Cooperating with Avatars Through Gesture, Language and Action

Pradyumna Narayana*, Nikhil Krishnaswamy†, Isaac Wang‡, Rahul Bangar*, Dhruva Patil*, Gururaj Mulay*, Kyeongmin Rim†, Ross Beveridge*, Jaime Ruiz‡, James Pustejovsky†, and Bruce Draper*
*Department of Computer Science, Colorado State University, Fort Collins, CO USA
†Department of Computer Science, Brandeis University, Waltham, MA USA
‡Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL USA

***Abstract*—Advances in artificial intelligence are fundamentally changing how we relate to machines. We used to treat computers as tools, but now we expect them to be agents, and increasingly our instinct is to treat them like peers. This paper is an exploration of peer-to-peer communication between people and machines. Two ideas are central to the approach explored here: *shared perception*, in which people work together in a shared environment, and much of the information that passes between them is contextual and derived from perception; and *visually grounded reasoning*, in which actions are considered feasible if they can be visualized and/or simulated in 3D.**

We explore shared perception and visually grounded reasoning in the context of blocks world, which serves as a surrogate for cooperative tasks where the partners share a workspace. We begin with elicitation studies observing pairs of people working together in blocks world and noting the gestures they use. These gestures are grouped into three categories: social, deictic, and iconic gestures. We then build a prototype system in which people are paired with avatars in a simulated blocks world. We find that when participants can see but not hear each other, all three gesture types are necessary, but that when the participants can speak to each other the social and deictic gestures remain important while the iconic gestures become less so. We also find that ambiguities flip the conversational lead, in that the partner previously receiving information takes the lead in order to resolve the ambiguity.

***Keywords*—Gesture Recognition, Human Computer Interfaces, Artificial Intelligence**

I. Introduction

Advances in artificial intelligence are fundamentally changing how we relate to machines. We used to treat computers as tools, but now we expect them to be agents, and increasingly our instinct is to treat them like peers [1]. For example, we talk to them and give them personal names (e.g. Alexa, Siri, Cortana). Unfortunately, the more familiar we become with artificial agents, the more frustrated we become with their limitations. We expect them to see and hear and reason like people. *No, no, Alexa, can't you see that I ...*

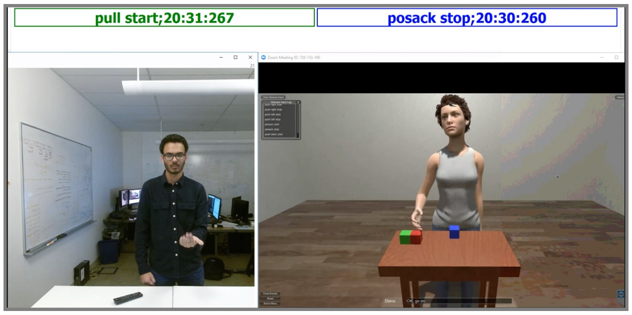


Fig. 1. Prototype peer-to-peer interface. The signaler on the left can communicate through gestures and words. The avatar on the right can communicate through gestures, words and actions.

This paper is an exploration of ideas in peer-to-peer communication between people and machines. It considers what capabilities a machine might need, and presents a prototype system with a limited form of peer-to-peer communication. Two ideas are central to the approach explored here. The first is *shared perception*. When people work together, much of the information that passes between them is contextual and derived from perception. Imagine, for example, two people cleaning a room. They might discuss high-level strategy ("you start here, I'll start over there"), but they don't describe every action they take to each other. They can just look at the room to see what the other person has or has not done. Only the high-level discussion is verbal, and even that is grounded in perception: the definitions of "here" and "over there" depend on knowing where the other person is. In general, when one person changes the state of the world, the other person can see it, and the goal of conversation is to provide additional information beyond what is provided by perception.

The second idea is that reasoning about physical objects is grounded in visualization. Imagine a simple command like "put the book on the table". Is this command feasible? Yes, if there exists both a book and a table, and there is a clear spot on the table at least the size of the book, and if there is a clear path from the book's position to the table. Most of the information needed to evaluate this command comes from perception, and in general the command can be understood if it can be visually simulated.

We explore these ideas in blocks world. In particular, we consider a scenario in which one person (the *builder*) has a table with blocks on it, and another person (the *signaler*) is given a target pattern of blocks. Only the builder can move the blocks, so the signaler has to tell the builder what to do. While blocks world is obviously not a real-world application, it serves as a surrogate for any cooperative task with a shared workspace.

We begin our exploration with elicitation studies similar to Wobbrock *et al*. [2], but with differences in how gestures are elicited. In the original elicitation study format, people are given specific actions (called referents), and asked to create a user-defined gesture (called signs) for the action. In our study, we take a more natural approach and simply present a pair of people with a task to complete and observe the actions and gestures that naturally occur. In our elicitation studies, the signaler and builder are both people. They are in separate rooms, connected by a video link. We vary the communication between them across three conditions: (1) the signaler and builder can both see and hear each other; (2) the signaler and builder can see but not hear each other; and (3) the signaler and builder can only hear each other (the signaler can see the builder's table and knows where the blocks are, but cannot see the builder).

Using gestures observed in the elicitation studies, we develop a prototype system in which the signaler is a person but the builder is an avatar with a virtual table and virtual blocks. The signaler can see a graphical projection of the virtual world, and communicate to the avatar through speech and gesture. The avatar communicates back through speech, gesture, and action, where an action is to move a block. Figure 1 shows the set up, with the signaler on the left and the avatar on the right in her virtual world.

Experience derived from using the prototype reveals important features of human-computer cooperation on shared physical tasks. For example, we learned that complex, gesture-based peer-to-peer conversations can be constructed from relatively few gestures, as long as the gesture set includes: (1) social gestures, for example acknowledgement and disagreement; (2) deictic gestures, such as pointing; and (3) iconic gestures mimicking specific actions, such as pushing or picking up a block. When the builder and avatar are allowed to speak, words can replace the iconic gestures, but the social and deictic gestures remain important. We learned that ambiguities arise in the context of conversations not just from questions of reference, i.e. which block to pick up, but also from options among actions, for example whether to put a block down on top of another block or next to it. Fortunately, these ambiguities are easily resolved if the conversational lead is allowed to switch from the signaler to the builder (in this case, the avatar). We also came to appreciate the importance of making two or more gestures at the same time, for example nodding (a social gesture) while signaling for the builder to pick up an object. Finally, we learned how important it was for the avatar to gesture back to the signaler, even when the avatar can speak and move blocks.

From an engineering perspective, we also confirmed that the combination of depth images from inexpensive sensors (Microsoft Kinect v2s) and deep convolutional neural networks is sufficient to recognize 35 common hand poses, and that with GPUs these hand poses can be recognized in real time. The directions of arm movements are also easily detected and are needed for deixis and for supplying directions to representational actions such as push or carry.

