Deictic Adaptation in a Virtual Environment

Nikhil Krishnaswamy(✉) and James Pustejovsky

Department of Computer Science, Brandeis University,
415 South Street, Waltham, MA 02453, USA
{nkrishna,jamesp}@brandeis.edu
http://www.voxicon.net/

Abstract. As human-computer interfaces become more sophisticated, people expect computational agents to behave more like humans. However, humans interacting make assumptions about mutual conceptual understanding that they may not make when interacting with a computational agent, where spatial cues in the environment affect their assumptions about the agent's knowledge. In this paper, we examine an interaction between human subjects and a virtual embodied avatar displayed on a screen, wherein a surface displayed on the screen is either "continued" in the real world by a physical surface or not. Subjects are, with minimal instruction, asked to indicate objects displayed in the shared environment to the agent in the course of a collaborative task. We then examine the subjects' adaptations, in aggregate, to the different configurations.

Keywords: Spatial cognition · Deixis · Virtual agent
Embodiment · Spatial reasoning

1 Introduction

In person-to-person interactions, assumptions about the interlocutor and the world influence everything from communication style or "message design" [13] to available concept vocabulary and modalities [4]. If two people jointly experience a localized event, they can be said to be *co-situated* and *co-perceptive*. Additionally, if engaged in a collaborative task, they *co-intend* to complete the task and must *co-attend* to the situation. Coordination between multiple agents becomes particularly advantageous when each agent may have incomplete knowledge of the situation, but can rely on their interlocutor(s) to clarify or provide instructions, which is facilitated by imagining the situation from a different perspective [8], or at a deeper level by neural structures like mirror neurons [3]. These parameters come together in a theory of *common ground* [5,11,36,41]. A rich, diverse literature exists on assumptions and presuppositions underlying human communication (e.g., [10,41,42]), and we have previously examined these factors from a computational perspective continued in this line of research [38].

S. Creem-Regehr et al. (Eds.): Spatial Cognition 2018, LNAI 11034, pp. 180–196, 2018.
https://doi.org/10.1007/978-3-319-96385-3_13

Some problems in a strictly presuppositional view of common ground have been raised by Abbott [2], which can be mitigated by the introduction of such mechanisms as "accommodation" of non-controversial information (a la Lewis [28]), or reminding the interlocutor of known but non-forefronted information.

At least some of the assumptions underlying common ground are not in force when a human interacts with a non-human. Just as the common ground between a human and an animal is limited [23], so too is it between a human and a robot or virtual agent, as no mechanism for accommodation or reminding exists in a computer system unless put there by the developers. However, unlike an animal, a robot or embodied agent is often created with the intent to approximate human behavior, and as they become more sophisticated, humans come to see them as human-analogue and expect them to behave as such [12,16]. How, then, do these conflicting cues—a perhaps subconscious expectation of an embodied agent's human or near-human capability, plus the agent's lack of some of the more sophisticated mechanisms to communicate its own situational perception—manifest in an interaction where some understanding of common ground is both present and required to complete a shared task?

This paper examines one such angle, using a platform which integrates a multimodal model of semantics (Multimodal Semantic Simulations, *MSS*) [25,37] with a realtime vision system for recognizing human gestures [45]. The result [24,33] creates an environment where a human interacts with a virtual agent to communicate spatially-grounded instructions in a collaborative task, and we examine how human users adapt their deictic techniques based on variant spatial cues in the experimental setup, as a proxy for the underlying assumptions they make about their virtual interlocutor, her embodiment, and understanding.

1.1 Deixis in Virtual Environments

Humans do intuitively understand virtual worlds to be different from the real one, particularly if said virtual world appears on a screen while the spatial cues of the real world remain visible [40]. This presents an interesting problem for the transfer of spatial cognitive tasks between the virtual space and real space. Many virtual interaction systems integrate the virtual space with the real space in order to make the transition as natural as possible. This inclination of course presages virtual reality (VR) and augmented reality (AR) systems, and thus we end up with tables and walls that act as tablet surfaces [22,39] or for content-sharing [21], and computer vision-tracked interaction with surfaces using gestures [29].

One of the most basic spatially-grounded gestures is deixis. Many (e.g., Hostetter [20]) argue that gestures are simulated action, but Clark [11], Volterra [43] and others view gesture as a more general mode of reference. While our multimodal semantic modeling language, VoxML, treats gestures as a special case of action programs for generating MSS, gestures may also simulate *objects* by decoupling object attributes (e.g., size, shape, relative location) from the object itself, and binding them to some denotational aspect of the gesture. In the case of deixis by pointing, this aspect is the location of the object as interpreted by the pointing agent. The pointing gesture binds a location—in most cases, the

location denoted by the vector of the pointer (e.g., the finger) intersecting some salient area (e.g., a real or imagined surface plane). The pointing gesture might then refer to a location, or to objects occupying it (cf. [6]).

Using an utterance S and a corresponding gesture G there are three ways for an agent a to perform a communicative act C [36]:

$$\text{a. } C_a = (G) \qquad \text{b. } C_a = (S) \qquad \text{c. } C_a = (S, G)$$

If gesture and speech are temporally aligned, the agent may point to an object and say "that one" or to a location and say "there," and the utterance may select for an object versus a location, while the gesture can be formally realized as a snippet of a context-free grammar, e.g., $Point_G \rightarrow Loc \mid Obj$.

