

# A Multimodal Real-Time Platform for Studying Human-Avatar Interactions

Hui Zhang, Damian Fricker, and Chen Yu  
Indiana University, Bloomington, USA  
{chenyu, huizhang, dfricker}@indiana.edu

**Abstract.** A better understanding of the human user’s expectations and sensitivities to the real-time behavior generated by virtual agents can provide insightful empirical data and infer useful principles to guide the design of intelligent virtual agents. In light of this, we propose and implement a research framework to systematically study and evaluate different important aspects of multimodal real-time interactions between humans 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 a pilot study based on the proposed framework as an example of the kinds of research questions that can be rigorously addressed and answered. This first study investigating human-agent joint attention reveals promising results about the role and functioning of joint attention in human-avatar interactions.

**Keywords:** embodied agent, multimodal interaction, visualization.

## 1 Introduction

Interacting embodied agents, be they groups of people engaged in a coordinated task, autonomous robots acting in an environment, or an avatar on a computer screen interacting with a human user, 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 (e.g., how much of an utterance they have processed, or how much of a situation they have comprehended)[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 (and experience) 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.

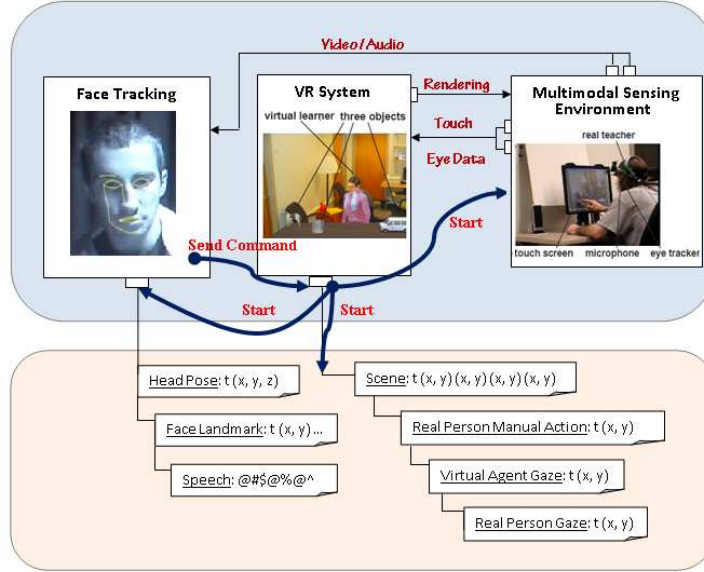
## 2 Related Work

The above requirement poses a particular challenge in the design of intelligent virtual agents as we need those agents to not only generate appropriate behaviors but also execute those actions at the right moment and with the right timing. For example, head nodding at the right moment may reflect a listener’s understanding as a back-channel feedback signal. In contrast, nodding at the unexpected moment may cause the speaker’s confusion in reading/accessing the listener’s attentional state. Similarly, a nodding action with abnormal timing may cause interruptions in communication. Indeed there is a growing research interest in studying real-time behaviors in human-computer interaction. In [3], an avatar generated reactive gaze behavior that is based on the user’s current state in an interview scenario. In [4], sequential probabilistic models were used to select multimodal features from a speaker (e.g. prosody, gaze and spoken words) to predict visual back-channel cues (e.g. head nods). [5] built an engagement estimation algorithm based on analyzing gaze transition patterns from users. [6] developed a real-time gaze model for embodied conversational agents that generated spontaneous gaze movements based on the agent’s internal state of cognitive processing.

## 3 A Real-Time Human-Agent Interaction Platform

The present work is specifically concerned with systematically studying the *exact timing* of *real-time interactions* between humans and virtual agents. To achieve this goal, we propose and implement a research framework for studying and evaluating different important aspects of multi-modal real-time interactions between humans and virtual agents, including establishment of joint attention via eye gaze coordination (an example application we will demonstrate by a pilot study described below), coupling of eye gaze, gestures, and utterances between virtual speaker and human listener in natural dialogues, and mechanisms for coordinating joint activities via verbal and nonverbal cues.

