User-Aware Shared Perception for Embodied Agents

David G. McNeely-White
Francisco R. Ortega
J. Ross Beveridge
Bruce A. Draper
Rahul Bangar
Dhruva Patil
Department of Computer Science
Colorado State University
Fort Collins, Colorado 80523
Email: Ross.Beveridge@colostate.edu

James Pustejovsky
Nikhil Krishnaswamy
Kyeongmin Rim
Department of Computer Science
Brandeis University
Waltham, Massachusetts 02453
Email: jamesp@cs.brandeis.edu

Jaime Ruiz
Isaac Wang
Department of Computer & Information
Science & Engineering
University of Florida
Gainesville, FL 32611
Email: jaime.ruiz@ufl.edu

***Abstract*—We present Diana, an embodied agent who is aware of her own virtual space and the physical space around her. Using video and depth sensors, Diana attends to the user's gestures, body language, gaze and (soon) facial expressions as well as their words. Diana also gestures and emotes in addition to speaking, and exists in a 3D virtual world that the user can see. This produces symmetric and shared perception, in the sense that Diana can see the user, the user can see Diana, and both can see the virtual world. The result is an embodied agent that begins to develop the conceit that the user is interacting with a peer rather than a program.**

I. Introduction

Complex computer systems are becoming a more prominent fixture in our everyday lives. To make using these systems easier, agents have begun to emerge that give complex systems more human-like interfaces. Apple's Siri and Amazon's Alexa are commercial examples of agents that help users interface to complex systems, including but not limited to the internet. Siri and Alexa are examples of conversational agents (CAs) that users can talk to and hear. These systems, while popular, often leave users frustrated and disappointed in the agent's capabilities [1]. In essence, users want them to be able to do what people can. As a simple example of where conversational agents fail, ask Siri or Alexa "what am I pointing at?"

With this in mind, many researchers have turned to embodiment in order to better meet user expectations[1]. Embodied conversational agents (ECAs) 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 ECAs 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. Like verbal conversations, visual communication mechanisms like gestures, expressions, and body language need to be two-way.

This paper presents Diana, an embodied agent (EA) who is aware not only of her own virtual space but of the physical space around her. As an avatar, Diana can speak, gesture, and emote. More importantly, however, Diana has inexpensive 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 joins verbal communication as a two-way process.

Diana herself is embedded in a virtual world that the user can also see. In the prototype described below, the objects are simple things like blocks and cups. What is important is that both Diana and the user can see them. Shared perception is a critical component of human communication. When people work together on a physical task, they can see what each other are doing and don't have to describe all their actions. Similarly, when Diana moves a (virtual) block, she doesn't 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.

II. Literature review

This paper builds on [2] to explore multi-modal, peer-to-peer communication between people and avatars in shared perceptual domains. Logically, our work here relates to many topics which have been well-studied, including human/avatar

[1] See Section II for additional discussion

978-1-7281-4125-1/19/$31.00 ©2019 IEEE
DOI 10.1109/HCC46620.2019.00015

interaction, embodied conversational agents, peer-to-peer communication, multi-modal interfaces, and semantics for reasoning in simulated environments.

It is well understood that people respond differently to embodied avatars then to non-embodied assistants [3], even performing better at certain tasks when embodied avatars are employed [4]. Users of avatars tend to take a more positive attitude, even attempting to make themselves appear better to the avatar, and they often assign the avatar a personality [5]. This motivates much of our work with Diana; however, there are downsides. When an avatar is not able to meet a user's goals as a human would, the user tends to get angry [6]. Anecdotally, as in [1], users of our system have described frustration when learning what Diana can and can't do, and especially when she doesn't recognize their communication properly. Additionally, while users may respond better to avatars than to non-embodied assistants, they respond better still to physically present robots [7]. While practical limitations preclude our presenting a system using a real robot, we expect much of what we are learning about embodied agents using sight, speech, gesture, physical manipulation, and shared perception to transfer to human-robot communication.

