Embodied Human-Computer Interactions through Situated Grounding

James Pustejovsky**
Brandeis University
Waltham, MA
jamesp@brandeis.edu

Nikhil Krishnaswamy
Colorado State University
Fort Collins, CO
nkrisha87@gmail.com

ABSTRACT

In this paper, we introduce a simulation platform for modeling and building *Embodied Human-Computer Interactions (EHCI)*. This system, VoxWorld, is a multimodal dialogue system enabling communication through language, gesture, action, facial expressions, and gaze tracking, in the context of task-oriented interactions. A multimodal simulation is an embodied 3D virtual realization of both the situational environment and the co-situated agents, as well as the most salient content denoted by communicative acts in a discourse. It is built on the modeling language VoxML [7], which encodes objects with rich semantic typing and action affordances, and actions themselves as multimodal programs, enabling contextually salient inferences and decisions in the environment. VoxWorld enables an embodied HCI by situating both human and computational agents within the same virtual simulation environment, where they share perceptual and epistemic common ground.

CCS CONCEPTS

• **Human-centered computing; HCI**;

KEYWORDS

multimodal embodiment, simulation, virtual agent, situated grounding

ACM Reference Format:
James Pustejovsky and Nikhil Krishnaswamy. 2020. Embodied Human-Computer Interactions through Situated Grounding. In *IVA '20: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20), October 19–23, 2020, Virtual Event, Scotland Uk.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3383652.3423910

1 INTRODUCTION

When humans engage in task-oriented dialogues, the conversation is an *embodied interaction* between agents, where language, gesture, gaze, and actions are situated within a common ground shared by all agents in the communication. As the demand for more sophisticated human-computer interactions grows, recent work in human-robot interactions (HRI) and communication has moved towards more serious dialogue modeling to facilitate deeper language understanding in context. In order to achieve these goals, human-computer/robot interactions requires robust recognition and generation of expressions through multiple modalities (language, gesture, vision, action); and the encoding of SITUATED MEANING: this entails three aspects of common ground interpretation: (a) the situated *grounding* of expressions in context; (b) an interpretation of the expression contextualized to the *dynamics* of the discourse; and (c) an appreciation of the *actions and consequences* associated with objects in the environment.

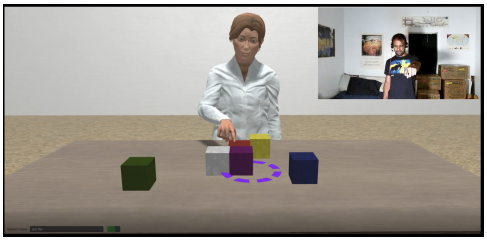

Figure 1: Diana, an Embodied Virtual Agent, engaging in an embodied HCI with a human user.

With this in mind, many HCI researchers have adopted the notion of "embodiment" in order to better understand user expectations when interacting with computational agents. Embodied agents or avatars add new dimensions to human/agent interactions compared to voice- or text-only conversational agents. Embodied agents can express emotions and perform gestures, two crucial non-verbal modes of human communication. Potentially, this enables such agents to have more human-like, peer-to-peer interactions with users. Unfortunately, embodiment alone does not avoid some of the key limitations of conversational agents. Even embedded in an avatar, most agents won't know what you are pointing at. As with verbal conversations, visual communication mechanisms like gestures, expressions, and body language need to be two-way.

2 VOXWORLD: AN EMBODIED INTERACTION PLATFORM

VoxWorld is an environment that supports embodied HCI with embodied virtual agents, who are aware not only of their own virtual space but of the physical space around them. One such avatar, Diana, can speak, gesture, track, move, and emote [2, 4]. Diana has video and depth sensors that let her sense the physical world around her, including the user. Diana observes the user, and knows when they are attending to her, as opposed to doing something else. She can observe the user's emotions, and most importantly she can understand the user's gestures. As a result, visual communication

*Both authors contributed equally to this research.

IVA '20, October 19–23, 2020, Virtual Event, Scotland Uk

ACM ISBN 978-1-4503-7586-3/20/09.
https://doi.org/10.1145/3383652.3423910

joins verbal communication as a two-way process. The current architecture of the VoxWorld system is shown in Figure 2.

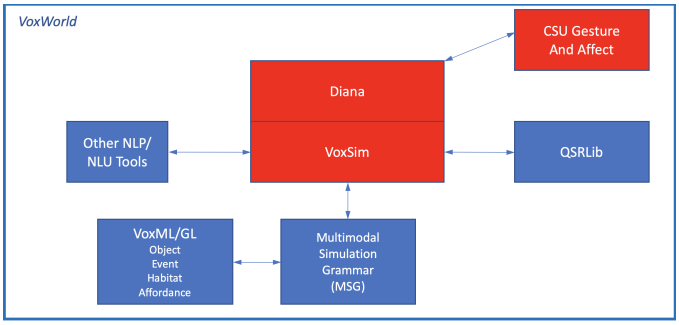

Figure 2: VoxWorld Architecture schematic

Diana herself is embedded in the virtual world of VoxWorld that the user can also see. Shared perception is a critical component of human communication. When people work together on a physical task, they can each see what the others are doing and do not have to describe all their actions. Similarly, when Diana moves a (virtual) block, she does not have to tell the user, the user can see it. This simplifies communication and makes it more natural. It also enables visually grounded reasoning, where the feasibility of actions is determined by the visualization/simulation of the action in the 3D environment perceived by both the person and the agent.