II. Related Work

This paper explores multi-modal peer-to-peer communication between people and avatars in shared perceptual domains. As such, it touches on multiple topics that have been studied before, including human/avatar interaction, multi-modal interfaces, and simulation semantics for reasoning, although we know of no previous system that integrates all of these components.

We have long known that people respond differently to avatars than to non-embodied interfaces. Users generally have a more positive attitude toward avatars, and often try to make themselves appear better to the avatar. They also tend to assign personality to avatars [3]. Although usually good, this can backfire: if the avatar is unable to meet a user's goals, the user is more likely to get angry [4]. People respond better to avatars and virtual robots than to non-embodied interfaces, but they respond better still to physically present robots [5]. Although the work here focuses on human/avatar interactions, it should extend interactions between humans and humnoid robots as well.

Multimodal interfaces combining language and gesture have been around since at least 1980, when Bolt introduced "Put-that-there" [6]. Bolt's work anticipated the use of deixis to disambiguate references. More importantly, it inspired a community of researchers to work on multimodal communication, as surveyed in [7] and [8].

Roughly speaking, there are two major motivations for multimodal interfaces. The psychological motivation, as epitomized by Quek *et al.* [9], holds that speech and gesture are coexpressive, and therefore compliment each other. People are able to process speech and gesture partially independently, so using both modalities to express information increases human working memory and decreases the cognitive load [7]. People therefore retain more information and learn faster when communicating multimodally.

Visual information has been shown to be particularly useful in establishing common ground [10], [11], [12], which is important in shared perception scenarios. Other research emphasizes the importance of video and shared visual workspaces in computer-mediated communication [13], [14], [15], [16], and highlights the usefulness of non-verbal communication to support coordination between humans. Thus, multimodal interfaces that leverage these human factors have the potential to be a more effective collaborators.

This second motivation for multimodal interfaces is practical, as epitomized by Reeves *et al.* [17]. They argue that multimodal interfaces increase the range of users and contexts. For example, a device that can be accessed by either voice or gesture command can be used both in the dark and in noisy environments. They also argue that multimodal interfaces improve security and privacy since, depending on the situation, voice commands might be overheard or gestures might be observed. In addition, Veinott *et al.* draw the implication that the inclusion of video (and gestural information) may be increasingly useful for communication in the presence of language barriers [18].

This paper concentrates on shared physical tasks. When people work together, their conversation consists of more than just words. They gesture and share a common workspace [19], [20], [21]. Their shared perception of this workspace supports simulation semantics, and it is this shared space that gives many gestures such as pointing their meaning [22].

If two beings communicate to complete a shared task, they can be considered agents, who are not only co-situated and co-perceiving but also act, together or individually, in response to communication. To coordinate action there must be agreement of a common goal between the agents, which can be called co-intent. Together, co-situatedness, co-perception, and co-intent are the first aspects of common ground. There is a rich and diverse literature on grounded communication [10], [23], [24], [25], [26]. However, in joint tasks, agents share an additional anchoring strategy—the ability to co-attend. This ability emerges as central to determining the denotations of participants in shared events. Experienced events differ from events as expressed in language, as language allow us to package, quantify, measure, and order our experiences, creating rich conceptual reifications and semantic differentiations. The surface realization of this ability is mostly manifest through linguistic utterances, but is also witnessed in gestures.

Simulation can play a crucial role in human computer communication by creating a shared epistemic model of the environment. Simulation also creates an environment where two parties may be co-situated and co-attend by giving the agent an explicit *embodiment* [27], and allows the agent to publicly demonstrate its knowledge, providing an additional modality to communicate shared understanding within object and situation-based tasks, such as those investigated by [28] [29] and [30].

The simulation environment provided includes the perceptual domain of objects, properties, and events. In addition, propositional content in the model is accessible to the discourse, allowing them to be grounded in event logic (a la [31]), and to be distinguished by the agents to act and communicate appropriately. This provides the non-linguistic visual and action modalities, which are augmented by the inherently non-linguistic gestural modality enacted within the visual context.

III. Case Study: Communicating with Gesture, Language and Action

To explore the role of gesture in peer-to-peer communication with shared perception, we first conducted human subject studies. The goal of these studies is to elicit common gestures and their semantic intents for the blocks world task in order to gain insight about how they might be used by people. We then built a prototype human/avatar system to explore communicating with computers.

A. Elicitation Studies

We begin with the human subject study design depicted in Figure 2. As mentioned before, our goal is to elicit a set of gestures and possible actions in blocks world by observing people as they naturally communicate and collaborate with each other to complete a task. Each trial has two subjects, a signaler and a builder. Both subjects stand at the base of a table with a monitor on the other end. A two-way video feed is set up to allow both people to interact as if they were at opposite ends of a long table. The builder is given a set of wooden blocks, while the signaler is given a block layout/pattern. The task is for the signaler to tell the builder how to recreate the pattern of blocks without showing the pattern to the builder. The use of a computer-mediated setup allows us to control the communication allowed based on condition.

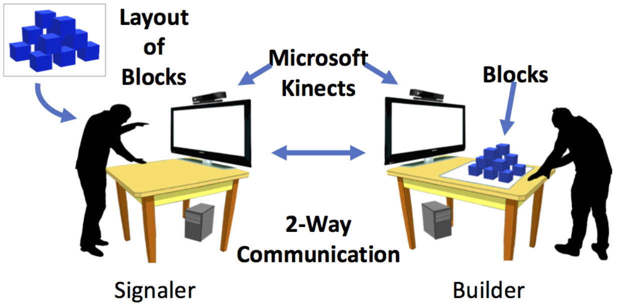


Fig. 2. Human subject study designed to elicit gestures for the blocks world domain and provide insight about how those gestures are used.

TABLE I. Time to completion (min:sec) for signaler/builder blocks world tasks under three conditions: audio+video, video only, and audio only.

condition	trials	min	median	mean	max
video+audio	188	0:26.2	0:48.1	1:06.0	6:07.1
video only	181	0:31.7	1:03.6	1:32.9	8:27.4
audio only	170	0:09.8	1:06.3	1:35.2	13:36.1

A total of 439 trials across 60 participants were conducted under the following three conditions:

- Audio + Video: Participants can both see and hear each other through the monitors.
- Video-only: Participants can see but not hear each other, requiring the use of non-verbal communication only.
- Audio-only: Participants can only hear each other (the shared workspace is maintained: the signaler can still see the builder's table and know where the blocks are, but only cannot see the builder).

In all three conditions, RGB-D video is captured of both the signaler and builder using a Microsoft Kinect v2; the Kinect also estimates the 3D coordinates of 17 visible body joints[1] in each frame.