Deixis serves as a method of directing attention. Being temporally aligned with speech, the object indicated by deixis is usually also the current topic of discussion or conversation [9]. This expectation is also in effect in a virtual world, or co-situated worlds mediated by virtualization technology, such as video conferencing. Therefore a disconnect between agents due to a misalignment in their respective frames of reference, or information available to one agent that is invisible to the other, makes it difficult to agree on or to communicate which object or coordinate is being indicated by deixis [18]. Research in both kinematics [34] and human-computer interaction tasks [49] points to speed of pointing as an inverse correlate of the difficulty of the pointing task being performed.

Pointing is one of the most basic communicative gestures, as demonstrated by various studies [31,32,46]. The gesture set used in this line of research comes from studies by Wang et al. [44], wherein one human, the *builder*, has a table with blocks on it, and another human, the *signaler*, is given a pattern of blocks to build, invisible to the builder. As only the builder can move the blocks, the signaler must instruct the builder on how to construct the target pattern. Further details about these elicitation studies in particular are given in [44].

Because in the elicitation studies, the subjects were standing before tables, the gestures elicited naturally used the table as a reference point. A subject might first indicate a spot on the table and then another to indicate a relative location. Because the subjects were physically separated from each other, the signalers naturally fell into a pattern of using points on their own table surface to indicate blocks or positions on the builder's table surface. This turns the $Point_G \rightarrow Loc \mid Obj$ interpretation into a mirroring exercise where $Point_G \rightarrow Loc\prime \mid Obj\prime$, and the location indicated on the signaler's table space is translated into the builder's table space. In less-constrained situations, without the presence of a common reference point such as a table, many studies have shown that subjects naturally default to pointing relative to another context. This might be a free-floating point situated within an immersive virtual reality environment [47], or, when relevant information is displayed on a screen, the screen [19,30].

This setup, where the system requirements for accurate/fluent deixis conflict with users' documented tendencies when interacting with technology, creates an opportunity to study if and how users adapt their use of deixis to the system.

2 Experimental Design

2.1 Scenario and Data Capture

In our experimental scenario, users collaborated with an avatar to build a test pattern. All users were asked to build a six-block, three-stepped staircase using the blocks available. The definition of "success" was left up to the user, as far as placement of specifically colored blocks, orientation of the staircase, exactness of the blocks' alignment, etc. Users were told to use gesture and speech to achieve the goal but were not given the vocabulary of gestures and words understood by the avatar.

The purpose of the previously-mentioned elicitation studies was to observe and catalog the use of naturalistic gestures in the given task. Thus the gesture set in use is adapted to the environment in which the task is conducted, and it is these uses that were used to develop the avatar-interaction system (hereafter referred to as HAB, or "human-avatar-blocks world"). The data evaluated hereafter was gathered as part of a larger study evaluating the coverage of the HAB system [27]. Focusing specifically on the pointing data here allows us to use the results of this larger evaluation study to examine the particularities of deixis in a virtual environment with a virtual interlocutor.

Our experiment recreates the experimental setup from the aforementioned elicitation studies, except the builder is not a human in a physical room but an embodied avatar in a 3D world rendered on a monitor. This creates a parallel to the original elicitation study, where the human and the avatar are "separated" by the computer screen, and only the avatar can access the blocks.

Fig. 1. Scene with embodied avatar.

This virtual world is created using VoxSim [25,26], a semantically-informed 3D event simulator used for experiments in communicating with computational agents. VoxSim is built on the Unity game engine and contains a sophisticated

model of object and event semantics based on VoxML and dynamic logic which allows the agent to access existing context to interpret input from a human user in multiple modalities. Here we focus on natural language and gesture.

Gestures are recognized in real time using depth data from a Microsoft Kinect® [50] which is classified using ResNet-style deep convolutional neural networks (DCNNs) [17] implemented in TensorFlow [1]. The system recognizes 35 independent gestures that represent attributes or programs with semantics encoded in VoxML [37]. The HAB avatar's contextual interpretations of each gesture type are enumerated in [24,33]. Here we focus primarily on pointing, with supplemental gestures to communicate affirmation or negation. Figure 2 shows the VoxML semantics for a pointing gesture.

$$
\left[\begin{array}{l}
\textbf{point} \\
\text{LEX} = \left[\begin{array}{l} \text{PRED} = \textbf{point} \\ \text{TYPE} = \textbf{assignment} \end{array}\right] \\
\text{TYPE} = \left[\begin{array}{ll}
\text{HEAD} = & \textbf{assignment} \\
\text{ARGS} = & \left[\begin{array}{ll} \text{A}_1 = & \textbf{x:agent} \\ \text{A}_2 = & \textbf{y:finger} \\ \text{A}_3 = & \textbf{z:location} \\ \text{A}_4 = & \textbf{w:physobj•location} \end{array}\right] \\
\text{BODY} = & \left[\begin{array}{ll} \text{E}_1 = & extend(x, y) \\ \text{E}_2 = & def(vec(x \rightarrow y \times z), as(w)) \end{array}\right]
\end{array}\right]
\end{array}\right]
$$

Fig. 2. VoxML semantics for a [[POINT]] gesture. A_4, w, shows the compound typing (a la Generative Lexicon [35]) of the indicated region and objects within that region.

