International Conference on Artificial Reality and Telexistence
Eurographics Symposium on Virtual Environments (2020)
K. Kim and B.-K. Seo (Editors)

Situational Awareness in Human Computer Interaction: Diana's World

Nikhil Krishnaswamy[1,2†], Ross. Beveridge[1], James Pustejovsky[2], Dhruva Patil[1], David G. McNeely-White[1], Heting Wang[1,3‡], Francisco R. Ortega[1]

[1]Department of Computer Science, Colorado State University, Fort Collins, CO, USA
[2]Department of Computer Science, Brandeis University, Waltham, MA, USA
[3]Department of Computer Science, University of Florida, Gainesville, FL, USA

Abstract

In this paper, we illustrate the role that situated awareness plays in modeling human interactions with Intelligent Virtual Agents (IVAs). Here we describe Diana, a multimodal IVA who exists within an embodied Human-Computer Interaction (EHCI) environment. Diana is a multimodal dialogue agent enabling communication through language, gesture, action, facial expressions, and gaze tracking, in the context of task-oriented interactions.

CCS Concepts
• ***Human-centered computing*** → *Natural language interfaces; HCI;*

1. Introduction

As interactive agents become more sophisticated, user expectations of these systems' capabilities grow similarly. Many commercial dialogue agents, such as Siri and Alexa, contain advanced natural language understanding functionality but cannot recreate the same seamless conversation and interaction that takes place between two humans. Some key missing components in these unimodal interactive agents are embodiment and situational awareness. They may be able to hear and interpret language, but lack the machinery to see and understand the surrounding environment and discuss and engage with it accordingly. Users want agents to engage in the same way that people can, but unimodal agents without situational awareness are incapable of answering such basic (to a human) questions as "what am I pointing at?" and agents without embodiment cannot respond symmetrically.

We present and demonstrate Diana, an interactive virtual agent (IVA) who facilitates embodied human-computer interaction (EHCI) through an articulated avatar, vision sensors, gesture recognition, and language interpretation capability [PK20]. Diana is aware of her environment and co-agent, and can interpret multi-channel inputs—including language, gesture, affect, and emotion—in real time, and has enough capabilities to establish the conceit that the user is interacting with a peer. Diana is a multimodal system designed for human-avatar collaboration, another unique aspect of our system.

2. Embodied Human-Computer Interaction

Diana's environment is built on the VoxSim platform [KP16] and the Unity game engine (Fig. 1). Here, Diana has access to the virtual objects in her world, which she can manipulate (including grasping, lifting, sliding, moving, etc.). Attached to the computer is a Kinect® RGBD camera and a microphone.

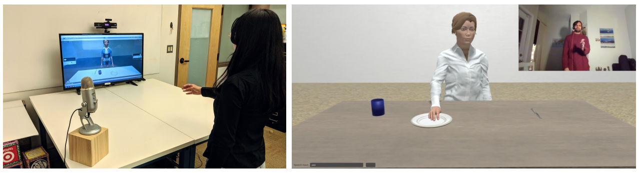

Figure 1: *Diana's interactive setup within the real world (L) and Diana's environment (human inset in upper right) (R)*

Custom gesture recognition algorithms process the Kinect® stream into 34 individual gestures which the human uses to indicate objects and actions to execute over them. Automatic speech recognition software allows Diana to consume full or partial sentences and mix information from the speech modality with information from the gestural modality. For instance, a human may 1) ask Diana to "put the knife next to the cup"; or 2) point to a red block, and say "(put it) on the green block"; or 3) say "the plate," make a "claw" gesture representing "grab it," and then guide Diana in the manner of how to grasp a plate appropriately.

Diana will ask questions or offer suggestions if she needs clarification. For instance, if the human points to a spot containing multiple objects, she will choose one but can be corrected, e.g., if she

[†] Work performed at Brandeis University and Colorado State University.
[‡] Work performed at Colorado State University.

selects the white block, “Wait! The yellow one," will cause her to select the yellow one instead. If the object is already part of an ongoing action, she can abort the action being taken with the white and apply it to the yellow one instead. If a specified action has multiple possible results, she may ask the user to choose, e.g., “Should I grasp [the cup] like this?" while demonstrating a possible action.