Diana is therefore more than an embodied conversational agent. She combines embodiment (i.e., an avatar) with visual perception to create a two-way conversational and visual agent. By being situated in a displayed visual world, she and the user also share perception. The combination results in an interface that feels qualitatively new. Even though the user knows that Diana is an artificial agent—and her avatar need not be particularly life-like—she has enough capabilities to establish the conceit that the user is interacting with a peer.

At the center of VoxWorld is the language VoxML [7] and the associated software, VoxSim [3]. VoxML (Visual Object Concept Modeling Language) is a modeling language for constructing 3D visualizations of concepts denoted by natural language expressions, and is used in the VoxWorld platform for creating multimodal semantic simulations in the context of human-computer and human-robot communication. VoxSim is the the software that handles visual event simulation in three dimensions, written with the Unity game engine.

VoxSim connects to a number of other default VoxWorld components, including some native natural language processing capabilities, VoxML encodings/GL knowledge as interpreted through the multimodal semantics [6], and 3rd-party libraries, e.g., QSRLib [1]. The interactive avatar Diana is an output interface that can also connect to 3rd-party endpoints; in the case of Diana, this is custom gesture and affect recognition [5].

The current implementation of VoxSim provides scenes in a Blocks World domain, augmented with a set of more complicated or interesting everyday objects (e.g., cups, plates, books, etc.). There are scenes without an avatar where the user can direct the computer to manipulate objects in space or with an avatar that can act upon objects and respond to the user's input. VoxWorld contains other software, models, and interfaces, e.g., to consume input from CNN-based gesture recognizers [5], and to track and update the agent's epistemic state or knowledge about what the human interlocutor knows.

Situational embodiment takes place in real time, so in the case of a situation where there may be too many variables to predict the state of the world at time t from a set of initial conditions at time 0, situational embodiment within the simulation allows the agent to reason forward about a specific subset of consequences of actions that may be taken at time t, given the agent's current conditions and surroundings. Situatedness and embodiment is required to arrive at a complete, tractable interpretation given any element of non-determinism. For example, an agent trying to navigate a maze from start to finish could easily do so with a map that provides complete, or at least sufficient, information about the scenario. If, however, the scene includes a disruptor (e.g., the floor crumbles, or doors open and shut randomly), the agent would be unable to plot a course to the goal. It would have to start moving, assess the current circumstances at every timestep, and choose the next move or next set of n moves based on them. Situated embodiment allows the agent to assess the next move based on the current set of relations between itself and the environment (e.g., ability to move forward but not leftward at the current state). This provides for reasoning that not only saves computational resources but performs more analogously to human reasoning than non-situated, non-embodied methods.

Figure 3 illustrates an embodied HCI, where deixis (pointing) and action-affordance gestures from the human are situated in an embodied space shared by both the IVA and the human. These are accompanied by aligned co-gestural language expressions, such as "that one", "the purple one", etc.

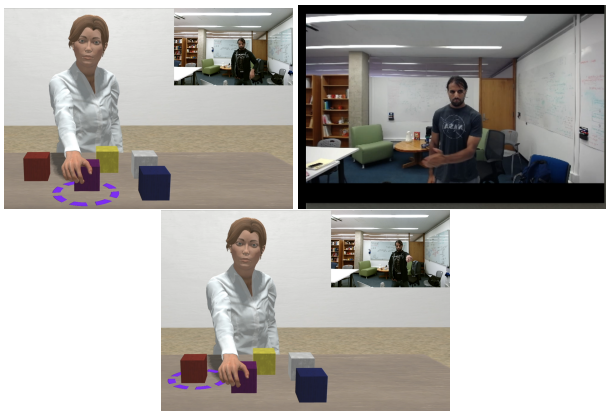

Figure 3: Embodied Interaction with Language and Gesture

Because the avatar Diana is embodied and situated within an embodied HCI environment, we have recently been able to apply transfer learning of object affordances between objects, as illustrated in Figure 4. For this configuration, we assume that Diana has no semantics for the object we recognize as a bottle. In embodied interaction with the human, Diana is able to observe certain similarities in the shape and habitats of the cup and the bottle (e.g., current upright orientation, similar symmetry and size constraints), and infer that they might share some behaviors, which leads her

to infer that a way to grasp the bottle would be like she grasps the cup. The close association between habitats and affordances and the structured encoding provided by VoxML allows us to perform this kind of transfer learning using methods similar to Skip-Gram word embedding models in natural language processing, by inferring a likely missing behavior given the current combination of circumstances.

Figure 4: Diana interacts with an unknown object through its recognized affordances

3 EMBODIED HCI AND ROBOT CONTROL

We are exploring an additional application of embodied HCI in the context of communication and control of a mobile robot. Specifically, we have adapted the VoxWorld interface for navigation in novel environments using coordinated gesture and language. We use a TurtleBot3 robot with a LIDAR and a camera, an embodied simulation of what the robot has encountered while exploring, and a cross-platform bridge facilitating generic communication. A human partner can deliver instructions to the robot using spoken English and gestures relative to the simulated environment, to guide the robot through navigation tasks.