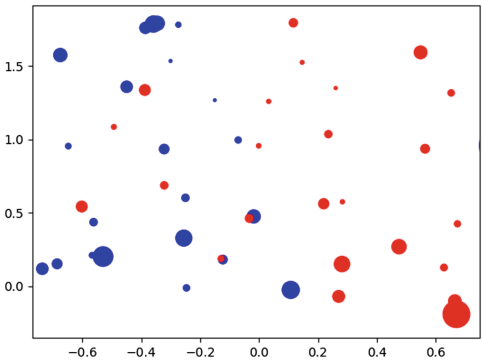

Fig. 3. The variance when a user points at a set of sampled points, using blue for left hand, red for right hand, and circle size proportionate to variance. The top edge of the figure represents space closer to the Kinect® while bottom edge represents space closer to the user. (Color figure online)

The HAB system uses a Kinect® positioned above and slightly behind the monitor displaying the avatar and the virtual world. Coordinates indicated on the user's table space are calculated relative to the Kinect® and then transposed into the virtual world. Therefore, if the user points at the coordinate represented in Fig. 3 as (0.0, 1.6), which approximately represents the point directly beneath the Kinect® on the table, this appears in the virtual world as the coordinate at the center of the table's "far" edge relative to the user (i.e., the edge closest to the avatar). The point represented in the figure as (0.6, 0.0) would appear in the virtual world as the point at the right-hand corner of the table closer to the human (rendered at the bottom of the screen). Coordinates were calculated by intersecting the vector extending through the forearm and extended finger with an infinite plane situated at the height of the physical table's surface. These coordinates were transposed into equivalent coordinates in the virtual world and visualized as a rotating purple target (as shown in Fig. 1). As the user moved their arm, this target also moved in real time, which allowed the user to situate the coordinate at which they were pointing in the virtual world. When the user was pointing at a coordinate interpreted as being on the surface of the virtual table, the target was displayed at full opacity, while if it wandered off the table's surface, the target decreased in opacity proportional to the distance of the interpreted coordinate to the table's surface. This cue allowed the user to see the target while they attempted to move it back onto the table.

We logged all interactions in the following format: `INDEX, SYMBOL, CONTENT, TIMESTAMP`. `SYMBOL` is a two letter code consisting of `H` (human) or `A` (avatar) followed by `S` (speech/utterance), `G` (non-pointing gesture), `P` (pointing) or `A` (action). `CONTENT` is the shallow semantic content of the move, including coordinates where applicable. `TIMESTAMP` is the time in seconds after the beginning of the interaction. Further details of logging and evaluation scheme may be found in [27]. For this study we focus only the circumstances surrounding human pointing gestures (prefixed `HP`). Table 1 shows an example snippet of a log file captured during the user studies.

Table 1. Sample trial log (abbreviated)

```
1 HG engage start 1.145281
2 AS "Hello." 1.145281
3 HP r,-0.25,-0.87 4.889832
4 HP r,-0.16,-1.21 4.928307
5 HP r,-0.07,-1.18 4.960413
6 HP r,-0.03,-1.06 5.040221
7 HP r,-0.09,-0.95 5.072867
8 HP r,-0.07,-0.27 5.15642
...
73 HP r,-0.08,11.69 8.552608
74 HG right point high,-0.02,5.45 8.588802
75 AS "Are you pointing here?" 8.588802
```

2.2 Experimentation and Evaluation

We ran user trials on 20 subjects taken from the university graduate student population. All subjects were computer-literate but had no knowledge of the particulars of interaction and had no prior experience using this integrated gesture system. Subjects were divided evenly in one of two environments, shown in Fig. 4. Kinects were placed at the same height in both setups and centered above and slightly behind the widescreen monitor. The test pattern was displayed on an iPad to the left of the screen for the subject's reference. In the setup with the table (**Environment A**), users stepped up to the table's edge to begin the interaction. In the setup without the table (**Environment B**), they stepped up to the blue tape line on the floor.

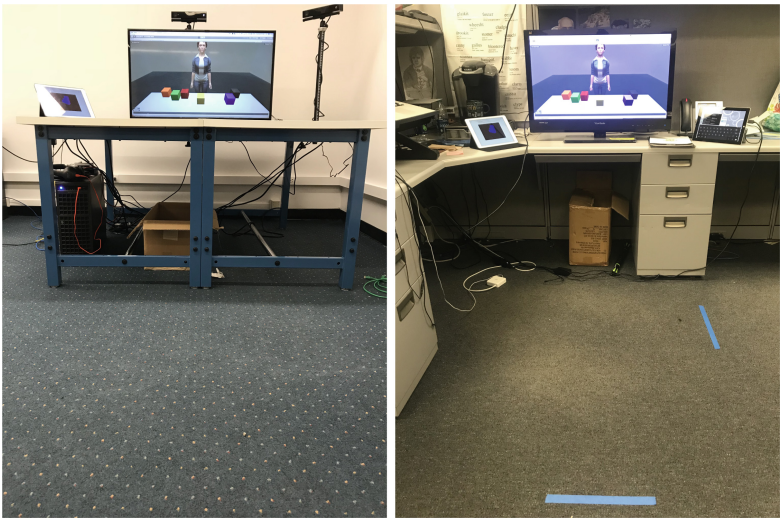

Fig. 4. Variant conditions with table, and monitor placed at the rear of the table (left) and without table, with monitor placed at the front edge of the supporting desk (right). The lines on the floor in the right-hand image demarcate the bounds (projected downward) of the imaginary table used for calculating pointing coordinates.