Diana is asynchronous, interruptable, and correctable, and can move the conversation forward similarly to a human, allowing the human user to instruct her in building configurations (e.g., staircase, pyramid, tower, place setting, etc.). Diana is compatible with multiple speech recognition platforms; the demo uses Google’s ASR platform. VoxSim is also compatible with certain virtual reality (VR) headsets such as the HTC VIVE$^{\text{TM}}$, so Diana can be deployed in VR applications for a more immersive experience.

3. Technical Functionality

Diana has evolved over time. Previous versions of Diana could interpret gestural instructions only [KNW*17], and later incorporated word-spotting [NKW*18]. Diana’s gesture vocabulary is derived from human-to-human elicitation studies conducted to understand gestures used by humans in the course of a collaborative task [WNP*17].[†] FOANet [NBD19], a ResNet-style deep convolutional neural net, allows Diana to recognize gestures that humans use to mean *grasp*, *lift*, *move*, *push*, as well as *yes*, *no*, *stop*, and more. She also recognizes iterative motion gestures, called “servo."

The VoxSim platform is built on top of the VoxML modeling language [PK16], which encodes the semantics of objects in Diana’s environment, including their habitats, or contextualized placement in the environment, and affordances, or typical use or purpose. Diana can therefore learn novel gestures not in the default set and reason about how to interact with novel objects by analogizing them to known ones. In Fig. 2, the cup (L) must be grasped differently from what the standard “claw down" grasp gesture implies, and by observing similarities in the structure, habitats, and affordances of a cup and the new object (a bottle), reasons that she should grasp the bottle similarly (R).

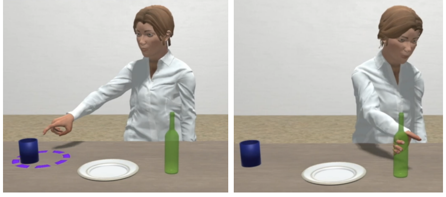

Figure 2: *Inferring grasping from object similarities*

4. Asynchrony and Affect

Human communication is asynchronous, as we attend to our interlocutor while continuing to speak and act in the interaction. Multimodal interactive agents should be the same. Diana is proactive, responsive, and interruptible with new information while attending to and acting upon the human’s multimodal cues. For instance, if Diana misinterprets the destination at which the human wants her to place an object, the human may interrupt and correct her with a statement like “wait—on the yellow block," perhaps with an accompanying stop gesture. Diana will then “rewind" the continuation, and reapply the new destination to the current action.

Diana can also take the human user’s emotional feedback into account. Diana’s emotional states include joy, sympathy, confusion, and concentration—key affective states in collaborative tasks. In particular, concentration was implemented to mimic the behavior of humans when solving tasks. If the user expresses anger or frustration, it should signal to Diana that the user is displeased with something she has done, and she should act accordingly. These decisions based on emotional feedback are situation-dependent, based on actions Diana or the user have taken preceding the affective cue.

5. Conclusion

Diana allows her interlocutor to accept the conceit that they are interacting with another human, achieving this through her embodiment, multimodal communication, and situational awareness. We believe the future of intelligent agents lies in situated communicative acts within embodied human-computer interaction. As intelligent agents become more widespread, it is crucial they be able to understand the environment they share with humans, and interact and communicate with them symmetrically. This demonstration of Diana showcases the methods developed toward that end.

References

[KNW*17] 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). 2

[KP16] 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* (2016), ACL. 1

[NBD19] Narayana P., Beveridge J. R., Draper B. A.: Continuous gesture recognition through selective temporal fusion. In *2019 International Joint Conference on Neural Networks (IJCNN)* (July 2019), pp. 1–8. doi:10.1109/IJCNN.2019.8852385. 2

[NKW*18] 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). 2

[PK16] Pustejovsky J., Krishnaswamy N.: Voxml: A visualization modeling language. *Proceedings of LREC* (2016). 2

[PK20] Pustejovsky J., Krishnaswamy N.: Embodied human-computer interactions through situated grounding. In *Proceedings of the 20th ACM International Conf. on Intelligent Virtual Agents* (2020). 1

[WNP*17] 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 *Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition* (2017). 2

[†] This dataset, EGGNOG, is available at https://www.cs.colostate.edu/~vision/eggnog/.