Indeed, considering both embodied virtual agents and real robots, evidence highlighting the benefits of multi-modal interfaces relative to speech-only approaches is easy to find. From cognitive psychology, we know that people tend to process speech and gesture somewhat independently, allowing humans using multi-modal systems to access greater working memory with less cognitive load than when using speech alone [8]. Combining language and gesture in computer interfaces is nothing new, first introduced by Bolt's "Put-that-there" [9]. Since then, many have followed suit as surveyed in [10]–[12].

What's more, multi-modal interfaces may offer additional practical advantages, as summarized by Reeves et al. [13]. Noisy or dark environments and settings where security or privacy is of importance are examples of circumstances in which having only a single mode of communication may be problematic. Further, Veinott et al. show that non-native English speakers benefit from video in English-language negotiation scenarios [14].

Sharing visual information has been shown to be particularly useful for establishing common ground [15]–[17]. Indeed, others have emphasized the importance of video and shared visual workspaces in computer-mediated communication, also highlighting the usefulness of non-verbal communication between humans [18]–[21]. This reinforces the need to provide a multi-modal interface to better encourage true peer-to-peer interactions.

The embodied conversational agent (ECA) [22] has logically risen as a combination of dialog-based systems and virtual avatars. There are many examples of ECAs, although we know of no previous systems which integrate embodied agents, two-way visual communication and shared perception. One such prominent example is the virtual classroom environment introduced by Barmaki et al. which was shown to be useful in studying the positive effects of non-verbal communication by teachers [23]. Their setup is similar to ours, with a Microsoft Kinect atop a large display watching for gestures. However, their agents are human-orchestrated, and the task is not facilitated with shared perception.

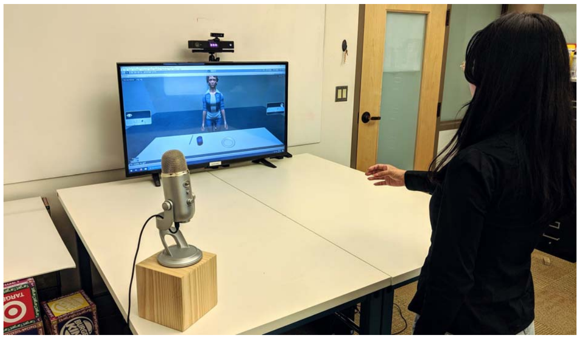


Fig. 1: System layout

Similarly in Carnell et al., a teaching and evaluation tool for doctors is presented [24]. In their work, virtual patients are hand-designed to provide a realistic doctor-patient interview, with successful prediction of real world medical domain performance. Perhaps most similar to our work, Mell et al. present a competition of automated negotiating agents with naive users [25]. These agents use emotion, deception, humor, and opponent modeling to attempt to win the most points in a negotiation game. While there are aspects of multi-modal communication and shared perception, these agents are fundamentally adversarial rather than cooperative and lack shared perception.

The work we are presenting here builds upon that presented by Narayana et al. [2]. As part of that earlier work, gesture elicitation studies were conducted in a peer-to-peer shared-perceptual environment. This step was key to us establishing how people expect other people to behave in the context of a shared task; typically that of building structures from blocks. The semantic gestures commonly used by people to solve problems are those employed here by our embodied agent, Diana.

III. Diana: An Embodied Agent

The best way to understand the impact of adding two-way visual communication and shared perception to an embodied agent is to walk up to Diana's table and interact with her, as in Figure 1. This section describes typical user interactions. The figures are taken from one of two videos of users with Diana included with the supplemental materials[2][3].

Users of our prototype system stand at one end of a table, as shown in Figure 1. A large monitor at the other end shows a virtual continuation of the table, with Diana standing at the end of it. The virtual half of the table has (virtual) blocks on it, and the user's task is to direct Diana to build block structures.