The data set collected and some initial observations have been described elsewhere [**?**]. Not previously reported, however, are the results below including the overall impact of the visual gestures. As shown in Table I, participants were able to finish the task in 1:06 (min:sec) on average in the audio+video condition. In the audio-only condition, however, where the signaler was only able to talk to the builder, the average time increased to 1:32.9. This is very similar to the average time to completion for the video-only case, which was 1:35.2. This suggests that gestures are almost as communicative in blocks world as words, and more importantly that words and gestures are not redundant. Their combination is better than either words or gestures alone, in alignment with [9], [18].

Overall, we collected about 12.5 hours of data. The audio+video and video-only trials were hand labeled at the level of left and right hand poses, left and right arm motions, and head motions. Summarizing these labels, we discovered 110 combinations of poses and motions that occurred at least 20

[1] The Kinect v2 estimates the positions of 25 joints, but the 8 lower-body joints are consistently obscured by the table.

TABLE II. SEMANTIC GESTURES PERFORMED AT LEAST 20 TIMES IN TOTAL BY AT LEAST 4 DIFFERENT HUMAN SUBJECTS.

Numeral	**Representational**	**Deictic**	**Social**
one	grab	point (that/there)	start
two	carry	tap (this/here)	done
three	push	this group	positive ack
four	push (servo)	column	negative ack
five	push together	row	wait for
	rotate		wait (pause)
			emphasis

times and were performed by at least 4 different subjects. Of these, 29 were determined to have no semantic intent, as when a subject drops their arms to their side. Of the 81 remaining gestures, many were either minor variations or enantiomorphs of each other. For example, a participant might make the “thumbs up” sign while raising their forearm or pushing it forward. Similarly, the “thumbs up” sign might be made with the right hand, the left hand, or both. We also grouped physically different but semantically similar poses, such as the “thumbs up” and “OK” signs. After grouping similar motions and poses, we were left with 22 unique semantic gestures, as shown in Table II.

The 22 semantic gestures fall into four categories. Deictic gestures, such as pointing or tapping the table, serve to denote objects or locations. Iconic gestures, such as grab, push or carry, mimic actions. Social gestures, such as head nods or thumbs up, address the state of the dialog. Numerals are a form of abstract plural reference. We note that there are broader and more inclusive schemes for categorizing gestures (see [32], chapter 6), but none are universally accepted and the simple categories above work well for describing the gestures we observed in blocks world.

B. Blocks World Prototype

The elicitation studies provide insights about gestures people use in blocks world. Our goal, however, is to explore peer-to-peer communication between people and computers using visually-grounded reasoning in the context of shared perception. To this end, we created a prototype system that replicates the experimental set-up in Figure 2, except that the builder is now an avatar and the blocks and table are virtual. The system operates in real time, and allows the human signaler to gesture and speak to the avatar. The avatar can gesture and speak in return, as well as move blocks in the virtual world. In some tests we turn off the audio channel, thereby eliminating words and limiting communication to gestures and observation.

This system gives us a laboratory for exploring peer-to-peer communication between people and computers. Using it, we have spent hours building and tearing down simple blocks world structures, and the lessons learned from this experience are summarized in the next section. The rest of this section describes the human/avatar blocks world (HAB) system itself.

HAB has many components. For the purposes of this paper, however, we concentrate on the perceptual module that implements gesture recognition, the grounded semantics module (VoxSim) which determines the avatar’s behavior, and the interplay between perception and reasoning. The perceptual module is described in subsection III-B1, VoxSim is described in subsection III-B2, and the interactions between perception and VoxSim are described in subsection III-B3.

1) Perception: We hypothesize that shared perception is the basis for peer-to-peer communication, particularly when working in a common workspace. To implement shared perception, the human signaler and avatar builder need to be able to see each other as well as the virtual table and its blocks. Perceiving the virtual table and blocks is relatively easy. The avatar has direct access to the virtual world, and can directly query the positions of the blocks relative to the table. The human builder sees the rendering of the virtual world, and therefore knows where the blocks are as well.

More challenging is the requirement for the human and avatar to see each other and interpret each other’s gestures. Based on our elicitation studies, we have a lexicon of commonly occurring gestures. We animate the avatar so that she can perform all of these gestures, and rely on the builder’s eyes to recognize them. We also have an RGB-D video stream of the human captured by the Microsoft Kinect v2. The rest of this section describes the real-time vision system used to recognize the builder’s gestures in real time.

Gesture recognition is implemented by independently labeling five body parts. The left and right hands are labeled according to their pose. The system is trained to recognize 34 distinct hand gestures in depth images, plus a 35th label (“other”) that is used for hands at rest or in unknown poses. The hand poses are directional, in the sense that pointing down is considered a different pose than pointing to the right. Head motions are classified as either nod, shake or other based on a time window of depth difference images. Finally, the left and right arms are labeled according to their direction of motion, based on the pose estimates generated by the Microsoft Kinect [33].

To recognize gestures in real time, the computation is spread across 6 processors, as shown in Figure 3. The processor shown on the left is the host for the Microsoft Kinect, whose sensor is mounted on top of the signaler’s monitor. It uses the Kinect’s pose data to locate and segment the signaler’s hands and head, producing three streams of depth images. The pose data also becomes a data stream that is used to label arm directions. The hand and head streams are classified by a ResNet-style deep convolutional neural network (DCNN) [34]. Each net is hosted on its own processor, with its own NVIDIA Titan X GPU. The arm labeling process has its own (non-GPU) processor. Finally, a sixth processor collects the hand, arm and head labels and fuses them using finite state machines to detect gestures.

2) VoxSim: The avatar’s reasoning system is built on the VoxSim platform [22], [35]. VoxSim is an open-source, semantically-informed 3D visual event simulator implemented in Unity [36] that leverages Unity’s graphics processing, UI, and physics subsystems.

VoxSim maps natural language event semantics through a dynamic interval temporal logic (DITL) [37] and the visualization modeling language VoxML [38]. VoxML describes and encodes qualitative and geometrical knowledge about objects and events that is presupposed in linguistic utterances but not made explicit in a visual modality. This includes information about symmetry or concavity in an object’s physical structure,

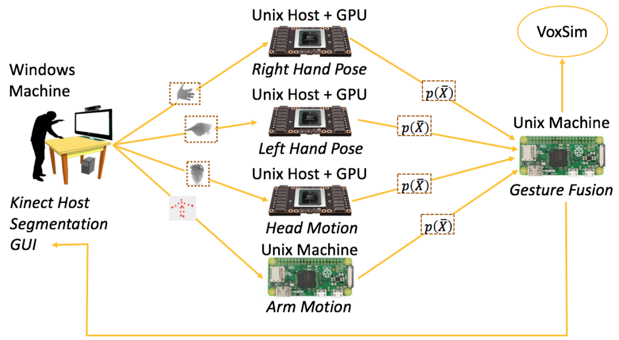


Fig. 3. The architecture of the real-time gesture recognition module.

the relations entailed by the occurrence of an event in a narrative, the qualitative relations described by a positional adjunct, or behaviors *afforded* by an object's *habitat* [39], [40] associated with the situational context that enables or disables certain actions that may be undertaken using the object. Such information is a natural extension of the lexical semantic typing provided within Generative Lexicon Theory [41], towards a semantics of embodiment. This allows our avatar to determine which regions, objects, or parts of objects may be indicated by deictic gestures, and the natural language interface allows for explicit disambiguation in human-understandable terms. The movement of objects and the movement of agents are compositional in the VoxML framework, allowing VoxSim to easily separate them in the virtual world, which means that the gesture used to refer to an action (or program) can be directly mapped to the action itself, establishing a shared context in which disambiguation can be grounded from the perspective of both the human and the computer program.

