subset of those gestures Sauppé and Mutlu would have categorized as *pointing*. As above, then, the difference in our categorization schemes is grounded in a difference in focus on *motion* vs *purpose*.

Figure 4: "Small Region Deictic Gesture": These are direct and spatial gestures to a target referent or referents, but do not have a clear unambigous target. This figure shows the participant using a "Small Region Deictic Gesture" to point in the direction of a red rectangle and a yellow triangle, where the exact target referent of the gesture underspecified by the gesture without additional context or linguistic accompaniment.

4.3 Large Region Deictic gesture

The final category of obviously Deictic gesture we observed what we term "Large Region Deictic Gestures". These gestures were very rare (1.2%), and were only used by some participants. Yet they were distinct enough in purpose and form to warrant separate consideration. The apparent purpose of this gesture was to direct attention to a large number of objects comprising multiple clusters, or to a large, general region of the task environment. This type of gesture physically manifested as a large, potentially full-arm gesture in the direction of the objects of interest, but without attempting to fit those objects into a saliency cone.

This category of gesture captures gestures that Sauppé and Mutlu (2014) would have categorized as grouping or sweeping gestures. Our categorization is more broad, however. Because we focus on on purpose rather than motion, we do not restrict this category to those that manifest as literal whole-arm sweeps, but rather include any gesture whose *intent* is to highlight a large region. For example, we include observed instances in which a speaker waves their hand across multiple clusters of blocks.

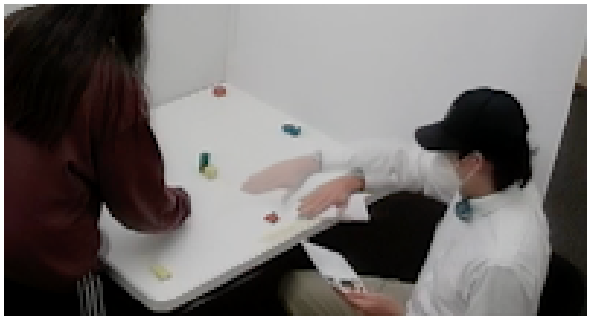

Figure 5: "Large Region Deictic gesture": These are large spatial gestures used to refer to many target referents. This figure shows the participant sweeping their hand over the top of the table demonstrating a "Large Region Deictic gesture" to refer to the red cube, the yellow cube, and the green cylinder.

4.4 Abstract Pointing Gesture

The next category of gesture we observed does not clearly fit into deictic gestures, nor does it clearly fit into another category, like metaphoric gesture. Rather, it represents a spectrum of gestures that fall somewhere between these categories. We term these gestures "Abstract Pointing Gestures". While not as common (around 9.7% of observed gestures) as deictic gestures, these gestures were consistently used at least once by all participants whose data was analyzed.

The apparent purpose of this gesture was to indicate that a target referent was "elsewhere", and *possibly* also to convey the direction in which the target referent was to be found. This gesture manifested in a variety of ways. In some cases, the gesture manifested as a point in the vague region of the target object, or in the direction of where the listener would need to go in order to begin traveling to the target. In either case, the gesture could be construed as casting an incredibly wide deictic cone; but we believe that the lack of precision in the gesture suggested that the speaker did not have a genuine expectation that the listener would follow their gaze, limiting the utility of modeling such a gesture as a cone.

Another reason for distinguishing this type of gesture from deictic gestures is other ways this could physically manifest that did not appear in this task-based dataset is due to the close distances between the speaker and all target referents, but which could manifest in other tasks with more varied out-of-context environments. A speaker trying to refer to something in another room, might wave generally or emblematically jerk their thumb over their shoulder. We would view these gestures as falling along the continuum of abstract gestures due to their shared intent.

This category of gesture captures those discussed by Stogsdill et al. (2021), McNeill et al. (1993)'s abstract deictic gestures, and some of the S-point gestures observed by Enfield et al. (2007). Our characterization of these gestures differs from that of Stogsdill et al. (2021), McNeill et al. (1993), and Enfield et al. (2007), however, in that we characterize them within a referential context, and ground them relative to other referential gestures within a comprehensive taxonomy. Additionally, they are grounded through our analysis of experimental data demonstrating how they are used in open-world task-based environments.