This setup is unique because it is multi-modal and symmetric. As with conversational agents, Diana can both hear

[2] Video #1: http://cs.colostate.edu/~vision/hcc19/video1.mp4

[3] Video #2: http://cs.colostate.edu/~vision/hcc19/video2.mp4

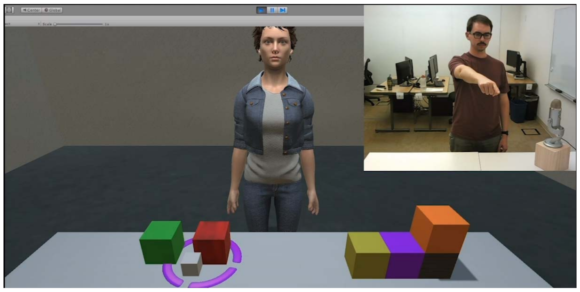


Fig. 2: Demonstration of signaler pointing.

and speak to the user. Unlike conversational agents, though, Diana can see the user and respond to the user's gestures, body language, and gaze[4]. She is aware of the user's presence in a way that conversational agents aren't. The user, of course, can also see Diana. And, since she is an avatar, she also gestures and directs her gaze. Perhaps most importantly, both the user and Diana can see the virtual blocks on the table, leading to *shared perception*. When Diana stacks the red block on the green block, she doesn't have to say what she is doing, the user can see what she is doing.

It is hard to overstate the importance of this second channel of communication. Contextual awareness is a general problem for embodied agents, and the introduction of a shared task space in which both person and computer can see each other and the physical manifestation of their shared attention profoundly expedites communication. It also necessitates new ways of formulating integrated speech, gesture, and visual human computer interaction.

As already alluded to above in our mention of elicitation studies, we have modeled Diana's behavior to the greatest extent possible after the results of human subject studies in which one person directs another to build block structures [2]. One observation from these studies was that users often wait to approach the table while they are planning how to build the block structure. At this point they stand back from the table and do not look at their partner on the other side. When they have a plan and are ready to begin, they step up to the table and look at their partner, who looks back. There is then an exchange of signals to confirm the start of the task. These signals can either be verbal (e.g. "ready?", "yes") or gestural (e.g. hand waves).

Interactions with Diana begin in roughly the same way. When the user steps up to the table and looks at Diana, in response Diana directs her attention, her gaze, back at the user. The user then waves to Diana, and Diana waves back while saying that she is ready. The user can then proceed with the task, knowing that Diana is watching and listening to them, and is ready to respond.

What happens next depends on the user and the block structure they want to build. Often, the next step is to reference a particular block that the user wants Diana to pick up or slide. In a multi-modal system, there are many ways to select a block. Verbally, Diana allows a block to be identified by its properties (e.g. "red block") or its location ("block on the left"). More often, however, users choose to point at the desired block (see Figure 2). Objectively, pointing is easier when there are many blocks on the table with similar properties, since verbal descriptions can quickly become complex and cumbersome: e.g. "The yellow block on the right of the table left of a green block and behind another yellow block." Interestingly, though, users prefer to point even when there are only a few uniquely colored blocks on the table. In practice, deixis [5] seems to be the preferred mode of reference in contexts with shared perception.

While pointing is a simple gesture for a person, it can be imprecise. For example, if two or more blocks are close together, Diana may be uncertain which block the user is pointing at. This is where the back and forth of a peer-to-peer dialog comes in. If the user points and Diana is unsure which block is being referenced, she starts asking the user questions. For example, she might ask "Do you mean the red block?" while gesturing toward that block. If the user says "yes" (or nods, or gives a thumbs up signal), the ambiguity is resolved. Otherwise Diana asks about the next most likely block until the ambiguity is resolved. If there are too many options, the user can interrupt by saying "nevermind".

Deixis is also the preferred mode of referencing locations. If the user wants Diana to move a block to an empty spot on the table, they can point at it, while perhaps saying something general like "put it there". In this case, the imprecision of pointing may mean that the block gets placed near but not exactly at the desired location. In this case, the easiest thing for the user to do is correct the error by having Diana slide the block a little, but the amount has to be carefully controlled. This can be done verbally ("slide to the left...a little more... little more... stop"), but again gestures appear to be the preferred mode, as shown in video #1.