3) Perception & VoxSim: To create a single, integrated system we connect the recognition module (and by extension, the human signaler) with VoxSim and its simulated world. VoxSim receives "words" from the gesture recognizer over a socket connection, and interprets them at a contextually-wrapped compositional semantic level. The words may be either spoken or gestured by the (human) builder. For the moment, we have seven multi-modal "words":

1) *Engage*. Begins when the signaler steps up to the table or says "hello", and ends when they step back or say "goodbye". Indicates that the signaler is engaged with the avatar.
2) *Positive acknowledge*. Indicated by the word "yes", a head nod, or a thumbs up pose with either or both hands. Used to signal agreement with a choice by the avatar or affirmative response to a question.
3) *Negative acknowledge*. Indicated by the word "no", a head shake, a thumbs down pose with any combination of hands, or a stop sign gestured with the hand closed, palm forward, and fingertips up. Signals disagreement with a choice by the avatar or negative response to a question.
4) *Point*. Gestured by extending a single finger, with the optional spoken words "this" or "that". The information given to VoxSim about pointing gestures includes the spot on the tabletop being pointed to. Signals either a block to be used for a future action, or an empty space.
5) *Grab*. Indicated by a claw-like pose of the hand that mimics grabbing a block or the word "grab". Tells the avatar to grab a block that was previously pointed to.
6) *Carry*. Indicated by moving the arm while the hand is in the grab position, with the optional spoken word "carry". The information given to VoxSim includes a direction, one of left, right, forward, back, up or down. A "carry up" can be thought of as pick up, and a "carry down" is equivalent to put down.
7) *Push*. Gestured with a flat, closed hand moving in the direction of the palm, with the optional spoken word "push". Similar to carry, it includes a direction, although *up* and *down* are not allowed. As a special case, a beckoning gesture signals the avatar to push a block toward the signaler.

In addition to the multi-modal "words", there are words that can only be spoken, not gestured. These words correspond to the block colors: black, red, green, blue, yellow and purple. Speech recognition is implemented by simple word spotting, so for example the phrase "the red one" would be interpreted as the single word "red."

The flow of information from the avatar/builder back to human/signaler is similar. The avatar can say the words and perform the gestures mentioned above, with the additional gesture of reaching out and touching a virtual block (a gesture not available to the human builder). One important difference, however, is that the avatar can also communicate through action. Because the human signaler can see the avatar and the virtual blocks world, they can see when the avatar picks up or moves a block.

Any time the recognition module determines that one of the known "words" begins or ends, VoxSim receives a message. VoxSim responds by parsing the meaning of the gesture in context. For example, if the gesture points to a spot on the right side of the table and the avatar is currently holding a block, then the gesture is a request to move the block to that spot. Alternatively, if the avatar is not holding a block, the same gesture selects the block nearest to the point as the subject of the next action.

Gestural ambiguities are common. If two blocks are near each other, a pointing gesture in their direction is an ambiguous. Which block did the signaler point to? Similarly, if the user says "the red one" when there are two red blocks on the table, the reference is ambiguous. Actions may also be ambiguous. If the user signals the avatar to put down one block near another, should it stack the blocks or put them side by side?

When presented with ambiguities, VoxSim assumes the initiative in the conversation and asks the user to choose among possible interpretations. In the case of pointing, for example, VoxSim might ask "do you mean the red block?". If the answer is negative, it might then try "do you mean the green block?". VoxSim orders the options according to a set of heuristics that favor interesting interpretations over less interesting ones. For example, if the options are to stack a red block on top of a blue block or put them next to each other, VoxSim favors the stacking option, because stacks are interesting.

IV. Lessons Learned

The motivation for the human studies and prototype system described in this paper is to gain first-hand experience with multi-modal peer-to-peer interfaces. We believe the prototype is unique, not because it recognizes natural gestures but because it interprets those gestures using simulation-based reasoning in the context of a shared perceptual task.

Our experience with HAB has taught us many lessons, and led us to quickly modify and improve it. The next subsection walks the reader through an example of a person and an avatar working together to build a block pattern. The remaining subsections capture and share some of the lessons we have learned through experience, with one important caveat: so far, the only users of the system are its design team. Usability studies with naive users will come later, when the system is more mature.