The 10 subjects in each environment were also divided into two conditions: half those in Environment A were explicitly told to consider the real table as an extension of the virtual table into the real world, as if the avatar were another person standing behind a glass screen or window, and half those in Environment B were explicitly told to imagine that the virtual table extended out of the screen into the real world, as if the avatar were another person standing behind a glass screen or window. This drew attention to the table or imagined table space and served as an implicit "hint" that it had some role to play in the interaction. The remaining subjects in each environment were not given any extra cues on considering the virtual table. This allowed us two small samples to examine how

this extra information affected the users' pointing strategies, and we end up with 4 distinct experimental conditions (Table 2).

Interactions were logged from start to finish, defined as the point at which the user decided that the test pattern was built to their satisfaction, and stepped away from the table's edge/demarcated line.

Table 2. 5 subjects were placed in each experimental condition.

Condition	Physical table	Supplemental information
1	Present (**A**)	None
2	Absent (**B**)	None
3	Present (**A**)	Physical table extends virtual table
4	Absent (**B**)	Virtual table extends into real world

In evaluating the data, we were interested in the time it takes a user in a given condition to settle on a particular position on the table after beginning their pointing gesture. As discussed in Sect. 1.1, deixis may either refer to a location or be coerced to a reference to objects that occupy that location. So, once the system has recognized that the subject has stopped moving their finger and is pointing at a specific location, the system then asks for a confirmation of the location. This question can take many forms based on context, e.g., "Are you pointing here/at this?" (if the user is just starting the interaction), "The <color> block?" (if deixis lands in an area containing one or more blocks), "Should I place something here?" (if deixis lands in a region empty of blocks but a block has been previously indicated), among others. No matter what the question or the context that prompted it, the user must answer it with a *positive acknowledgment* (the word "yes" or a thumbs-up gesture) or a *negative acknowledgment* (the word "no" or a thumbs-down). We define a *successful pointing* as a pointing followed by a positive acknowledgment (that is, the user pointed to a spot that the system recognized and the user confirmed), and a *failed pointing* as a pointing followed by a negative acknowledgment (the user pointed to a spot, the system recognized a different spot, and the user denied that this was correct). The time taken to successfully point is extracted from the log file as the time from the commencement of pointing (e.g., move 3 in Table 1) to the recognition of the location (move 74 in Table 1), but only in those blocks where the pointing event is succeeded by a positive acknowledgment. If a user adapts their deictic strategy to the system, intuitively these times to complete a successful pointing should decrease as the user proceeds further into the interaction. We can model the adaptation in pointing times as a *learning rate* and examine in which conditions users adapt a strategy more quickly.

3 Results and Discussion

We aggregated the data from all sessions of all users in a single condition, and removed outliers, defined as those times lying outside the interquartile range (IQR), for the distribution of all times logged, independent of experimental condition. Since each session may span a different length of time from start to finish, we cannot use the raw duration of an interaction as the independent variable when plotting results, so we normalized by plotting a user's pointing times against the *percentage* of the total interaction completed to that point. We plotted the postprocessed data in two ways:

(1) The raw times taken to complete successful pointing events against the percentage of interaction completed. This allows us to assess a learning curve (see below) for an average user in a given condition and see if the raw time to successfully complete a pointing declines over the course of an interaction, stays flat, or increases.
   – According to Wright's cumulative average model [48], a *learning curve* is modeled as a power law: $y_n = ax^b$. y_n is the average time to "produce the first n units" (in this context, the average time to successfully point to a location on the table the first n times for each trial subject in a given condition). Therefore a is the time to successfully point the first time, and b is the natural slope of the learning curve (over raw times we will denote this b_ρ), which reflects whether learning proceeds rapidly or slowly. A percentage, $s = 2^b$, can be used to express how much the time to point in that environment can be expected to increase or decrease each time the number of pointing events doubles. $s < 1$ (negative b_ρ) indicates increasing adaptation as the interaction proceeds in the condition under examination, as successive points take less time overall. $s > 1$ (positive b_ρ) indicates increasing confusion or difficulty in successfully pointing.

(2) The *ratio* between the time to complete a logged successful pointing event and the *geometric mean* time to complete a successful pointing in that condition, against the percentage of interaction completed.
   – Since we aggregate all data, and since individual users might take longer or shorter on average to indicate a location than others, taking the difference from the mean allows us to normalize some of the variation due to a given subject's natural level of aptitude with the system. Using a ratio rather than a difference allows us to use the geometric mean of the recorded values and thereafter plot the line of best fit using a linear regression, which represents a more intuitive analogue of the learning curve achieved by taking the log of both sides: $log\ y_n = log\ a + b\ log\ x$. The slope of the line, b_μ, reflects changes in a user's ability to successfully point relative to their normalized mean pointing time—regarded as a proxy for the user's natural "set point" ability to successfully indicate a location to the system—in the condition under examination. As above, negative b_μ indicates adaptation to the system and positive b_μ indicates increasing confusion over time.

The Figs. 5 and 6 below show the aggregate data plotted for each experimental condition (see Table 2). In all graphs the X-axis displays the users' progress through their interaction trials, represented as a percentage. In the graphs on the left, the Y-axis shows the time to complete a successful pointing event. On the right, the Y-axis shows the time to complete a successful pointing event, as a ratio to the geometric mean of all recorded pointing times for the user whose time is plotted. Line of best fit is shown as a least-squares fitted power law (left), and a linear regression (right), with b for each curve displayed in the caption.