Although gestures are the preferred mode for referencing visual objects and certain types of positions, the role of language in our multi-modal system should not be underestimated. Although numbers can be indicated gesturally by raising the appropriate number of fingers, users far prefer to say the words "one", "two", etc. More significantly, requests for actions are more often spoken than gestured, even though mimic-like gestures exist. This may be because there are a small number of well-defined actions in this domain, like grab, slide and pick up. (In human subject studies, the verb "rotate" was often accompanied by a gesture, perhaps to convey the axis of rotation.) It may also be that the greater ambiguity of language is useful. Users often say something like "put the red block next to the green one". This statement doesn't say how the block should be moved, for instance should it be slid or carried, because the user doesn't care. They just want the red block next to the green one.

Language is also useful for sequencing, because of its ability

[4] In the near future, Diana will also respond to facial expressions.

[5] From a practical perspective the term "deixis" is used here as a synonym for pointing, but our use of the term reflects a deep connection with disambiguation in the context of language understanding.

to make references across sentences. For example, users might say “Slide the red block forward. Now put the blue one next to it”. In this case, “it” is a reference to the red block. This is hard to do with gestures. After sliding the red block, the user has to point to the blue one and then back to the red one to have the same effect. Interestingly, in our system references are often cross-modal. A user can say “put this there”, while using deixis (pointing) to indicate “this” and “there”.

Once a person is engaged with Diana, a wide array of interactions arise drawing upon different combinations of speech and gestures, not to mention the goals naive users bring to each interaction. Since versions of the Diana system are up and running at all three partner institutions, i.e. Colorado State, Brandeis and Florida, hundreds of subjects have been introduced to, and allowed to play with, Diana. These sessions include both highly structured user studies and more informal sessions with visitors. From these interactions we’ve observed some broad trends. Perhaps most important, the majority of naive users intuitively understand the system well enough to successfully build structures: towers, staircases, or sometimes more imaginative structures. That naive users are able to successfully build structures shows us that task focused communication with an embodied agent capable of shared perception is both engaging and useful.

Another trend concerns how different users blend speech and gesture. As Diana understands sentences such as “Place the purple block on the red block”, some users quickly gravitate toward a language-dominant mode of communication. At the other extreme, Diana is able to interpret over 30 distinct gestures and some users rely almost exclusively upon gesture. This ability for users to have the system adapt to their favored mode of interaction is itself huge. For one thing, it supports the intuition that not all users seek the same modes of interaction. It will allow us in the near future to begin quantifying differences in users’ preferences.

Figure 3 provides a quick visual introduction to the different gestures used by a person and the actions of Diana. The person may be seen waving in greeting, pointing, indicating a block should be slid over, that a block should be grasped, and finally a thumbs up acknowledging successful completion of the staircase. Diana is shown waving in greeting, acknowledging understanding of the slide gesture, preparing to grasp a block, placing blocks, and finally with the hand over the white block suggesting understanding that the white block is the one that the user wants moved. The purple dashed circle is the only aspect of this entire interaction not readily reproduced with a real robot; it shows in real time where Diana thinks the user is pointing.

So far our examples concern blocks world, but clearly solving tasks in other spatial domains is key to understanding embodied agents with shared perception. A second domain is illustrated in Figure 4. This table setting domain currently includes plates, knives and cups. Objects such as these introduce new challenges because there is an arguably richer set of implied semantics along with more specialized ways to interact with these objects. For example, Diana can grab a cup from above with her fingers around the rim, like she grabs a block. This is fine for picking up a cup, but bad if the goal is to hold the cup to pour something into it, since in this position her hand covers the opening. For pouring, holding the cup by the side or the handle is better. Diana has a model for how to interact with objects based on affordances. The best way to grab the cup is determined by what you (most likely) intend to do with it. However, this specialization calls out for the need to teach Diana new grab gestures. Currently users can teach Diana new gestures online, and Diana will associate the new gesture with the particular grasp (see video #2). Technical details about one-shot gesture learning can be found in Yu [26]. See Figure 4 for examples of object affordances for which new gestures must be learned.

