An Interactive Virtual Reality Platform for Studying Embodied Social Interaction

Hui Zhang (huizhang@cs.indiana.edu)
Department of Computer Science; Indiana University
Chen Yu (chenyu@indiana.edu)
Department of Psychology and Brain Sciences; Indiana University
Linda B. Smith (smith4@indiana.edu)
Department of Psychology and Brain Sciences; Indiana University

Abstract

We present an interactive virtual reality platform for studying the role of embodied social interaction in the context of language learning. The virtual environment consists of virtual objects, a virtual table, and most importantly, a set of virtual students with different social-cognitive skills. Real users are asked to serve as language teachers and teach virtual learners object names. They can interact with virtual learners via gazing, pointing at and moving virtual objects as well as speech acts. Since both the virtual environment (what users see) and the virtual humans (whom users interact with) are controlled (pre-programmed), this provides a unique opportunity to study how real teachers perceive different social signals generated by virtual learners and how they adjust their behaviors accordingly. One primary result is that real people feel comfortable to interact with virtual humans in the virtual environment and treat them as social partners. Moreover, the platform allows us to record real people's multimodal behavioral data and analyze the data across individual participants to extract shared behavioral patterns. Overall, this work demonstrates the usefulness of virtual reality technologies in studying both human-human and human-machine social interactions.

Introduction

A better understanding of human-human interaction in language learning has long been a subject of fascination. Language learning is a social event between teachers and learners. Nonverbal communication, including body language, gaze, gesture, facial expression, is crucial for both smooth communication and effective learning. More specifically, body language signaled by a language teacher provides useful cues for a language learner to infer what the speaker intends to refer to in unknown (yet) language. For example, a deictic pointing action would single out one object from multiple ones in a natural scene and indicate the speaker's referential intentions [14]. Meanwhile, body language signaled by a language learner indicates his/her attentional state so that the language teacher can adjust behaviors accordingly to enhance interaction and learning. For instance, if the language teacher realizes that the learner is not engaged in the interaction, she would generate some actions to attract the learner's attention. On the other hand, if the learner is fully engaged, then the teacher would focus more on using body language to facilitate language learning (but not on engaging the language learner).

Although previous research demonstrates the importance of social cues in the laboratory environment [1], quantitative analyses of the role of social cues in real world is very difficult without interfering with the interaction itself. What is really needed is an approach to controlling dynamic interactions between the language teacher and the language learner. By doing so, we can decouple the social interactions between two agents and manipulate the parameters in the interaction dynamically and systematically in a well-controlled way. The present paper addresses this challenge by using state-of-art technologies in computer graphics and virtual reality.

In the past decade, applications of virtual reality (VR) technology have been rapidly developed with the advance of computer graphics software and hardware. Virtual Reality techniques provide a unique way to enable people to interact efficiently with 3D computerized characters in a computer-rendered environment in real time using their natural senses and skills. Recently there is a growing trend that VR can play an important role in basic research in a variety of disciplines including cognition[2], education [9, 4] and perception[11]. Among others, Jasso and Triesch presented a virtual reality platform for developing and evaluating embodied models of cognitive development in [6]. Turk et al.[13] introduced a paradigm for studying multimodal and nonverbal communication in collaborative virtual environment where a user's communication behaviors can be filtered and re-rendered in a VR environment to change the nature of social interaction.

In light of this, we present a new experimental paradigm that exploits VR technologies to decouple complex social interactions between two agents and to study the role of embodied social cues in language learning. Specifically, we hypothesize that naturalistic social influence can occur within immersive virtual environments as a function of two additive factors, behavioral realism and social presence. This paper takes the first steps towards this goal by designing and implementing a novel interactive virtual reality platform by asking real users to interact with virtual humans through various embodied social interactions. We report a case study of using this virtual reality platform with the evaluations of this platform in the context of a language learning task.

System Framework

Overview

We build virtual humans equipped (pre-programmed) with different kinds of social cognitive skills and ask real people to interact with virtual humans in a virtual environment.

Our VR interaction system consists of four components as shown in Figure 1:

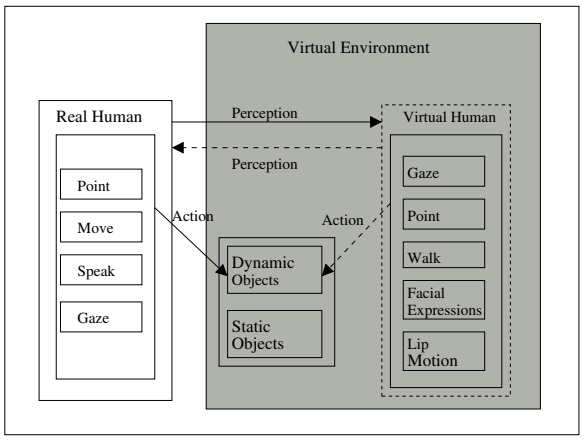

Figure 1: Overview of system architecture.