4.5 Iconic Reference Gesture

Thus far, we have been discussing gestures that are either clearly deictic, or that appear deictic in physical manifestation. We will now describe *iconic* gestures that we term "Iconic Reference Gestures", which we observed to also play a key role in referential communication within the open-world task-based interactions that we analyzed. This category

Figure 6: "Abstract Pointing Gesture": These are non-spatial pointing gestures used to create a gestural representation of a referent. This figure shows a participant using an "Abstract Pointing Gesture", pointing away from the table to refer to a block which has not seen before.

Figure 7: "Iconic Reference Gesture": These are non-spatial gestures used to mimic properties of the target referent. This figure shows the participant using an "Iconic Reference Gesture" to refer to a rectangular block by making a rectangular shape with their hands.

comprises a subset of the iconic gestures as delineated by McNeill and Levy (1982). However, we believe they are worth highlighting here as a separate category due to the referential purposes they achieve.

These gestures accounted for over one-tenth (11.8%) of the total gestures used. The apparent purpose of this gesture was to provide semantic content regarding the referent, to help disambiguate the semantic content of the speaker's speech. As such, this type of gesture physically manifested as mimicry of the shape of a referenced block to make the shape of a block (such as a rectangle, or semi-circle), or tracing out the shape of the block in the air with an index finger.

These gestures play an especially important role in open-world communication. In fact, these gestures were the most common method we observed when speakers intended to refer to objects that were not currently visible. Despite the observed priamcy of these gestures for open-world task-based reference, these iconic gestures are not well studied in task-based HRI.

5 Discussion

5.1 What types of referring gestures did we observe?

Our first research question was "How might we better categorize the types of referring gestures used in open-world task-based environments?" Our results show that a variety of different referring gestures are used in open-world task-based environments, including multiple types of deictic gestures, iconic gestures, and abstract gestures that trouble the previously delineated boundaries of traditional gesture categories.

As demonstrated above, our results suggest that deictic gestures may be best split into three categories, based on the intended specificity of the gesture. This suggests a need to shift from a focus on the physical form of different types of deictic gestures, to the way that speaker intent shapes gesture specificity.

Our results also highlight the need to consider iconic gestures when analyzing referring gestures. As demonstrated above, participants used a variety of iconic gestures to help communicate the properties of referents. We argue that these types of gestures are uniquely important to open-world task-based interactions, both to help describe the properties of previously-seen or as-yet-unseen objects – or to signal through the use of iconic gesture that those objects are not currently visible.

Finally, our results demonstrate the importance of Abstract Pointing Gestures; how these gestures are uniquely used in open-world task-based environments; and how imprecision and abstractness serve as tools to communicate this open-world status. These gestures, which do not fit cleanly into traditional gesture categories, demonstrate a need to think differently about gestures in open-world task-based environments, in a way that moves beyond traditional frameworks for categorizing gestures.

5.2 With what prevalence were different referring gestures observed?

Our second research question was "How prevalent are these different categories of gestures?", Table 1 shows the distribution of how the gestures observed in the analyzed dataset. While "Deictic Precise

Gestures" are nearly the only type of referring gesture explored in the task-based human-robot interaction literature, they comprised only one-third of the gestures we observed (36.6%), making them only the second most common gesture type observed. In contrast "Deictic Small Region Gestures" were the most common gesture, constituting nearly half of observed gestures (40.6%), and one-fifth (21.5%) of gestures used were non-deictic gestures. This troubles the dominant perspective that nearly exclusively focuses on precise deictic gestures.

5.3 Is the HRI community's focus on Precise Deictic Gestures warranted?

Our final research question was "Is the overwhelming focus of the HRI research community on precise, deictic, B-point gestures justified by this distribution of observed gestures". Based on the types of gestures observed in the experiment, and their frequency of observation, we believe this overwhelming and myopic focus on precise deictic gestures is not justified. While the 37% frequency of use for "Precise Deictic Gestures" represents a meaningful quantity, it does not justify near-exclusive focus. These results suggest that the HRI research community should dramatically expand its scope of work to consider a wider variety of referring gestures.