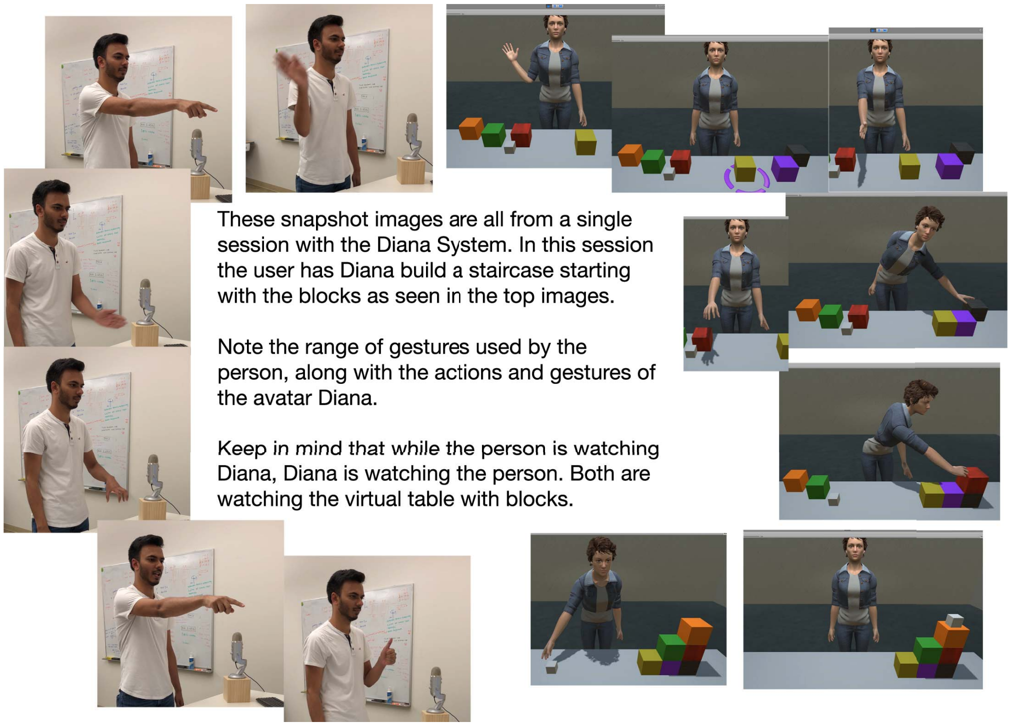
IV. System Design

Underneath the hood, Diana is a complex combination of independent processes. The processes involved in recognizing gestures, body position, and gaze are shown in Figure 5. A host process drives a Kinect sensor and uses its pose determination (a.k.a. skeleton) feature to determine the position of the user’s body, left hand, right hand, and head. It produces four data streams: three RGB-D video streams focused on the left hand, right hand, and head, and a pose stream of joint position data. Convolutional neural nets (CNNs) analyze the video streams for poses and motions consistent with known gestures [27]–[29], while a Long-Short Term Memory (LSTM) network analyses the pose stream for arm motions. The head stream is also used to analyze emotions and gaze. Finally, a state-machine-based fusion process collects the per-frame labels from all four streams to recognize gestures and significant body language (e.g., stepping toward or away from the table).

VoxSim [30], [31], the oval in the upper right of Figure 5, is the visual semantic reasoning engine containing Diana’s virtual world. Among other components, it contains a language model for processing/parsing verbal input, a nondeterministic push-down automaton architecture for tracking the visual, situational, and dialogue context available to both Diana and her interlocutor, and a semantic model of objects and events built on the VoxML modeling language [32] which extends the semantic typing structure of Pustejovsky’s Generative Lexicon theory [33]. This semantic model grounds object and event affordances to the particulars of the current situation throughout an interaction with Diana, and judges the feasibility and plausibility of potential actions based on affordances and the ability of an action to be successfully simulated.

V. Significance and Design Implications

As mentioned in the literature survey, there are many ECAs with capabilities that overlap Diana’s capabilities. What sets Diana apart from other state-of-the-art ECAs is her ability to facilitate peer-to-peer communication. Diana is aware of the physical world (i.e. the human user’s speech, gestures, gaze, and emotion), the visually-shared virtual world, and her own gestures and emotions. Diana suggests how future agents need to develop, including the need for physical embodiment,



These snapshot images are all from a single session with the Diana System. In this session the user has Diana build a staircase starting with the blocks as seen in the top images.