Overall, the data tends to be dispersed when plotted against best-fit lines, even when removing IQR outliers. Nevertheless, most points tend to cluster near the bottom of the plotted distributions, between 0–2 s for the raw pointing times, and close to a 1:1 ratio of individual pointing times to geometric mean. We can observe a few trends that contrast between experimental conditions, and we expect these trends, where they appear, would be more pronounced with larger sample size, possibly with a higher r^2 value.

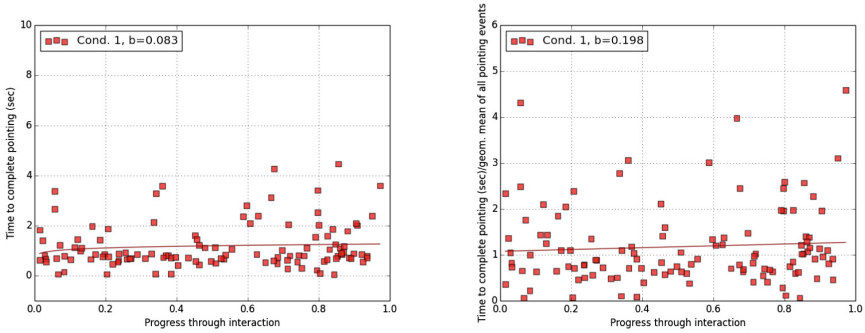

Fig. 5. Results in condition 1. $b_\rho \approx 0.083$, $s \approx 1.059$; $b_\mu \approx 0.198$

In Condition 1, both lines are almost flat, with a very slight upward curve ($b_\rho \approx 0.083$, $s \approx 1.059$; $b_\mu \approx 0.198$), suggesting that users did not adapt a more

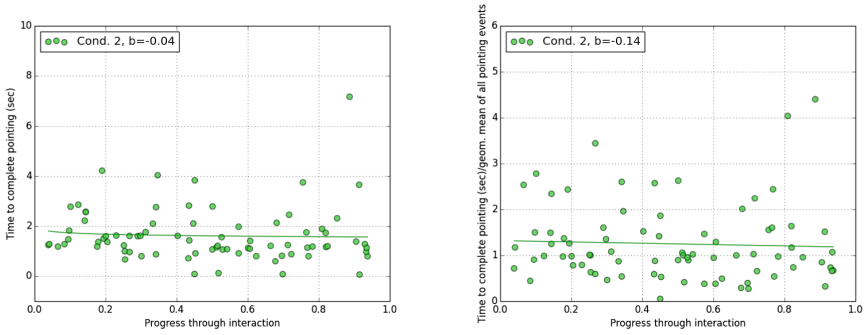

Fig. 6. Results in condition 2. $b_\rho \approx -0.044$, $s \approx 0.970$; $b_\mu \approx -0.144$

efficient pointing strategy by the end of an interaction when compared to the start. The positive b values suggest that users in this condition became slightly *less* efficient at pointing, and although the values are not large enough to be very significant, it does suggest that using the system in this condition (in the presence of the physical table, with no additional information about its role in the setup) may create some confusion for the user.

In Condition 2, both lines are about as flat the best-fit lines in Condition 1, trending very slightly downward ($b_\rho \approx -0.044$, $s \approx 0.970$; $b_\mu \approx -0.144$), but also not enough to draw a firm conclusion. Users appear not to adapt a significantly more efficient pointing strategy in this condition (without the physical table, with no additional information about the table's role), or if so, the learning rate was not fast enough to make an apparent difference over a single interaction (Fig. 7).

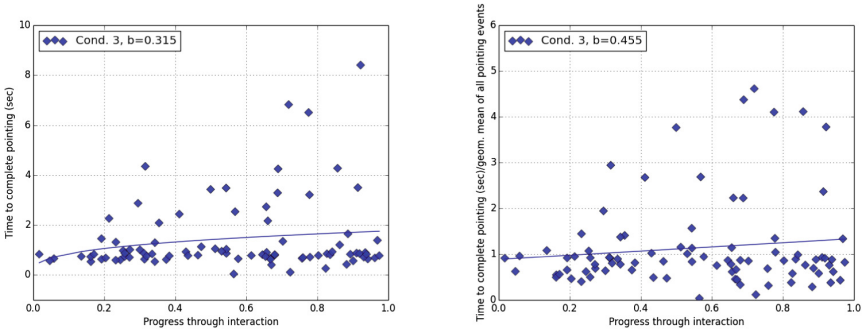

Fig. 7. Results in condition 3. $b_\rho \approx 0.315$, $s \approx 1.245$; $b_\mu \approx 0.455$

Condition 3 demonstrates a negative learning rate (increasing time to point successfully as the interaction goes on). As the interactions proceed, the pointing times get notably more dispersed and the divergent values trend away from the

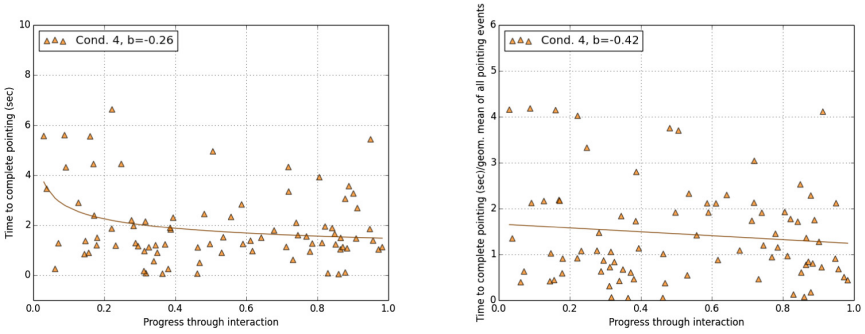

Fig. 8. Results in condition 4. $b_\rho \approx -0.265$, $s \approx 0.832$; $b_\mu \approx -0.427$

geometric mean of users' pointing times ($b_\rho \approx 0.315$, $s \approx 1.245$; $b_\mu \approx 0.455$). Users in this condition (with the physical table, told to regard it and the virtual table as extensions of each other) appear to display increasing difficulty in indicating a location successfully (Fig. 8).