Specifically, we recommend focusing on narrowing down the broader category of deictic gestures into different categories of deixes based on target specificity, and to explore the use of non-deictic gestures and the role they play in open-world referential communication. Similarly, while saliency cones are a good way to detect the target of a pointing gesture, as we start to incorporate non-pointing gestures, it is no longer sufficient. Understanding and generation of gestures which cannot be modeled by a saliency cone may be difficult, but it is also necessary for effective communication the open-world tasks that are prevalent within HRI. Overall, we suggest the HRI community should move away from the coarse, generic gesture categorization schemes they have relied on for so long.

6 Conclusion

Clear communication is critical for both human-human and human-robot task-based interaction; and clear communication in many task-based domains requires effective referential gestures. While current gesture research in task-based HRI is nearly exclusively focused on precise deictic gestures, as we transition to more realistic open-world tasks, we will need to move beyond these gestures.

Our results show that precise deicitic gestures only play a small part in task-based communication, suggesting that HRI research perhaps needs to be "pointed" in a new direction for both gesture generation and gesture understanding. We propose a new gesture taxonomy that can refocus the conversation about gesture to include gestures that are more suitable for open-world environments. We argue that this taxonomy is more effective at capturing the types of gestures used in tasks for both open-world and closed-world environments.

Limitations

While our research shows the need to reevaluate gesture categorization, our results were limited due to the constraints of the experiment. One major limitations was the difficulty of seeing the direction of deictic gestures from the camera's perspective, and thus, of differentiating "Precise Deictic Gestures", "Small Region Deictic Gestures" and sometimes "Abstract Pointing Gestures". Another major limitation was that certain gestures may have been under- or over-represented in the dataset we analyzed. Specifically, large deictic gestures may have been under-represented due to the task requiring only a single item at time, reducing the need for simultaneous reference for multiple blocks.

Ethics Statement

While this research itself may not have overt ethical implications, work on gesture understanding and generation does present ethical implications. Specifically, work on gesture in robotics brings along risks to privacy via perception, and risks of over-trust due enhanced anthropomorphic morphology. A major requirement for gesture generation and understanding is advanced capabilities in robotic perception. These perception methods bring along risks of data privacy and security as it can be unclear if and how perceptual data can be used during and after the robotic interaction. Additionally generation of gestures enhances the anthropomorphic morphology of the robot, which can lead to a higher and potentially false perception of the robot's intelligence. This can lead to an overtrust in the robot's intelligence and capabilities, and can lead to potentially dangerous outcomes (Robinette et al., 2016).

Acknowledgements

This work has been supported in part by the Office of Naval Research grant N00014-21-1-2418.

References

Jens Allwood, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. 2007. The mumin coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, 41:273–287.

Cynthia Breazeal, Cory D Kidd, Andrea Lockerd Thomaz, Guy Hoffman, and Matt Berlin. 2005. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *2005 IEEE/RSJ international conference on intelligent robots and systems*, pages 708–713. IEEE.

Rehj Cantrell, Paul Schermerhorn, and Matthias Scheutz. 2011. Learning actions from human-robot dialogues. In *2011 RO-MAN*, pages 125–130. IEEE.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Will Culpepper, Thomas A Bennett, Lixiao Zhu, Rafael Sousa Silva, Ryan Blake Jackson, and Tom Williams. 2022. Ipower: Incremental, probabilistic, open-world reference resolution. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Nele Dael, Marcello Mortillaro, and Klaus R Scherer. 2012. The body action and posture coding system (bap): Development and reliability. *Journal of Nonverbal Behavior*, 36:97–121.

Antonella De Angeli, Walter Gerbino, Giulia Cassano, and Daniela Petrelli. 1998. Visual display, pointing, and natural language: the power of multimodal interaction. In *Proceedings of the working conference on Advanced visual interfaces*, pages 164–173.

Jan de Wit, Paul Vogt, and Emiel Krahmer. 2022. The design and observed effects of robot-performed manual gestures: A systematic review. *ACM Transactions on Human-Robot Interaction*.

Nick J Enfield, Sotaro Kita, and Jan Peter De Ruiter. 2007. Primary and secondary pragmatic functions of pointing gestures. *Journal of Pragmatics*, 39(10):1722–1741.