A. *Example*

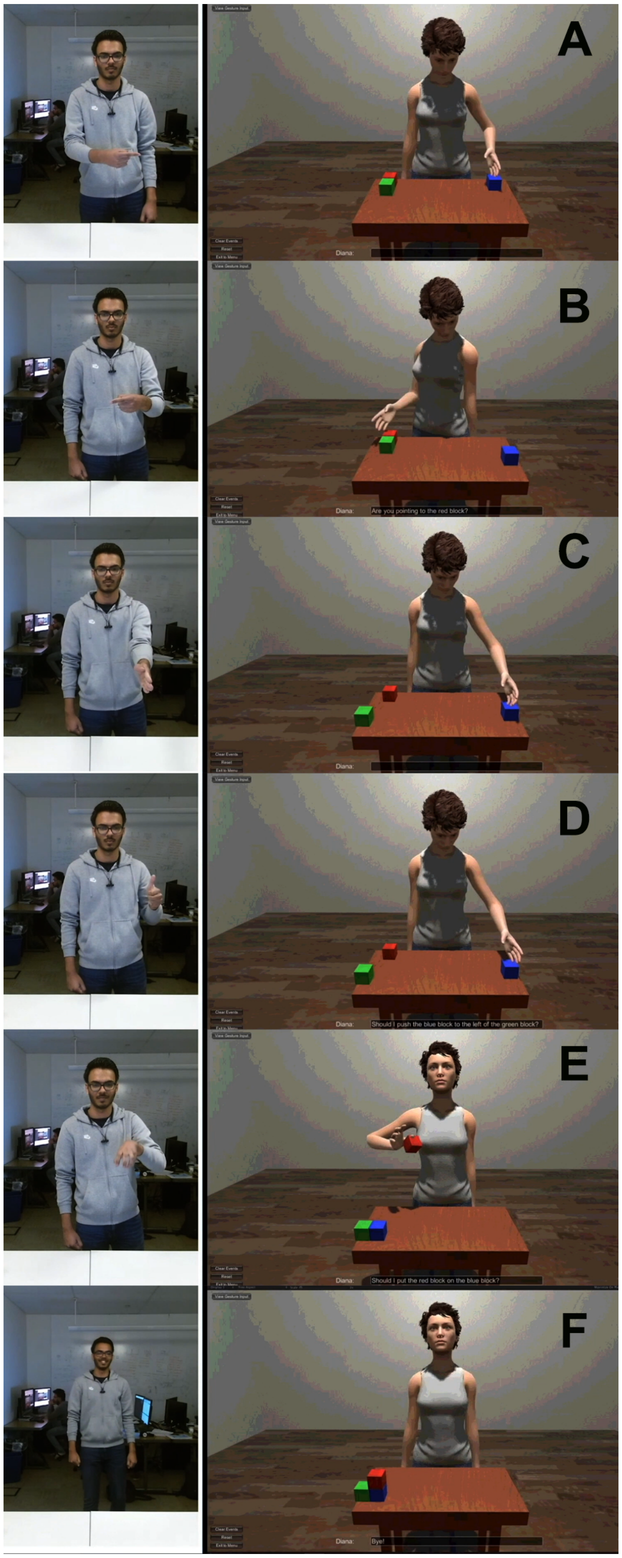
We illustrate HAB with an example in which the audio has been turned off, so all communication happens through gestures and actions. The example begins with three blocks on the table: a green block and a red block to the right of the signaler, and a blue one on the left. The signaler's goal is to arrange the blocks in a staircase. The conversation begins when the signaler steps up to the table, causing an *engage* gesture to be recognized and sent to VoxSim.

The signaler points to the left, as shown in Frame A of Figure 4. VoxSim, interprets this gesture as selecting the blue block for the next action. The avatar moves its hand toward the blue block in anticipation; this is a gesture that serves as a form of positive acknowledgement, since it lets the signaler know what the avatar understood. The signaler then beckons, and the avatar pushes the block away from itself and toward the signaler.

Next the signaler points to his right where the red and green blocks are (Frame B of Figure 4). This is an ambiguous reference, so the avatar reaches toward the red block as a way of asking whether the signaler means the red block. The signaler shakes his head, sending a negative acknowledgement, so the avatar motions toward the green block. This time the signaler nods, resolving the ambiguity. The signaler then beckons again, and the avatar pushes the green block toward the signaler.

Continuing with the example, the signaler points toward the blue block and gestures to slide it to the right (Frame C). The slide gesture is ambiguous, however. Should the avatar slide the block a little ways to the right, or slide it all the way to the green block? Sliding it to the green block is the more interesting option, so this is the one the avatar suggests, and since it is what the signaler wants, he gives a thumbs up (Frame D) and the avatar slides the block.

This style of interaction continues. The signaler selects the red block by pointing and then mimics a grabbing motion. Both the reference and action are unambiguous, so the avatar complies. Next the signaler raises his arm while keeping his hand in the grabbing pose, and brings his arm forward. The avatar responds as shown in Frame E. The signaler then lowers his arm and releases his grip, asking the avatar to put the block



Diana:

Diana: Are you pointing to the red block?

Diana:

Diana: Should I push the blue block to the left of the green block?

Diana: Should I put the red block on the blue block?

Diana: Bye!

Fig. 4. An example of building a staircase in HAB using only gestures and actions (no spoken language).

down. The placement is ambiguous – should the red block go on the blue block, the green block, or the table top? – but some gestural back and forth quickly clear this up, and the staircase is completed, as shown in Frame F.

B. The Uses of Gesture Types (Qualitative Observations)

What did we learn from many interactions like these? We learned qualitative lessons about the roles of different gesture types in dialogs with shared perceptual domains. We were then able to make comparisons between these qualitative conclusions and quantitative results from our human subject studies, suggesting similarities between our human/avatar prototype and true human/human interactions. Finally, we were able to measure the accuracy with which our system recognizes natural human gestures.

The gestures elicited from our human subjects are divided into four categories: numeric, deictic, iconic, and social, as listed in Table II. HAB currently recognizes 1 deictic gesture (pointing), 3 iconic gestures (grab, carry, and push), and 3 social gestures (engage/disengage, positive acknowledge, and negative acknowledge). The gesture recognition component also recognizes the numbers one through five, although these are not currently supported by the reasoning module (VoxSim).

The iconic gestures for grab, push and carry were among the first gestures we thought to integrate into the system, and more iconic gestures will be added in future work, e.g. stack and rotate. After all, in a physical domain like blocks world, iconic gestures tend to directly represent the underlying actions. Not all action gestures are representational: the beckoning motion used to draw a block toward the user is a social convention, and doesn't mimic a human action. Nonetheless, iconic gestures tend to correspond to actions and be representational.

Interestingly, users feel less comfortable with iconic gestures than deictic ones. The prototype can be run in two modes, with and without the audio channels (i.e. words). We do not have quantitative measures because the prototype is not yet robust enough for naive user studies, but when the audio channel is on, users tend to say the words *grab, push* or *carry* rather than make the gestures (sometimes they do both). This is true even though the words push and carry have to be expanded with directional phrases (e.g. "to the left").

Pointing, on the other hand, seems completely natural. Users do it almost without thinking, and do it whether or not the audio channel is available. This may be because the alternative often requires a description ("the red block" or "the red block on the left" if there are two), although we noted above that actions may also require descriptions in form of directions.

Social gestures turn out to be critically important. They maintain the structure of the dialog. In an early version of HAB, the human users could only gesture and the avatar could only act and ask disambiguating questions. The system was uncomfortable to use, because the user would point to a block and then not be sure whether the avatar had seen the gesture or if they should point again. Ironically, it was almost better when the pointing gesture was ambiguous, because then the avatar would ask a clarifying question. We then gave the avatar her first gesture, reaching toward a block when it was referenced as a form of positive acknowledgement. Immediately the users became more comfortable and tasks were completed more quickly.

The timing of seeking and giving acknowledgement is important. We tested a version of the prototype in which the avatar always waited for confirmation before taking actions. This was slow and frustrated the users. More importantly, because acknowledgements were so common the conversation would often hit an impasse when the avatar was waiting for an acknowledgement that the signaler thought they had already given.