Note the range of gestures used by the person, along with the actions and gestures of the avatar Diana.

Keep *in mind* that *while* the person is watching Diana, Diana is watching the person. Both are watching the virtual table with blocks.

Fig. 3: Sample snapshots of both the user and Diana during a single session.

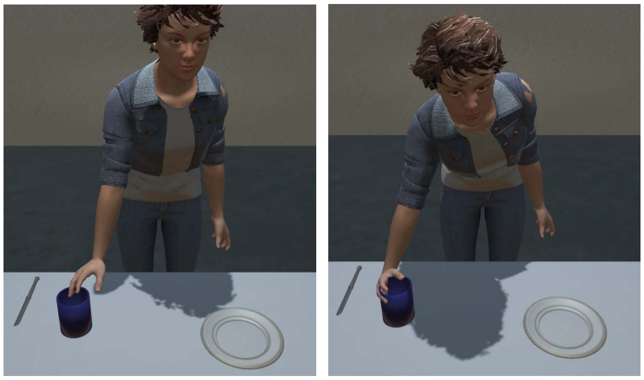


Fig. 4: Examples of non-standard object affordances

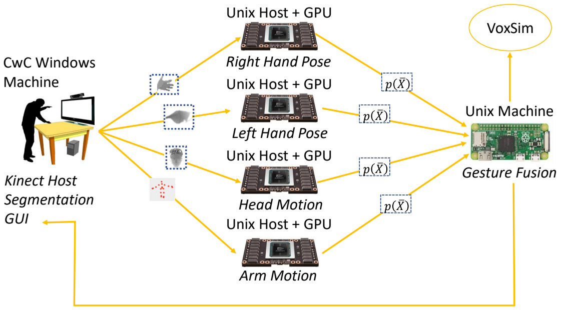


Fig. 5: System Design with multiple GPUs

visual as well as audio (speech) communication, and shared perception of elements in both physical and virtual reality. Down this path of agent development lie agents able to support the conceit that we are interacting with another human, or at least with an agent with human-like capabilities.

By design this paper focuses on what Diana as an Agent does, leaving aside technical descriptions of her implementation. Technical descriptions of her components appear elsewhere [2], [27]–[36]. However, there are a few aspects of the Diana system we wish to briefly highlight.

- **Elicitation**: Diana's behavior was modeled after the results of an observational study of human-to-human communication. Basing Diana's actions on huma studies supports the conceit that Diana is person-like.
- **Gesture Recognition**: Real-time gesture recognition is now feasible using banks of Convolutional Neural Networks (CNNs). It is well understood that the user experience degrades with as little as a 50ms delay [37] and that perceptible latency dramatically reduces human satisfaction [38], [39], but CNN technology allows gesture recognition within these constraints. Gesture recognition in turn re-enforces the conceit that Diana is seeing and responding as a person might.
- **Shared Perception**: When working on shared tasks, an important part of the dialog is what *doesn't* have to be said because both participants can see the work space. The fact that both Diana and the user can see the shared virtual world is critical to her success.

VI. Conclusion

We present Diana, an embodied agent. Diana is able to see the user's environment, understand her own environment, and manage the overlap of the virtual and real worlds. In the broad quest to humanize computing, few avenues of development are likely to prove more important in the long run than embodied agents that can see their users and display their virtual world to their users. User awareness and shared perception humanize our interactions with computers by successfully enabling people to interact with an avatar as if they were a person. In the near future, we predict people will come to expect agents to watch them, listen to them, and understand their shared surroundings.

Acknowledgment

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] E. Luger and A. Sellen, "Like having a really bad pa: the gulf between user expectation and experience of conversational agents," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2016, pp. 5286–5297.

[2] P. Narayana, N. Krishnaswamy, I. Wang, R. Bangar, D. Patil, G. Mulay, K. Rim, R. Beveridge, J. Ruiz, J. Pustejovsky *et al.*, "Cooperating with avatars through gesture, language and action," in *Proceedings of SAI Intelligent Systems Conference*. Springer, 2018, pp. 272–293.