Brian Gleeson, Karon MacLean, Amir Haddadi, Elizabeth Croft, and Javier Alcazar. 2013. Gestures for industry intuitive human-robot communication from human observation. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 349–356. IEEE.

Susan Goldin-Meadow. 1999. The role of gesture in communication and thinking. *Trends in cognitive sciences*, 3(11):419–429.

Charles Goodwin. 2003. Pointing as situated practice. In *Pointing*, pages 225–250. Psychology Press.

Stephanie Gross, Brigitte Krenn, and Matthias Scheutz. 2017. The reliability of non-verbal cues for situated reference resolution and their interplay with language: implications for human robot interaction. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 189–196.

Zhao Han, Polina Rygina, and Thomas Williams. 2022. Evaluating referring form selection models in partially-known environments. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 1–14.

Zhao Han and Tom Williams. 2022. A task design for studying referring behaviors for linguistic hri. In *2022 ACM/IEEE International Conference on Human-Robot Interaction (HRI), Late-Breaking Report (LBR)*.

Adam Kendon. 1997. Gesture. *Annual review of anthropology*, 26(1):109–128.

Aelee Kim, Jooyun Han, Younbo Jung, and Kwanmin Lee. 2013. The effects of familiarity and robot gesture on user acceptance of information. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 159–160. IEEE.

Sotaro Kita. 2003. *Pointing: Where language, culture, and cognition meet*, chapter Pointing: A Foundational Building Block of Human Communication. Psychology Press.

Alfred Kranstedt, Andy Lücking, Thies Pfeiffer, Hannes Rieser, and Ipke Wachsmuth. 2005. Deixis: How to determine demonstrated objects using a pointing cone. In *International Gesture Workshop*, pages 300–311. Springer.

Andy Lücking, Thies Pfeiffer, and Hannes Rieser. 2015. Pointing and reference reconsidered. *Journal of Pragmatics*, 77:56–79.

Nikolaos Mavridis. 2015. A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems*, 63:22–35.

David McNeill. 1985. So you think gestures are nonverbal? *Psychological review*, 92(3):350.

David McNeill. 2003. Pointing and morality in chicago. In *Pointing*, pages 301–314. Psychology Press.

David McNeill, Justine Cassell, and Elena T Levy. 1993. Abstract deixis.

David McNeill and Elena Levy. 1982. Conceptual representations in language activity and gesture. *Speech, place, and action*, pages 271–295.

Raedy Ping and Susan Goldin-Meadow. 2010. Gesturing saves cognitive resources when talking about nonpresent objects. *Cognitive Science*, 34(4):602–619.

Paul Robinette, Wenchen Li, Robert Allen, Ayanna M Howard, and Alan R Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*, pages 101–108. IEEE.

Maha Salem, Stefan Kopp, Ipke Wachsmuth, Katharina Rohlfing, and Frank Joublin. 2012. Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics*, 4:201–217.

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

Boris Schauerte and Gernot A Fink. 2010. Focusing computational visual attention in multi-modal human-robot interaction. In *International conference on multimodal interfaces and the workshop on machine learning for multimodal interaction*, pages 1–8.

Boris Schauerte, Jan Richarz, and Gernot A Fink. 2010. Saliency-based identification and recognition of pointed-at objects. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4638–4643. IEEE.

Adam Stogsdill, Grace Clark, Aly Ranucci, Thao Phung, and Tom Williams. 2021. Is it pointless? modeling and evaluation of category transitions of spatial gestures. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pages 392–396.

Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. 2020. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3(1).

Stefan Waldherr, Roseli Romero, and Sebastian Thrun. 2000. A gesture based interface for human-robot interaction. *Autonomous Robots*, 9:151–173.

Tom Williams, Saurav Acharya, Stephanie Schreitter, and Matthias Scheutz. 2016. Situated open world reference resolution for human-robot dialogue. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 311–318. IEEE.

Tom Williams and Matthias Scheutz. 2015a. A domain-independent model of open-world reference resolution. In *CogSci*.

Tom Williams and Matthias Scheutz. 2015b. Power: A domain-independent algorithm for probabilistic, open-world entity resolution. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1230–1235. IEEE.