For comfort, peer-to-peer dialogs require that when one partner provides information, the other acknowledges receiving it. This is sometimes called *backchannel* communication. If the information is a request for an action, such as grab, then performing the action is sufficient acknowledgement. If the information is ambiguous, asking to clarify it also serves as acknowledgement. In all cases, however, some form of positive acknowledgement is required from whichever partner, human or avatar, receives the information.

Because acknowledgements are so common, it is important that they be unintrusive. Too many verbal acknowledgements quickly become annoying. The ability to positively acknowledge through head nods is important because it can be done without interrupting the audio stream and without interrupting other gestures by the hands.

Acknowledgements (positive or negative) in response to ambiguity are important, as they allow the system to engage in the conversational act of repair [42], or the use of clarifying and re-referring to correct misunderstandings in discourse. The recognition of the social gestures mentioned before are key to the system's ability to tackle ambiguity. These gestures enable natural and rapid feedback in order to complete tasks and allow the system to function as a human-like conversational agent.

C. Human/Human vs Human/Avatar Gesture Usage

The observations above were qualitative, based on our own experience. Studies with naive users are being planned, but further development is required to make sure that system artifacts don't distract naive users and invalidate the data. What we can do now, however, is test if our qualitative predictions match quantitative data from the human/human interaction studies. If so, this supports our predictions and suggests that there are at least similarities between human/humann interaction and our prototype human/avatar system.

As shown in Section III-A, we have labeled data from over 180 trials of human/human blocks world interactions in both the audio+video and video-only conditions. Based on our qualitative predictions, social gestures – particularly positive and negative acknowledgements – should outnumber iconic and deictic gestures in both conditions. Iconic gestures should appear more often in the video-only condition than the audio+video condition. Deictic gestures should appear with similar frequencies in both conditions.

Table III shows the frequencies of gestures in the human/human studies, organized by condition and gesture type.

TABLE III. FREQUENCIES OF GESTURES IN HUMAN DATA IN THE VIDEO ONLY AND VIDEO+AUDIO CONDITIONS, ORGANIZED BY GESTURE CATEGORY. THE RATIO IS FREQUENCIES BETWEEN THE TWO CONDITIONS IS A MEASURE OF HOW MUCH MORE LIKELY WAS A GESTURE TO BE MADE WHEN THE AUDIO WAS TURNED OFF, VERSUS WHEN WORDS ARE AVAILABLE.

Gesture	Video Only	Video + Audio	Total	Ratio
Iconic Gestures				
Translate	234	73	307	3.2
Rotate	225	67	292	3.4
Separate	126	125	251	1.0
Servo Translate	206	44	250	4.7
Bring Together	83	38	121	2.2
Servo Toward	35	20	55	1.7
Iconic Total	**909**	**367**	**1,276**	**2.5**
Deictic Gestures				
This Block	240	94	334	2.6
That/There	150	106	256	1.4
Here/This	111	42	153	2.6
This Group	86	28	114	3.1
This Column	52	13	65	4.0
This Stack	38	16	54	2.4
These Blocks	24	20	44	1.2
Deictic Total	**701**	**319**	**1,020**	**2.2**
Social Gestures				
Pos. Acknowledge	693	225	918	3.1
Wait (Pause)	400	246	646	1.6
Start	100	51	151	2.0
Done	81	47	128	1.7
Neg. Acknowledge	109	11	120	9.9
Emphasis	17	27	44	0.6
Social Total	**1,371**	**607**	**1,978**	**2.3**
Numeric Gestures				
One	85	18	103	4.7
Two	64	9	73	7.1
Three	26	4	30	6.5
Four	21	5	26	4.2
Numeric Total	**196**	**36**	**232**	**5.4**

The gesture labels do not match up with the gestures recognized by HAB. There are many more gestures in the human data, and the gesture labels are more semantic. All gestures that are intended to make a partner wait, for example, are grouped into one category. *Servo* gestures are the continual gestures as in “a little more... a little more...”. For a complete explanation of gestures, see [43].

Our first prediction based on HAB was that social gestures should be more common than iconic or deictic gestures. According to Table III this is true, although not by a large margin: 1,371 social gestures versus 1,276 iconic. Social gestures sometimes went unnoticed by the labelers, however, so the true disparity may be larger.

Our second prediction was that iconic gestures should appear more often in the video-only condition than the video+audio condition. It turns out that *all* gestures are more common in the video-only condition, because when people can’t hear each other they gesture more. This is why we included the ratio of frequencies between the two conditions in Table III. While the ratio is positive for all four gesture types, it is higher for iconic (2.5) than for social (2.3). Deictic gestures have the lowest ratio at 2.2. Although not implemented in HAB, numbers have by far the highest ratio at 5.4. Apparently people rarely gesture a number if they can simply say it.

D. Coordinate Systems

To ground deixis, the signaler and builder must agree on a coordinate system. In some dimensions the coordinate system is obvious; up is always against gravity, and down is always with it. The other dimensions aren’t as clear, however.

Our initial mental model was that the signaler and builder were at different ends of a shared table; think of Figure 2 without the gap between tables. In this model, the signaler’s left is the builder’s right, and vice-versa. This seems natural in practice, and re-examining the human subjects data it is indeed the coordinate system that our test subjects used.

This same model would predict that the edge of the table closest to the signaler is farthest from the builder and vice-versa. Thus if a signaler gesturally pushes a block away, the avatar should pull the block toward herself. Conversely, if a signaler pulls a block toward themself, the avatar should push the block away. But somehow, this doesn’t seem as natural in practice.

Re-examining the human subjects data, human builders are inconsistent when a human signaler pushes a block away. About half the time, builders pull the block toward themselves. The other half of the time they push the block away. In the video-only trials, this was a source of confusion. In the video+audio trials, signalers and builders resolved the coordinate system verbally, using phrases like “toward you” or “toward me”. This suggests an interesting interaction between speech and gesture in the context of deixis.

E. Recognition Accuracy

One concern that arose while designing HAB was whether gesture recognition would be accurate enough to support peer-to-peer communication. While there have been many previous 3D gesture-based interfaces, most have been designed to detect large-scale gestures, for example gaming motions or scuba diving signals, as in the ChaLearn challenge [44]. We needed to recognize natural gestures elicited from naive users in the context of blocks world, and were counting on new sensors and recogntion techniques to make this possible.

In particular, we were counting on the Kinect sensor to extract accurate depth maps, and the Kinect pose data (a.k.a. skeleton) to reliably identify the locations of the hands and head. We were also relying on ResNet-style deep convolutional neural networks (DCNNs) [34] to recognize hand poses in segmented depth images, and to recognize head motions given a time window of differences of depth images. A significant risk factor with regard to DCNNs was whether we had enough training data to train reliable networks.