- A virtual environment includes a virtual laboratory with furniture and a set of virtual objects that real people can manipulate in real time via a touch screen mounted on a computer screen.
- Virtual humans can demonstrate different kinds of social skills and perform actions in the virtual environment.
- Multimodal interaction between virtual humans and real people includes speaking, eye contact, pointing at, gazing at and moving virtual objects.
- Data recording monitors and records a participant's body movements including pointing and moving actions on virtual objects, eye gaze, and speech acts in real time.

Building Virtual Humans

Appearance and Behavior One of the most important issues in our design is the "behavioral realism" of the virtual agents, which means that virtual humans should act and respond like a human, or in other words, they should be believable (see [12]), in both the physical actions of the agents themselves, and their social interactions with the human users.

The implementation at perceptual and motor levels is based on a human animation software package called $DI-Guy$, which is commercially available from Boston Dynamics Inc. It provides textured human characters with basic motor skills, such as standing, strolling, walking, running, sitting, etc. The actions of $DI-Guy$ characters can be scripted manually using an interactive tool called $DI-GuyScenario$. The other option, which is the one we use, is based on $DI-Guy$ SDK, allowing external C/C++ programs to control a character's basic motor repertoire. This SDK enables us to interface $DI-Guy$ to our extensive, high-level attentional and cognitive control software. A sample virtual human is shown in Figure 2, suggesting that using DI-Guy can result in smooth and lifelike movements being generated automatically.

Attentional State In our system, virtual humans can be programmed to behave to be engaged or disengaged in the

Figure 2: Interacting with virtual agent. The virtual lady is paying attention to the attentional objects on the virtual table.

interaction. If she is engaged, she will generate a set of actions, such as following the visual attention of a real person, paying attention to the objects that the real person is manipulated, and showing positive facial expressions. If she is not engaged, she would look somewhere irrelevant the real person's actions and generate negative facial expressions. We suggest that eye gaze plays a pivotal role in face-to-face interaction. Therefore, the simulation of cognitive skill is based primarily on avatar's eye gaze and pointing models evident in the psychological literature, and our simulation takes advantage of many techniques that have been widely used in other avatar interfaces (see [8], [10], [3], [7] and [5]).

The highest level of our eye gaze model is based on transitions between the two states (i.e., gazing at attentional objects and gazing away from attentional objects). The transition is triggered primarily by the passing of time in the current state, which is controlled by the level of engagement. And when the virtual human is engaged in a social conversation, he should gaze at the attentional object the human user is attending to. A further example in Figure 3 shows various engagement levels on multiple agents can be modeled to simulate a teaching and learning environment.

Figure 3: Modeling students with different levels of engagement.

Interaction and Data Recording

As shown in Figure 1, a user and a virtual human can interact through multiple channels including pointing at and moving virtual objects via hands, gazing at objects via eyes, and generating facial expressions. We have developed a multimodal data recording program that collects participants' speech, gaze movement on the computer screen, and actions on the touch screen mounted the display computer monitor. Speech

signals were sampled at 8000Hz and the sampling rate of both actions on the touch screen and eye gaze is 60Hz.

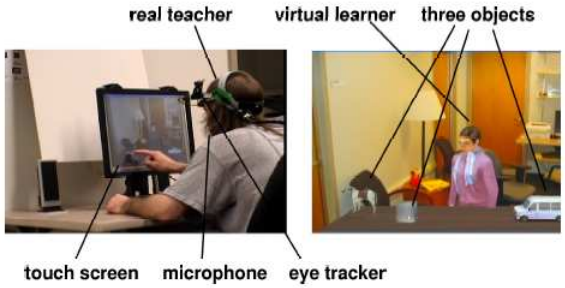

Figure 4: Left: a participant wearing an eye tracker and a microphone interacts with the virtual human in a virtual environment through a touch screen. Right: the VR scene consists of a virtual human and three objects on a table in each trial.

