Modeling Real-time Multimodal Interaction with Virtual Humans

Hui Zhang, Damian Fricker, and Chen Yu
Indiana University, Bloomington
{huizhang,chenyu,dfricker}@indiana.edu
http://www.indiana.edu/~dll/

Abstract—Natural social interactions can be very complex and comprise multiple levels of coordination, from high-level linguistic exchanges, to low-level couplings and decouplings of bodily movements. A better understanding of how these are coordinated can provide insightful empirical data and infer useful principles to guide the better design of human-computer interaction interface. In light of this, we propose and implement a research framework to model real-time multimodal interaction between real people and virtual humans. The value of doing this is that we can systematically study and evaluate different important aspects of multimodal real-time interactions between human and virtual agents. Our platform allows the virtual agent to keep track of the user’s gaze and hand movements in real time, and adjust his own behaviors accordingly. Multimodal data streams are collected in human-avatar interactions including speech, eye gaze, hand and head movements from both the human user and the virtual agent, which are then used to discover fine-grained behavioral patterns in human-agent interactions. We present an experiment based on the proposed framework as an example of the kinds of research questions that can be rigorously addressed and answered.

Index Terms—embodied agent, multimodal interaction, virtual human, visualization.

I. INTRODUCTION

Interacting agents engaged in a smooth coordinated task must seamlessly coordinate their actions to achieve a collaborative goal. The pursuit of a shared goal requires mutual recognition of the goal, appropriate sequencing and coordination of each agent’s behavior with others, and making predictions from and about the likely behavior of others. Such interaction is multimodal as we interact with each other and with intelligent artificial agents through multiple communication channels, including looking, speaking, touching, feeling, and pointing. In the case of human-human communication, moment-by-moment bodily actions are most of the time controlled by subconscious processes that are indicative of the internal state of cognitive processing in the brain [1]. Indeed, both social partners in the interaction rely on those external observable behaviors to read the other person’s intention and to initiate and carry on effective and productive interactions [2]. In the case of human-agent interaction, human users interacting with virtual agents perceive them as “intentional” agents, and thus are automatically tempted to evaluate the virtual agent’s behaviors based on their knowledge of the real-time behavior of human agents. Hence, to build virtual agents that can emulate smooth human-human communication, intelligent agents need to meet with the human user’s expectations and sensitivities to the real-time behaviors generated by virtual agents and perceive them in the similar way just as the user interacts with other humans.

II. A REAL-TIME HUMAN-AGENT INTERACTION PLATFORM

Our virtual experimental environment renders a virtual scene on a computer screen with a virtual agent sitting on the sofa and partially facing toward the computer screen so that she can have a face-to-face interaction with the human user. There are a set of virtual objects on the table in the virtual living room that both the virtual agent and the human user can move and manipulate. The virtual human’s manual actions toward those virtual objects are implemented through VR techniques and the real person’s actions on the virtual objects are performed through a touch-screen. There are several joint tasks that can be carried out in this virtual experimental environment. For example, the real person can be a language teacher while the virtual agent can be a language learner. Thus, the communication task is for the real person to attract the virtual agent’s attention and then teach the agent object names so the virtual agent can learn the human language through social interaction. For another example, the virtual agent and the real user can collaborate on putting pieces together in a jigsaw puzzle game. In such collaborative tasks, they can use speech and gesture to communicate and refer to pieces that the other agent can easily reach. In order to get insight into the sensorimotor level behaviors in human-agent interaction, our multimodal system records non-redundant system-wide time-stamped behavioral information from both the real person and the virtual agent (see e.g., Figure 1).

III. EXPERIMENT

Our study focuses on joint attention between a human user and a virtual agent. The joint task employed requires the human participant to teach the virtual agent a set of the (fictitious) names of various objects. We manipulated the engagement levels of the virtual agent to create three
interaction conditions – engaged 10%, 50%, or 90% of total interaction time. When the virtual agent is engaged, she would share visual attentions with the human agent. When she is not engaged, she would look at one of the other objects that the real teacher is not attending to with negative facial expressions. In this study, there are in total 18 learning trials, each of which consists of 3 to-be-taught objects. The human teachers can manually move any of the three objects through the touch-screen to attract the virtual learner’s attention first and then name it.

Multimodal data from both interacting partners were recorded and analyzed to examine how real people adapted their behavior to interact with virtual agents possessing different level of social skills. For instance, we are interested in how frequently the real teacher checks the virtual agent’s visual attention in three engagement conditions, how the virtual agent’s engagement may influence what the real teacher says and what actions s/he generates toward the to-be-learned objects, what are sequential multimodal behavioral patterns within an agent, and how the real agent may generate coupled adaptive actions based on the virtual agent’s state. More importantly, this new platform allows us to answer those questions based on moment-by-moment micro-level behavioral patterns as an objective way to access the smoothness of human-avatar interaction. We have developed a visual data mining system that allows us to analyze rich multimodal datasets to search for detailed time-course patterns exchanged between the human user and the virtual agent [3]. Fig.2 shows an example data analysis of our study. The first three streams are derived from raw action data from both the user and the agent. The first one is the Region-Of-Interest (ROI) stream from the virtual agent’s eye gaze, indicating which object the virtual agent attends to. The second stream is the ROI stream (three objects and the virtual agent’s face) from the human user’s gaze and the third one encodes which object the real person is manipulating and moving. The bottom four data streams show the speeds of objects and the user’s hand actions (on the touch-screen) in the interaction. We highlight 4 momentary interactive behaviors from those data streams (labeled from left to right on the top) to illustrate the kinds of patterns we investigate using our multimodal real-time platform: 1) joint attention: both agents visually attend to the same object (colored in blue); 2) eye-hand coordination: the human user gazes at an object while moving it through the touch-screen; 3) adaptive following: this sequential pattern starts with the situation that the virtual agent and the human user attend to different objects (step 1), and then the human user checks the virtual agent’s gaze (step 2) and follows the virtual agent attention to the same object (step 3) and finally reach to that object (step 4); 4) face-to-face: the virtual agent and the human user look towards each other’s face.

Fig. 1. An overview of system architecture. Top: A real person and a virtual human are engaged in a joint task with a set of virtual objects in a virtual environment. Bottom: multiple data streams are recorded from human-avatar interactions which are used to discover fine-grained behavioral patterns and infer more general principles.

Fig.2. Multimodal data streams and data analysis.

IV. CONCLUSION

Our proposed platform allows real people to treat virtual humans as social partners and provides a way for us to study and evaluate fine-grained behavioral data from both agents. We report our first steps to systematically and quantitatively study moment-by-moment micro-level behavioral patterns. This study investigating human-agent joint attention also reveals promising results about the role and functioning of joint attention in human-avatar interactions.

REFERENCES