Training samples were extracted from the human subjects studies described in Section III-A, which we hand labeled [43]. For 25 hand poses, this yielded a significant number of training samples. There were other hand poses, such as thumbs down, that appeared less often but that we still wanted to include in the system. We therefore supplemented the training samples for 10 more hand poses by having volunteers perform the poses in front of the Kinect. Unfortunately, the data collected in this way turned out to be exaggerated compared to naturally occurring poses.

When the prototype was completed, signalers reported satisfaction with the gesture recognition. In practice, the system rarely missed gestures or inserted false gestures. There are many possible explanations for this, however, including

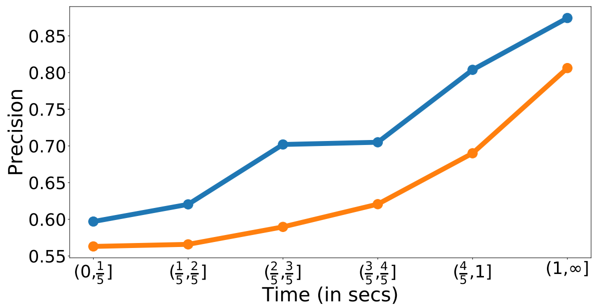


Fig. 5. Precision of hand pose detection. The horizontal axis represents durations of detected hand poses. The vertical axis is the percent of true detections as judged by naive raters. The blue line shows the precision of the 25 hand poses trained on the human subjects data. The orange line adds 10 more poses for which additional, exaggerated training data was collected.

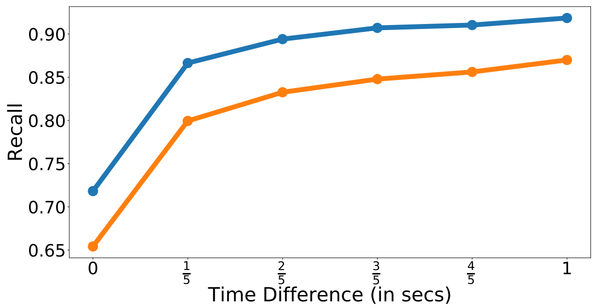


Fig. 6. Recall of hand pose detection. The horizontal axis is the time difference between the frame selected by a naive rater and the automatic detection. The vertical axis is the percent of recall. The blue and orange lines indicate the same distinction between 25 and 35 poses as in Figure 5.

that the signalers were system designers with an interest in gesturing clearly. Furthermore, although the gesture recognition system detects 35 hand poses, only a subset are fused by the finite state machines into the 7 multi-modal gestures integrated with VoxSim. The underlying performance of hand pose recognition was therefore still unknown.

To evaluate the accuracy of hand pose recognition, we collected new data from 14 naive human subjects, using the same experimental setup and protocol as in Section III-A. We didn't have the resources to hand label every frame, so instead we adopted a sampling methodology. The new videos were processed through the DCNN hand pose classifier. To estimate precision, we randomly sampled 60 instances of each hand pose as identified by the DCNN. We then brought in naive raters and asked them whether the detected gestures actually occurred where the DCNN said they did.

The precision results are shown in Figure 5. The units of the horizontal axis are seconds, so that the leftmost data point is the precision for gesture detections that lasted for $\frac{1}{5}$ of a second or less. The second data point represents detections with a duration between $\frac{1}{5}$ and $\frac{2}{5}$ of a second, and so on. The orange line represents the precision across all 35 poses, while the blue line shows the precision for the 25 poses trained on data from the human subjects studies. The plot shows that even for gestures that last for $\frac{1}{5}$ of a second or less, over 55% of the DCNNs detections are correct. Long duration gestures (one second or more) have a precision of 80%. When we limit the evaluation to the 25 naturally trained poses, these numbers go up to 60% for short gestures and 87% for long ones.

Recall was estimated using a similar procedure. In this case, naive raters were given portions of videos, and asked to label any hand poses they saw. For each pose, they selected one frame. (Although not instructed to do so, they usually selected the first frame in which the pose appeared.) We then measured how often the DCNN detected the hand pose at that frame, within a fifth of a second of that frame, within two fifths of a second, and so on, up to a second. The resulting plot is shown in Figure 6, with the same color scheme as in Figure 5 in terms of recall for 25 or 35 poses.

Interestingly, almost half the recall omissions are the result not of mistakes by the DCNN, but of segmentation failures. The depth images passed to the DCNN are windows of the full image centered on the hand, as identified by the Microsoft skeleton. When the hand position is incorrect, failure is inevitable.

V. Conclusion

This paper investigates peer-to-peer communication between people and avatars in the context of a shared perceptual task. In our experiments, people communicate with avatars using gestures and words, and avatars communicate back through gestures, words, and actions. Together, they complete tasks through mixed initiative conversations. The human signaler has the initial goal and tells the avatar what to do, but when ambiguities arise the initiative shifts and the avatar asks the human for clarification. Social cues make this process flow naturally. Overall, we demonstrate an example of peer-to-peer cooperation between people and machines, through shared perception and perceptually-grounded reasoning.

VI. Acknowledgements

This work was supported by the US Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO) under contract #W911NF-15-1-0459 at Colorado State University and the University of Florida and contract #W911NF-15-C-0238 at Brandeis University.

References

[1] D. Küster, E. Krumhuber, and A. Kappas, "Nonverbal behavior online: A focus on interactions with and via artificial agents and avatars," in *The Social Psychology of Nonverbal Communication*. Springer, 2015, pp. 272–302.

[2] J. O. Wobbrock, M. R. Morris, and A. D. Wilson, "User-defined Gestures for Surface Computing," ser. CHI '09. New York, NY, USA: ACM, 2009, pp. 1083–1092. [Online]. Available: http://doi.acm.org/10.1145/1518701.1518866

[3] L. Sproull, M. Subramani, S. Kiesler, J. H. Walker, and K. Waters, "When the interface is a face," *Human-Computer Interaction*, vol. 11, no. 2, pp. 97–124, 1996.

[4] M. Dastani, E. Lorini, J.-J. Meyer, and A. Pankov, "Other-condemning anger = blaming accountable agents for unattainable desires," in *Proceedings of the 16th Conference on Autonomous Agents and Multi-Agent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2017, pp. 1520–1522.

[5] J. Li, "The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents," *International Journal of Human-Computer Studies*, vol. 77, pp. 23–37, 2015.

[6] R. A. Bolt, *"Put-that-there": Voice and gesture at the graphics interface*. ACM, 1980, vol. 14, no. 3.

[7] B. Dumas, D. Lalanne, and S. Oviatt, "Multimodal interfaces: A survey of principles, models and frameworks," *Human machine interaction*, pp. 3–26, 2009.

[8] M. Turk, "Multimodal interaction: A review," *Pattern Recognition Letters*, vol. 36, pp. 189–195, 2014.

[9] F. Quek, D. McNeill, R. Bryll, S. Duncan, X.-F. Ma, C. Kirbas, K. E. McCullough, and R. Ansari, "Multimodal human discourse: gesture and speech," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 9, no. 3, pp. 171–193, 2002.