In our system, a human user and a robot exist in a co-situated space that is mediated by a virtual environment displayed on a screen, such that the human can see a virtual rendition of the environment the robot has explored, and of the robot's current perspective view. The human can then gesture to objects and locations on the screen, either in a perspective or omniscient view, and speak about them in English, e.g., "*go there*," "*go to that wastebasket and turn around*," or "*find the blue block*." Deictic gestures are grounded to coordinates on the screen which are transformed to equivalent coordinates in the robot's ROS environment, allowing the robot to execute native navigation commands, e.g, $go\_to(x, y)$. The robot can likewise communicate status updates back to the human which are then spoken out through text-to-speech.

Figure 5 shows the virtual rendition of the robot's environment from its perspective (main panel), an omniscient view (top left), and the visualized LIDAR data (bottom left).

The purple circle shows where the user (top right panel) is pointing, as in the Diana implementation. In the top of the figure, the robot might here the instruction "go here/to that one," be able to see which object the user is indicating, and go to it. In another scenario, imaging the user is pointing to the pink block as shown in the bottom view, and gives the same instruction. There, the denotation of "here" or "that one" is not available to the robot in the common ground, because the demonstrative has not been grounded to a location. The robot will have to ask for clarification ("I can't see where you're pointing") or turn around to scan until it see the location of deixis in order to interpret the instruction.

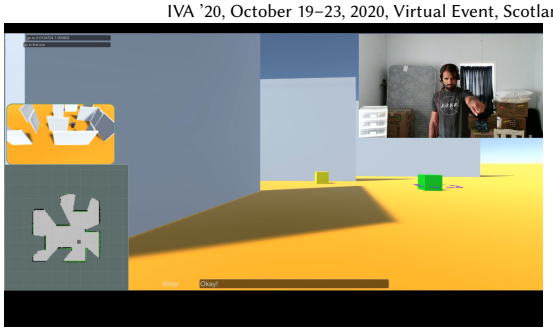

Figure 5: Communicating with Mobile Robot

4 CONCLUSION

We believe that VoxWorld will facilitate experimentation with IVAs in embodied HCI contexts, using multiple modalities in diverse settings. An embodied HCI, such as that enabled by the simulation environment VoxWorld, provides a venue for the human and computer or robot to share an epistemic space, and any communicative modality that can be expressed within that space (e.g., linguistic, visual, gestural) enriches the number of ways that a human and a computer or robot can communicate regarding objects, actions, and situation-based tasks.

5 ACKNOWLEDGMENTS

We would like to thank the reviewers for their helpful comments. We would also like to thank Ross Beveridge, Bruce Draper, Francisco Ortega, and their team at CSU working with Brandeis, without whose contribution the Diana System would not be a reality. This work was supported by the US Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO) under contract #W911NF-15-C-0238 at Brandeis University. The points of view expressed herein are solely those of the authors and do not represent the views of the Department of Defense or the United States Government. Any errors or omissions are, of course, the responsibility of the authors. All errors and mistakes are, of course, the responsibilities of the authors.

REFERENCES

[1] Yiannis Gatsoulis, Muhannad Alomari, Chris Burbridge, Christian Dondrup, Paul Duckworth, Peter Lightbody, Marc Hanheide, Nick Hawes, DC Hogg, AG Cohn, et al. 2016. QSRlib: a software library for online acquisition of Qualitative Spatial Relations from Video. (2016).

[2] Nikhil Krishnaswamy, Scott Friedman, and James Pustejovsky. 2019. Combining Deep Learning and Qualitative Spatial Reasoning to Learn Complex Structures from Sparse Examples with Noise. In *33rd AAAI Conference on Artificial Intelligence*. AAAI.

[3] Nikhil Krishnaswamy and James Pustejovsky. 2016. VoxSim: A Visual Platform for Modeling Motion Language. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. ACL.

[4] David G McNeely-White, Francisco R Ortega, J Ross Beveridge, Bruce A Draper, Rahul Bangar, Dhruva Patil, James Pustejovsky, Nikhil Krishnaswamy, Kyeongmin Rim, Jaime Ruiz, and Isaac Wang. 2019. User-Aware Shared Perception for Embodied Agents. In *2019 IEEE International Conference on Humanized Computing and Communication (HCC)*. IEEE, 46–51.

[5] Pradyumna Narayana, Nikhil Krishnaswamy, Isaac Wang, Rahul Bangar, Dhruva Patil, Gururaj Mulay, Kyeongmin Rim, Ross Beveridge, Jaime Ruiz, James Pustejovsky, and Bruce Draper. 2018. Cooperating with Avatars Through Gesture, Language and Action. In *Intelligent Systems Conference (IntelliSys)*.

[6] James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.

[7] James Pustejovsky and Nikhil Krishnaswamy. 2016. VoxML: A Visualization Modeling Language. *Proceedings of LREC* (2016).