Condition 4 is the only condition in which users display a marked ability to adapt a more efficient pointing strategy over the course of the interaction ($b_\rho \approx -0.265$, $s \approx 0.832$; $b_\mu \approx -0.427$). Users in this condition (without the physical table, told to imagine that the virtual table extends into the real world) appear, in aggregate, to be able to successfully point about 17% faster by each point at which their cumulative number of successful pointing events has doubled (i.e., from 1 to 2 or from 2 to 4, etc.).

4 Conclusion

When we examine the minimal pair of conditions 1/3 vs. 2/4 (that is, the conditions with the physical table and the conditions without the physical table, regardless of users' knowledge of the table's role), we see a trend of increasing difficulty in successfully pointing in conditions with the table, and a trend of more efficient pointing in conditions without the table. This is actually the opposite of what we expected, in that the presence of the table did not seem to provide the users with a reference point with which to ground their deictic gestures and in fact seemed to make pointing more difficult, indicating that it introduced a measure of confusion to the interaction, perhaps causing them uncertainty about which was the valid reference point, the table or the screen.

In Conditions 3 and 4, the difference between the "table" condition and "tableless" condition is more pronounced than in Conditions 1 and 2. The nearly flat lines in Conditions 1 and 2 suggest that users barely changed their pointing strategies in reaction to the system at all. We can speculate that in both conditions they tended to settle on a particular strategy (most likely pointing at the monitor screen/toward the avatar, as suggested by the literature, which was also anecdotally observed during the trials[1]), and persisted with it through the trials, making minimal changes despite difficulties encountered. Presented with no suggestion about how the table might be used in the interaction, subjects adhered to their initial pointing strategies.

Meanwhile, subjects in Conditions 3 and 4, where they are given a prompt about the table's role, display either marked adaptation or marked confusion. Subjects in Condition 4 typically demonstrate appreciable adaptation in pointing strategies over the course of their interactions, while in Condition 3 the time subjects took to successfully point *increased* regularly as the interactions went on. We can hypothesize that, when their attention was drawn to the (physically present) table, users set about trying to use it in the interaction, and if they did not meet with initial success (e.g., were unable to quickly figure out that they should use it to mirror positions on the virtual table), grew confused. In Condition 4, without the physical table present, users could perhaps more

[1] No personal information or video of the subjects was captured.

easily imagine the virtual table extended into a space in their world and point to coordinates relative to their own bodies (and body-centered coordinate systems do seem to be natural and ingrained in spatial representation [15]) without the distraction of another object filling the space between themselves and the virtual world. Put simply, it's possible that the presence of the physical table imposed extra cognitive load on the task of trying to imagine the virtual table extending out of the monitor. All this points to difficulties in situating oneself in "mixed-reality" environments [7,14], perhaps due to the cognitive load involved in transforming one's embodied coordinate system to that of the virtual world. Further research would be needed to examine whether the precise phrasing of the relationship between the physical and virtual tables might that have any influence in the results, such as on the mental transformations being performed.

Due to the nature of the experimental setup (e.g., distractors not accounted for in the surrounding environment or that emerge during the course of the interaction), we should forward some caveats and possible alternate explanations for some phenomena. In some conditions, pointing became more difficult in the later stages of the trial. It may be that in some cases, as the target structure emerged, it became more difficult to accurately point at the desired location with the greater need for precision and increased density of blocks. However, this increasing difficulty only emerges in some conditions, so in the others, it may be overridden by the learning adaptation. In addition, the test subjects were allowed free reign to adapt their overall strategy for the building task (i.e., for actions supervenient on the individual vocabulary items and gestures such as pointing), so if pointing at a particular location proved difficult, they adapted their overall plan by (for example), moving objects to new locations entirely (by pointing) or by loosening the constraints on what they determined to be successful actions (e.g., allowing spaces between the blocks so that block location could be easier to indicate by pointing).

In both physical conditions, providing explicit instructions on how to conceive of the task led to more marked results than not providing any guidance. This suggests that the person's model of the situation matters, as well as the physical situation itself.

Due to the partition of subjects into four conditions, we were only able to run five subjects in each condition. As such, these results should be considered tentative. Nonetheless, the conclusions suggested by the results are intriguing, worth further examination, and may become more pronounced in studies with more subjects.

The data we have gathered suggests that when interacting with a virtual environment on a screen, humans have a strong preference for indicating positions relative to that screen, even when physical cues are present that imply that the displayed scene is not the entirety of the environment involved. When these factors are merely implicit, they seem to have little effect on subjects' behavior, but when more attention is drawn to them, subjects are either able to adapt their behavior or the environment, or find the physical cues in conflict with their assumptions about the virtual world. Although deixis is just one part

of interacting with a virtual world, it is an important one, and this insight into how humans treat deixis in a virtual environment should be useful to developers seeking to build intelligent systems capable of interacting fluently with humans.