[10] H. H. Clark and S. E. Brennan, "Grounding in communication," in *Perspectives on Socially Shared Cognition*, L. Resnick, L. B., M. John, S. Teasley, and D, Eds. American Psychological Association, 1991, pp. 13–1991.

[11] H. H. Clark and D. Wilkes-Gibbs, "Referring as a collaborative process," *Cognition*, vol. 22, no. 1, pp. 1–39, Feb. 1986. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0010027786900107

[12] P. Dillenbourg and D. Traum, "Sharing solutions: Persistence and grounding in multimodal collaborative problem solving," *The Journal of the Learning Sciences*, vol. 15, no. 1, pp. 121–151, 2006.

[13] S. R. Fussell, R. E. Kraut, and J. Siegel, "Coordination of Communication: Effects of Shared Visual Context on Collaborative Work," in *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, ser. CSCW '00. New York, NY, USA: ACM, 2000, pp. 21–30. [Online]. Available: http://doi.acm.org/10.1145/358916.358947

[14] S. R. Fussell, L. D. Setlock, J. Yang, J. Ou, E. Mauer, and A. D. I. Kramer, "Gestures over Video Streams to Support Remote Collaboration on Physical Tasks," *Hum.-Comput. Interact.*, vol. 19, no. 3, pp. 273–309, Sep. 2004.

[15] R. E. Kraut, S. R. Fussell, and J. Siegel, "Visual Information As a Conversational Resource in Collaborative Physical Tasks," *Hum.-Comput. Interact.*, vol. 18, no. 1, pp. 13–49, Jun. 2003.

[16] D. Gergle, R. E. Kraut, and S. R. Fussell, "Action As Language in a Shared Visual Space," in *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, ser. CSCW '04. New York, NY, USA: ACM, 2004, pp. 487–496. [Online]. Available: http://doi.acm.org/10.1145/1031607.1031687

[17] L. M. Reeves, J. Lai, J. A. Larson, S. Oviatt, T. Balaji, S. Buisine, P. Collings, P. Cohen, B. Kraal, J.-C. Martin *et al.*, "Guidelines for multimodal user interface design," *Communications of the ACM*, vol. 47, no. 1, pp. 57–59, 2004.

[18] E. S. Veinott, J. Olson, G. M. Olson, and X. Fu, "Video Helps Remote Work: Speakers Who Need to Negotiate Common Ground Benefit from Seeing Each Other," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '99. New York, NY, USA: ACM, 1999, pp. 302–309. [Online]. Available: http://doi.acm.org/10.1145/302979.303067

[19] A. Lascarides and M. Stone, "Formal semantics for iconic gesture," in *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue (BRANDIAL)*, 2006, pp. 64–71.

[20] A. S. Clair, R. Mead, M. J. Matarić *et al.*, "Monitoring and guiding user attention and intention in human-robot interaction," in *ICRA-ICAIR Workshop, Anchorage, AK, USA*, vol. 1025, 2010.

[21] C. Matuszek, L. Bo, L. Zettlemoyer, and D. Fox, "Learning from unscripted deictic gesture and language for human-robot interactions." in *AAAI*, 2014, pp. 2556–2563.

[22] N. Krishnaswamy and J. Pustejovsky, "Multimodal semantic simulations of linguistically underspecified motion events," in *Spatial Cognition X: International Conference on Spatial Cognition*. Springer, 2016.

[23] M. Gilbert, *On social facts*. Princeton University Press, 1992.

[24] R. Stalnaker, "Common ground," *Linguistics and philosophy*, vol. 25, no. 5, pp. 701–721, 2002.

[25] N. Asher and A. Gillies, "Common ground, corrections, and coordination," *Argumentation*, vol. 17, no. 4, pp. 481–512, 2003.

[26] M. Tomasello and M. Carpenter, "Shared intentionality," *Developmental science*, vol. 10, no. 1, pp. 121–125, 2007.

[27] B. K. Bergen, *Louder than words: The new science of how the mind makes meaning*. Basic Books (AZ), 2012.

[28] K.-y. Hsiao, S. Tellex, S. Vosoughi, R. Kubat, and D. Roy, "Object schemas for grounding language in a responsive robot," *Connection Science*, vol. 20, no. 4, pp. 253–276, 2008.

[29] J. Dzifcak, M. Scheutz, C. Baral, and P. Schermerhorn, "What to do and how to do it: Translating natural language directives into temporal and dynamic logic representation for goal management and action execution," in *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*. IEEE, 2009, pp. 4163–4168.

[30] A. Cangelosi, "Grounding language in action and perception: from cognitive agents to humanoid robots," *Physics of life reviews*, vol. 7, no. 2, pp. 139–151, 2010.

[31] J. M. Siskind, "Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic," *J. Artif. Intell. Res.(JAIR)*, vol. 15, pp. 31–90, 2001.

[32] A. Kendon, *Gesture: Visible Action as Utterance*. Cambridge University Press, 2004.

[33] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE MulitMedia*, vol. 19, pp. 4–10, 2012.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[35] N. Krishnaswamy and J. Pustejovsky, "VoxSim: A visual platform for modeling motion language," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. ACL, 2016.

[36] W. Goldstone, *Unity Game Development Essentials*. Packt Publishing Ltd, 2009.

[37] J. Pustejovsky and J. Moszkowicz, "The qualitative spatial dynamics of motion," *The Journal of Spatial Cognition and Computation*, 2011.

[38] J. Pustejovsky and N. Krishnaswamy, "VoxML: A visualization modeling language," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, N. C. C. Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Paris, France: European Language Resources Association (ELRA), May 2016.

[39] J. Pustejovsky, "Dynamic event structure and habitat theory," in *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon (GL2013)*. ACL, 2013, pp. 1–10.

[40] D. McDonald and J. Pustejovsky, "On the representation of inferences and their lexicalization," in *Advances in Cognitive Systems*, vol. 3, 2014.

[41] J. Pustejovsky, "The generative lexicon," 1995.

[42] G. Hirst, S. McRoy, P. Heeman, P. Edmonds, and D. Horton, "Repairing conversational misunderstandings and non-understandings," *Speech Communication*, vol. 15, no. 3, pp. 213–229, Dec. 1994. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0167639394900736

[43] I. Wang, P. Narayana, D. Patil, G. Mulay, R. Bangar, B. Draper, R. Beveridge, and J. Ruiz, "Eggnog: A continuous, multi-modal data set of naturally occurring gestures with ground truth labels," in *Io appear in the Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition*, 2017.

[44] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera, "Chalearn lap 2016: First round challenge on first impressions - dataset and results," in *ECCV*, 2016, pp. 400–418.