[3] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan, "Embodiment in conversational interfaces: Rea," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 1999, pp. 520–527.

[4] R. E. Mayer and C. S. DaPra, "An embodiment effect in computer-based learning with animated pedagogical agents." *Journal of Experimental Psychology: Applied*, vol. 18, no. 3, p. 239, 2012.

[5] 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.

[6] 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 MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2017, pp. 1520–1522.

[7] 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.

[8] 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.

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

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

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

[12] S. Saunderson and G. Nejat, "How robots influence humans: A survey of nonverbal communication in social human–robot interaction," *International Journal of Social Robotics*, pp. 1–34, 2019.

[13] 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.

[14] "Video Helps Remote Work: Speakers Who Need to Negotiate Common Ground Benefit from Seeing Each Other," ser. CHI '99, New York, NY, USA.

[15] 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.

[16] H. H. Clark and D. Wilkes-Gibbs, "Referring as a collaborative process," *Cognition*, vol. 22, no. 1, pp. 1–39, 1986.

[17] 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.

[18] "Coordination of Communication: Effects of Shared Visual Context on Collaborative Work," ser. CSCW '00, New York, NY, USA.

[19] "Gestures over Video Streams to Support Remote Collaboration on Physical Tasks," vol. 19, no. 3.

[20] "Visual Information As a Conversational Resource in Collaborative Physical Tasks," vol. 18, no. 1.

[21] "Action As Language in a Shared Visual Space," ser. CSCW '04, New York, NY, USA.

[22] J. Cassell, J. Sullivan, E. Churchill, and S. Prevost, *Embodied conversational agents*. MIT press, 2000.

[23] R. Barmaki and C. E. Hughes, "Embodiment analytics of practicing teachers in a virtual immersive environment," *Journal of Computer Assisted Learning*, vol. 34, no. 4, pp. 387–396, 2018.

[24] S. Carnell, B. Lok, M. T. James, and J. K. Su, "Predicting student success in communication skills learning scenarios with virtual humans," in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*. ACM, 2019, pp. 436–440.

[25] J. Mell, J. Gratch, T. Baarslag, R. Aydoğran, and C. M. Jonker, "Results of the first annual human-agent league of the automated negotiating agents competition," in *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. ACM, 2018, pp. 23–28.

[26] Z. Yu, "One-shot learning with pretrained convolutional neural networks," Master's thesis, Colorado State University, 5 2019.

[27] P. Narayana, J. R. Beveridge, and B. Draper, "Continuous gesture recognition through selective temporal fusion," in *2019 International Joint Conference on Neural Networks (IJCNN). IEEE*, 2019.

[28] P. Narayana, R. Beveridge, and B. Draper, "Gesture recognition: Focus on the hands," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[29] P. Narayana, J. R. Beveridge, and B. Draper, "Analyzing multi-channel networks for gesture recognition," in *2019 International Joint Conference on Neural Networks (IJCNN). IEEE*, 2019.

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

[31] 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.

[32] J. Pustejovsky and N. Krishnaswamy, "VoxML: A visualization modeling language," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.

[33] J. Pustejovsky, *The Generative Lexicon*. Cambridge, MA: MIT Press, 1995.

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

[35] 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 *12th IEEE International Conference on Automatic Face & Gesture Recognition*, 2017.

[36] I. Wang, P. Narayana, D. Patil, G. Mulay, R. Bangar, B. Draper, R. Beveridge, and J. Ruiz, "Exploring the use of gesture in collaborative tasks," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '17, 2017, pp. 2990–2997.

[37] C. D. Wickens, "The effects of control dynamics on performance." 1986.

[38] S. Friston and A. Steed, "Measuring latency in virtual environments," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 4, pp. 616–625, April 2014.

[39] J. J. LaViola Jr, E. Kruijff, R. P. McMahan, D. Bowman, and I. P. Poupyrev, *3D user interfaces: theory and practice*. Addison-Wesley Professional, 2017.