Acknowledgments. The authors would like to thank the reviewers for their helpful comments. We would also like to thank our colleagues at Colorado State University and the University of Florida for developing the gesture recognition systems: Prof. Bruce Draper, Prof. Jaime Ruiz, Prof. Ross Beveridge, Pradyumna Narayana, Isaac Wang, Rahul Bangar, Dhruva Patil, Gururaj Mulay, Jason Yu, and Jesse Smith; and our Brandeis University colleagues, Tuan Do and Kyeongmin Rim, for their work on VoxSim. Additional thanks to Jason for providing Fig. 3. This work is supported by a contract with the US Defense Advanced Research Projects Agency (DARPA), Contract W911NF-15-C-0238. Approved for Public Release, Distribution Unlimited. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: TensorFlow: a system for large-scale machine learning. In: Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI), Savannah, Georgia, USA (2016)
2. Abbott, B.: Presuppositions and common ground. Linguist. Philos. **31**(5), 523–538 (2008)
3. Arbib, M., Rizzolatti, G.: Neural expectations: a possible evolutionary path from manual skills to language. Commun. Cogn. **29**, 393–424 (1996)
4. Arbib, M.A.: From grasp to language: embodied concepts and the challenge of abstraction. J. Physiol. Paris **102**(1), 4–20 (2008)
5. Asher, N., Gillies, A.: Common ground, corrections, and coordination. Argumentation **17**(4), 481–512 (2003)
6. Ballard, D.H., Hayhoe, M.M., Pook, P.K., Rao, R.P.: Deictic codes for the embodiment of cognition. Behav. Brain Sci. **20**(4), 723–742 (1997)
7. Benford, S., Greenhalgh, C., Reynard, G., Brown, C., Koleva, B.: Understanding and constructing shared spaces with mixed-reality boundaries. ACM Trans. Comput.-Hum. Interact. (TOCHI) **5**(3), 185–223 (1998)
8. Bergen, B.K.: Louder than Words: The New Science of How the Mind Makes Meaning. Basic Books, New York (2012)
9. Brooks, A.G., Breazeal, C.: Working with robots and objects: Revisiting deictic reference for achieving spatial common ground. In: Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction, pp. 297–304. ACM (2006)
10. Clark, H.H., Brennan, S.E.: Grounding in communication. In: Resnick, L., Levine, B., John, M., Teasley, S.D. (eds.) Perspectives on Socially Shared Cognition, pp. 13–1991. American Psychological Association (1991)
11. Clark, H.H., Schreuder, R., Buttrick, S.: Common ground at the understanding of demonstrative reference. J. Verbal Learn. Verbal Behav. **22**(2), 245–258 (1983)

12. David, N., Bewernick, B.H., Cohen, M.X., Newen, A., Lux, S., Fink, G.R., Shah, N.J., Vogeley, K.: Neural representations of self versus other: visual-spatial perspective taking and agency in a virtual ball-tossing game. J. Cogn. Neurosci. **18**(6), 898–910 (2006)
13. Edwards, A., Shepherd, G.J.: Theories of communication, human nature, and the world: associations and implications. Commun. Stud. **55**(2), 197–208 (2004)
14. Flintham, M., Benford, S., Anastasi, R., Hemmings, T., Crabtree, A., Greenhalgh, C., Tandavanitj, N., Adams, M., Row-Farr, J.: Where on-line meets on the streets: experiences with mobile mixed reality games. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 569–576. ACM (2003)
15. Fogassi, L., Gallese, V., Di Pellegrino, G., Fadiga, L., Gentilucci, M., Luppino, G., Matelli, M., Pedotti, A., Rizzolatti, G.: Space coding by premotor cortex. Exp. Brain Res. **89**(3), 686–690 (1992)
16. Fussell, S.R., Kiesler, S., Setlock, L.D., Yew, V.: How people anthropomorphize robots. In: Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction, pp. 145–152. ACM (2008)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
18. Hindmarsh, J., Fraser, M., Heath, C., Benford, S., Greenhalgh, C.: Object-focused interaction in collaborative virtual environments. ACM Trans. Comput. Hum. Interact. (TOCHI) **7**(4), 477–509 (2000)
19. Hindmarsh, J., Heath, C.: Embodied reference: a study of deixis in workplace interaction. J. Pragmatics **32**(12), 1855–1878 (2000)
20. Hostetter, A.B., Alibali, M.W.: Visible embodiment: gestures as simulated action. Psychon. Bull. Rev. **15**(3), 495–514 (2008)
21. Izadi, S., Brignull, H., Rodden, T., Rogers, Y., Underwood, M.: Dynamo: a public interactive surface supporting the cooperative sharing and exchange of media. In: Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology, pp. 159–168. ACM (2003)
22. Johanson, B., Hutchins, G., Winograd, T., Stone, M.: PointRight: experience with flexible input redirection in interactive workspaces. In: Proceedings of the 15th Annual ACM Symposium on User Interface Software and Technology, pp. 227–234. ACM (2002)
23. Kirchhofer, K.C., Zimmermann, F., Kaminski, J., Tomasello, M.: Dogs (canis familiaris), but not chimpanzees (pan troglodytes), understand imperative pointing. PLoS ONE **7**(2), e30913 (2012)
24. Krishnaswamy, N., Narayana, P., Wang, I., Rim, K., Bangar, R., Patil, D., Mulay, G., Ruiz, J., Beveridge, R., Draper, B., Pustejovsky, J.: Communicating and acting: understanding gesture in simulation semantics. In: 12th International Workshop on Computational Semantics (2017)
25. Krishnaswamy, N., Pustejovsky, J.: Multimodal semantic simulations of linguistically underspecified motion events. In: Barkowsky, T., Burte, H., Hölscher, C., Schultheis, H. (eds.) Spatial Cognition/KogWis - 2016. LNCS (LNAI), vol. 10523, pp. 177–197. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68189-4_11
26. Krishnaswamy, N., Pustejovsky, J.: VoxSim: a visual platform for modeling motion language. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. ACL (2016)
27. Krishnaswamy, N., Pustejovsky, J.: An evaluation framework for multimodal interaction. In: Proceedings of LREC (2018, forthcoming)