Platform Evaluation: Real humans teach virtual learners

As a first step to evaluate the usability of this platform, we designed an experiment in which real people were asked to teach virtual learners object names. We control the behaviors of virtual learners to create different learning situations, and measure how real people perceive the social-cognitive skills of different virtual people and how they adjust both their interactive behaviors and teaching strategies based on their perception of virtual learners.

Design and Procedure

As shown in Figure 4, real people were asked to teach virtual foreigners the names of several everyday objects. They were allowed to point to, gaze at and move those objects through a touch screen. There was no constraint about what they have to say or what they have to do. There were three conditions in this experiment wherein three virtual agents demonstrated different levels of engagement in interaction - engaged in 10%, 50% or 90% of total interaction time. When a virtual human is fully engaged in interaction, she would share visual attention with a real teacher by gazing at the object attended by a real teacher and generating positive facial expressions (e.g. smile, trust, etc.). While she is not engaged, she would look at somewhere else with negative facial expressions (e.g. sad, conniving, etc.). The objects attended by a real person are detected based on where he is looking as well as his actions on those objects through the touch screen. The attentional information is then sent to the virtual human so that she can switch her attention to the right objects in real time when she is in the engaged state.

We recruited 26 subjects who received course credits for participation. They were asked to interact with three virtual humans in total and one per condition. We randomly assigned the virtual humans to three levels of engagement, counterbalancing across participants.

There were six trials in each engagement condition and three virtual objects were introduced in each trial. Thus, participants needed to teach $3\times$ 6=18 objects in each condition and 54 objects in all of the three conditions. Whenever they thought that the virtual learner already acquired three object names in the current trial, they could move to the next trial. We recorded real people's behaviors in interaction including their pointing and moving actions, speech acts and eye gaze. Moreover, they were asked to complete questionnaires at the end of the experiment. The questionnaires measured social intelligence of three virtual learners. They were also asked to provide their estimates of the percentage of time the virtual humans followed the human teacher's attention.

Measure and Results

A 5-point Likert scale was used for a set of 10 questions in our questionnaire. Those questions focus on different aspects of participants' perception of the social-cognitive skills of three virtual humans:

- **Joint attention and eye contact** We measured how much the participants felt that eye movements of virtual humans were natural, social and friendly. A representative question contributed to this measure is "I felt that the agent did not look enough at me".
- **Social intelligence/engagement** We calculated a score to measure how much the participants felt that virtual learners were engaged during interaction (0-not engaged at all, 5-fully engaged). A representative question in this measure is "the agent and I interacted very smoothly".
- **Overall intelligence** We calculated a score to measure participants' estimates of virtual learners' intelligence. An example question used here is "the agent is smart".
- **Gaze time estimation**: Participants were also asked to estimate the amount of time (on a scale of 0 to 100 percent) that virtual humans paid attention to their behaviors.

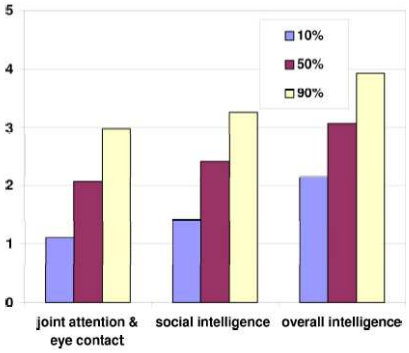

Figure 5: A comparison of participants' evaluation of three virtual humans.

Table 1: The estimated engagement times of virtual humans

	10%	50%	90%
gaze time	M= 22.50% SD= 22.10%	M=54.37% SD= 23.89%	M= 86% SD=16.1%

Figure 5 shows a comparison of the results of three virtual humans with different engagement levels. Clearly, participants

were aware of social behaviors of virtual humans and provided quite consistent estimates of their social sensitivities. Thus, the significant differences between three conditions are not surprising. We note that even when the virtual human almost fully engaged in interaction by following the real person's actions in 90% of the total time, most people were still not satisfied with the virtual human's social behaviors. Another observation is that they gave more credits to the high-level questions such as the overall intelligence of the virtual humans, but were less satisfied with more concrete issues, such as eye contact. This is true in all of the three conditions.

Table I shows the estimated times that virtual humans pay attention to participants' behaviors. Although the means of two out of three estimated times are close to 50% and 90% separately. Surprisingly, participants provided quite different estimates in all of three conditions. For instance, the low limit for the estimates in the 10% condition is 0%, indicating that some participants significantly overestimated the virtual human's engagement time. Meanwhile, some of them underestimate the times in the 90% condition as well. Further investigation is needed to explain this observation.

