Diana's World: A Situated Multimodal Interactive Agent

Nikhil Krishnaswamy,[1] Pradyumna Narayana,[2] Rahul Bangar,[2] Kyeongmin Rim,[1] Dhruva Patil,[2] David McNeely-White,[2] Jaime Ruiz,[3] Bruce Draper,[4] Ross Beveridge,[2] James Pustejovsky[1]
[1]Brandeis University Department of Computer Science, Waltham, MA, USA
[2]Colorado State University Department of Computer Science, Fort Collins, CO, USA
[3]University of Florida Department of Computer & Information Science & Engineering, Gainesville, FL, USA
[4]DARPA Information Innovation Office, Arlington, VA, USA*
{nkrishna, jamesp}@brandeis.edu; ross.beveridge@colostate.edu; jaime.ruiz@ufl.edu; bruce.draper@darpa.mil

Abstract

State of the art unimodal dialogue agents lack some core aspects of peer-to-peer communication—the nonverbal and visual cues that are a fundamental aspect of human interaction. To facilitate true peer-to-peer communication with a computer, we present Diana, a situated multimodal agent who exists in a mixed-reality environment with a human interlocutor, is situation- and context-aware, and responds to the human's language, gesture, and affect to complete collaborative tasks.

Introduction

Sophisticated language models trained over large amounts of data have allowed dialogue agents to have convincing conversations using spoken or written language. But even a state-of-the-art unimodal system will be unable to answer such basic questions as "What am I pointing at?" Unimodal language agents are not *environmentally aware* of their interlocutor's embodiment or of objects and actions in the situation. They are not *co-situated* with their interlocutors and lack machinery to integrate a visual stream of the environment with the interlocutor's language. Here, we present and demonstrate Diana, a multimodal interactive agent with awareness of her environment and coagent, who can interpret multi-channel inputs—including language, gesture, affect, and emotion—in real time, and plays a proactive role in collaborative interactions with a human.

Situated Multimodal Interaction

Diana exists in a virtual world built on the Unity game engine (Fig. 1), where she can manipulate virtual objects by grasping, lifting, moving, and sliding them. This world is displayed by a computer attached to a Kinect® RGB+depth camera and a microphone, through which she consumes inputs from the human user (Fig. 2).

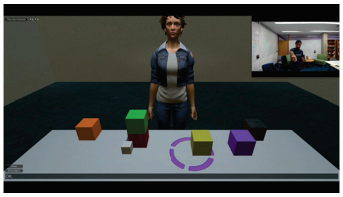

Figure 1: Diana's world (human inset in upper right)

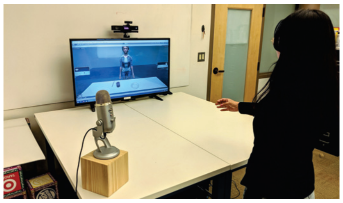

Figure 2: Diana's world within the real world

Diana understands up to 34 individual gestures which the human can use to both indicate specific objects as well as what to do with them. She also understands spoken language in the form of full or partial sentences, allowing these modalities to be mixed and matched in real time in the manner of two humans interacting. For instance, a human may 1) ask Diana to "put the red block on the green block"; or 2) point to the red block, and say "(put it) on the green block"; or 3) say "the red block," make a "claw" gesture representing "grab it," and then point to the green block. All of these and other mixes of modality and order prompt the same action.

Diana will ask questions or offer suggestions if she needs clarification. For instance, if the human points to a spot containing multiple objects, Diana may suggest one, e.g., "Do you mean the purple block?", or if a specified action has multiple possible results, may ask the user to choose, e.g., "Should I push the orange block left of the black block?"

Through accumulating and integrating partial information, prompting, and clarifying, Diana moves the conversation forward similarly to a human, allowing the human user to instruct her in building structures (e.g., staircase, pyramid, tower, etc.) or executing routines, i.e., laying a place setting.

Interpretive Machinery

Diana in her current state is an updated version of a previous system that could interpret coarse, purely gestural instructions (Krishnaswamy et al. 2017), and later simple natural

*Work performed at Colorado State University.

language (i.e., word-spotting) (Narayana et al. 2018).