28. Lewis, D.: Scorekeeping in a language game. J. Philos. Logic **8**(1), 339–359 (1979)
29. Malik, S., Ranjan, A., Balakrishnan, R.: Interacting with large displays from a distance with vision-tracked multi-finger gestural input. In: Proceedings of the 18th Annual ACM Symposium on User Interface Software and Technology, pp. 43–52. ACM (2005)
30. Moeslund, T.B., Störring, M., Granum, E.: A natural interface to a virtual environment through computer vision-estimated pointing gestures. In: Wachsmuth, I., Sowa, T. (eds.) GW 2001. LNCS (LNAI), vol. 2298, pp. 59–63. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-47873-6_6
31. Morris, M.R., Huang, A., Paepcke, A., Winograd, T.: Cooperative gestures: multi-user gestural interactions for co-located groupware. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1201–1210. ACM (2006)
32. Morris, M.R., Wobbrock, J.O., Wilson, A.D.: Understanding users' preferences for surface gestures. In: Proceedings of Graphics Interface 2010, pp. 261–268. Canadian Information Processing Society (2010)
33. Narayana, P., Krishnaswamy, N., Wang, I., Bangar, R., Patil, D., Mulay, G., Rim, K., Beveridge, R., Ruiz, J., Pustejovsky, J., Draper, B.: Cooperating with avatars through gesture, language and action. In: Intelligent Systems Conference (IntelliSys) (2018, forthcoming)
34. Papaxanthis, C., Pozzo, T., Schieppati, M.: Trajectories of arm pointing movements on the sagittal plane vary with both direction and speed. Exp. Brain Res. **148**(4), 498–503 (2003)
35. Pustejovsky, J.: The Generative Lexicon. MIT Press, Cambridge (1995)
36. Pustejovsky, J.: From actions to events: communicating through language and gesture. Interact. Stud. **19**(1) (2018)
37. Pustejovsky, J., Krishnaswamy, N.: VoxML: a visualization modeling language. In: Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, May 2016
38. Pustejovsky, J., Krishnaswamy, N., Draper, B., Narayana, P., Bangar, R.: Creating common ground through multimodal simulations. In: Proceedings of the IWCS Workshop on Foundations of Situated and Multimodal Communication (2017)
39. Scott, S.D., Grant, K.D., Mandryk, R.L.: System guidelines for co-located, collaborative work on a tabletop display. In: Kuutti, K., Karsten, E.H., Fitzpatrick, G., Dourish, P., Schmidt, K. (eds.) ECSCW 2003, pp. 159–178. Springer, Heidelberg (2003). https://doi.org/10.1007/978-94-010-0068-0_9
40. Spence, I., Feng, J.: Video games and spatial cognition. Rev. Gen. Psychol. **14**(2), 92 (2010)
41. Stalnaker, R.: Common ground. Linguist. Philos. **25**(5–6), 701–721 (2002)
42. Tomasello, M., Carpenter, M.: Shared intentionality. Dev. Sci. **10**(1), 121–125 (2007)
43. Volterra, V., Caselli, M.C., Capirci, O., Pizzuto, E.: Gesture and the emergence and development of language. Beyond nature-nurture: Essays in honor of Elizabeth Bates, pp. 3–40 (2005)
44. Wang, I., Narayana, P., Patil, D., Mulay, G., Bangar, R., Draper, B., Beveridge, R., Ruiz, J.: EGGNOG: a continuous, multi-modal data set of naturally occurring gestures with ground truth labels. In: To appear in the Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition (2017)

45. Wang, I., Narayana, P., Patil, D., Mulay, G., Bangar, R., Draper, B., Beveridge, R., Ruiz, J.: Exploring the use of gesture in collaborative tasks. In: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, CHI EA 2017, pp. 2990–2997. ACM, New York (2017). https://doi.org/10.1145/3027063.3053239
46. Wobbrock, J.O., Morris, M.R., Wilson, A.D.: User-defined gestures for surface computing. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1083–1092. ACM (2009)
47. Wraga, M., Creem-Regehr, S.H., Proffitt, D.R.: Spatial updating of virtual displays. Mem. Cogn. **32**(3), 399–415 (2004)
48. Wright, T.P.: Learning curve. J. Aeronaut. Sci. **3**(1), 122–128 (1936)
49. Zhai, S., Kong, J., Ren, X.: Speed-accuracy tradeoff in fitts' law tasks-on the equivalency of actual and nominal pointing precision. Int. J. Hum. Comput. Stud. **61**(6), 823–856 (2004)
50. Zhang, Z.: Microsoft kinect sensor and its effect. IEEE MultiMedia **19**, 4–10 (2012)