The purpose of these measures is to investigate whether the participants believe that they have been interacting with the representation of a real other (i.e., "social presence"). According to our experiments, social influence that occurs in the interacted virtual reality is accepted by the real participants. Our investigation shows that as far as those primitive actions generated by virtual humans look realistic, real people would treat them as social partners and are willing to interact with them.

Conclusion

Compared with using a real robot in a real environment, virtual humans are easy to implement and use mainly because we can neglect low-level technical problems, such as motor control of joint angles, which perfectly matches our research purposes. We are most interested in high-level social-cognitive skills in language learning. We attempt to answer how the behavioral-level actions, such as gazing and pointing, generated from both a language teacher and a language learner, are dynamically coupled in real time to create the social learning environment, and how the language learner appreciates those social cues signaled by the teacher. Moreover, the virtual platform has several special advantages in the study of social interaction: (1) Various virtual environments can be easily created and we can dynamically change or switch between different virtual scenes easily during an experiment; (2) the degree to fully control both virtual humans' behaviors and the virtual environment that real users and virtual humans share cannot be achieved with neither real robots nor human experimenters, which allows us to systematically study what aspects of the social environment are crucial for learning; and (3) we can easily maintain the consistency of the experimental environment and perfectly reproduce the experiments across multiple participants.

In summary, the present study proposes and implements a new experimental paradigm to study learning from multimodal interaction. We build virtual humans and control their behaviors to create different social partners that real people interacted with. We measured how well real people interact with virtual humans and how they shape their behaviors to adapt to different social-cognitive skills that virtual humans possess. We found that real people treat virtual humans as social partners when they interact with them, suggesting that we can further apply this experimental setup to create different interaction conditions by systematically manipulating the virtual human's behaviors.

References

[1] D.A. Baldwin. Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental Psychology*, 29:832–843, 1993.

[2] D.H. Ballard, M. M. Hayhoe, P.K. Pook, and R. P. Rao. Deictic codes for the embodiment of cognition. *Behavioural and Brain Science*, 1996.

[3] A. Colburn, M. Cohen, and S. Drucker. The role of eye gaze in avatar mediated conversational interfaces, 2000.

[4] S. D. Craig, B. Gholson, and D. M. Driscoll. Animated pedagogical agents in multimedia educational environments: Effects of agent properties, picture features, and redundancy. *Journal of Educational Psychology*, 94(2):428–434, June 2002.

[5] M. Garau, M. Slater, S. Bee, and M. A. Sasse. The impact of eye gaze on communication using humanoid avatars. In *CHI '01: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 309–316, New York, NY, USA, 2001. ACM Press.

[6] H. Jasso and J. Triesch. A virtual reality platform for modeling cognitive development. In G. Palm and S. Wermter, editors, *Biomimetic neural learning for intelligent robots*. Springer, 2005.

[7] K.F. MacDorman, T. Minato, M. Shimada, S. Itakura, S. Cowley, and H. Ishiguro. Assessing human likeness by eye contact in an android testbed. In *Proceedings of the XXVII Annual Meeting of the Cognitive Science*, page 226, Stresa, Italy, 2005.

[8] C. Peters and C. O'Sullivan. Bottom-up visual attention for virtual human animation. In *CASA*, pages 111–117, 2003.

[9] Jeff Rickel and W. Lewis Johnson. Task-oriented collaboration with embodied agents in virtual worlds. pages 95–122, 2000.

[10] T. Rist, M. Schmitt, C. Pelachaud, and M. Bilvi. Towards a simulation of conversations with expressive embodied speakers and listeners. In *CASA*, pages 5–10, 2003.

[11] C. S. Sahm, S. H. Creem-Regehr, W. B. Thompson, and P. Willemsen. Throwing versus walking as indicators of distance perception in similar real and virtual environments. *ACM Trans. Appl. Percept.*, 2(1):35–45, 2005.

[12] N. M. Thalmann, H. Kim, A. Egges, and S. Garchery. Believability and interaction in virtual worlds. In *MMM*, pages 2–9, 2005.

[13] M. Turk, J. Bailenson, A. Beall, J. Blascovich, and R. Guadagno. Multimodal transformed social interaction. In *ICMI '04: Proceedings of the 6th international conference on Multimodal interfaces*, pages 46–52, New York, NY, USA, 2004. ACM Press.

[14] C. Yu, D. H. Ballard, and R. N. Aslin. The role of embodied intention in early lexical acquisition. In *25th Annual Meeting of Cognitive Science Society (CogSci 2003)*, 2003.