The gestures Diana understands were gathered from human-to-human elicitation studies conducted to better understand communicative gestures used by humans in the course of a collaborative task (Wang et al. 2017). Diana recognizes gestures using ResNet-style deep convolutional neural nets and classifies them by aggregating such information as hand position, arm motion, and body pose. She can recognize gestures that humans use to mean *grasp*, *lift*, *move*, *push*, as well as *yes*, *no*, *stop*, and more. She also recognizes iterative versions of motion gestures, that we call "servo."

We take a modular approach to speech recognition and parsing, allowing interfaces with a number of APIs, including IBM Watson or custom Kaldi models, and parsers like Stanford Dependency or spaCy. Text-to-speech is facilitated through a Unity API to the system TTS.

Diana uses a continuation-passing style semantics to interpret and aggregate inputs from multiple modalities and a blackboard architecture to pass inputs from module to module, such as generated responses to the the TTS, or gestural and language inputs to the continuation-style interpreter.

When human users witness an avatar speaking fluently, they frequently assume that she can understand equally fluent language, along with the knowledge of entities being discussed. To semantically reason over both the objects she manipulates and the actions she takes, Diana has been implemented on top of the VoxSim platform (Krishnaswamy and Pustejovsky 2016), a dynamic semantic reasoning engine built on the VoxML modeling language (Pustejovsky and Krishnaswamy 2016), which provides the semantics of the entities in Diana's world, including their *habitats*, or contextualized placement in embedding spaces, and *affordances*, or typical use or purpose. This means that Diana can *learn* novel gestures not in the default set. For instance in Fig. 3, the standard "claw down" can be used to signal the standard grasp (L), and a gesture as if miming grasping a cup can be taught to her to signal her to grasp the cup in a way more appropriate to the object geometry and affordances (R).

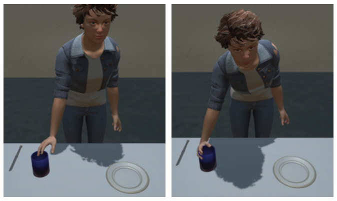

Figure 3: Alternate affordances for "grasp the cup"

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.

Another extension to Diana currently under development is the ability to take the human user's emotional feedback into account. Using the Affectiva emotional measurement technology, we are making Diana recognize facial expression and take emotional state into account. For instance, 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. This might include undoing a recent action, or using a different strategy for disambiguation or clarification. These decisions based on emotional feedback are situation-dependent, based on actions Diana or the user have taken that preceded the affective cue.

Conclusion

Diana is an interactive agent that, like a human, is embodied, communicates multimodally, is capable of multichannel interpretation in real time, and is situationally context-aware. We believe the future of intelligent agents lies in situated communicative acts within a *common ground* that facilitates peer-to-peer communication. As intelligent agents become more widespread and integrated with everyday life, it is crucial that they be able to understand the environment they share with the humans they interact with. We present Diana, a *d*ynamic, *i*nteractive, *a*sy*n*chronous *a*gent, to showcase methods we provide of doing just that.

References

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

Krishnaswamy, N.; Narayana, P.; Wang, I.; Rim, K.; Bangar, R.; Patil, D.; Mulay, G.; Ruiz, J.; Beveridge, R.; Draper, B.; and Pustejovsky, J. 2017. Communicating and acting: Understanding gesture in simulation semantics. In *12th International Workshop on Computational Semantics*.

Narayana, P.; Krishnaswamy, N.; Wang, I.; Bangar, R.; Patil, D.; Mulay, G.; Rim, K.; Beveridge, R.; Ruiz, J.; Pustejovsky, J.; and Draper, B. 2018. Cooperating with avatars through gesture, language and action. In *Intelligent Systems Conference (IntelliSys)*.

Pustejovsky, J., and Krishnaswamy, N. 2016. 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.; and Piperidis, S., eds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris, France: European Language Resources Association (ELRA).

Wang, I.; Narayana, P.; Patil, D.; Mulay, G.; Bangar, R.; Draper, B.; Beveridge, R.; and Ruiz, J. 2017. 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*.