Specifically, there are three primary goals of building and using such a framework: 1) to test and evaluate moment-by-moment interactive behavioral patterns in human-agent interaction; 2) to develop, test and evaluate cognitive models that can emulate those patterns; 3) to develop, test and design new human-agent interfaces which include the appearance of the virtual agent, the control strategy as well as real-time adaptive human-like behaviors. More importantly, we expect to use this platform to discover fundamental principles in human-agent interaction which can be easily extended to various scenarios in human-computer interactions. Two critical requirements for our framework are that it be able to collect, in an unprecedented way, fine-grained multi-modal sensorimotor data that can be used for discovering coupled behavioral patterns embedded in multiple data streams from both the virtual agent and the human user, and that



**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. The platform tracks the user’s gaze and hand movements in real time and feeds the information to the virtual agent’s action control system to establish real-time perception-action loops with real and virtual agents. Bottom: multiple data streams are recorded from human-avatar interactions which are used to discover fine-grained behavioral patterns and infer more general principles.

the virtual agent can monitor the user’s behaviors moment by moment, allowing the agent to infer the user’s cognitive state (e.g. engagement and intention) and react to it in *real time*. To meet these requirements, we propose a framework consisting of four components (as shown in Figure 1): 1) a virtual experimental environment; 2) a virtual agent control system; 3) multi-modal sensory equipment; and 4) data recording, processing and mining. In the following, we will provide some details for each of the 4 components.

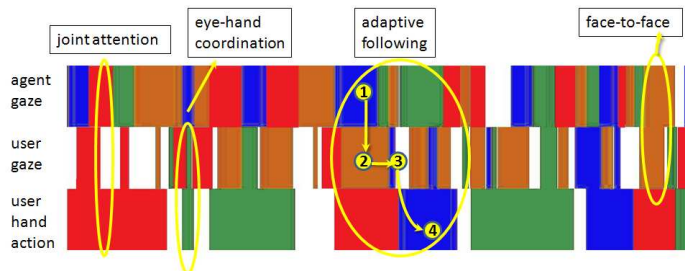
***The Virtual Experimental Environment.*** This virtual environment consists of a virtual living room with everyday furniture, e.g. chairs and tables. This virtual scene is rendered on a computer screen with a virtual agent sitting on the sofa and partially facing toward the computer screen so that she can have a face-to-face interaction with the human user. There are a set of virtual objects on the table (or on the floor) 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 which is covered on the computer monitor. There are several joint tasks that can

be carried out in this virtual 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 this collaborative task, they can use speech and gesture to communicate and refer to pieces that the other agent can easily reach.

***The Virtual Agent.*** In our implementation, we use Boston Dynamics DI-Guy libraries to animate virtual agents that can be created and readily programmed to generate realistic human-like behaviors in the virtual world, including gazing and pointing at an object or a person in a specific 3D location, walking to a 3D location, and moving lips to synchronize with speech while speaking. In addition, the virtual human can generate 7 different kinds of facial expressions, such as smile, trust, sad, mad and distrust. All these combine to result in smooth behaviors being generated automatically. A critical component in the virtual human is to perceive the human user’s behavior in real time and react appropriately within the right time span. As shown in Figure 1, this human-like skill is implemented by a combined mechanism including real-time eye tracking and motion tracking on the human side, real-time data transfer in the platform and real-time action control on the virtual human’s side.

***Multi-modal Sensory Equipment.*** As shown in Figure 1, our platform collects fine-grained behavioral data from both the virtual agent and the human user. On the human side, a Tobii 1750 eye tracker is used to monitor the user’s eye movements at the frequency of 50Hz. The user’s manual actions on virtual objects through the touch-screen is also recorded with timing information on a dedicated computer. Meanwhile, the system also records the user’s speech in the interaction. More recently, we added a video camera pointing to the face of the user and a faceAPI package from seeingmachine ([www.seeingmachine.com](http://www.seeingmachine.com)) is deployed and integrated into the whole system to record 38 3D face landmarks plus head rotation and orientation. On the virtual human side, our VR program not only renders a virtual scene with a virtual agent but also records gaze and manual actions generated by the virtual agent and his facial expressions. In addition, we also keep track of the locations of objects on the computer screen. As a result, we gather multimodal multi-stream temporal data streams from human-avatar interactions. All data streams are synchronized via system-wide timestamps.

***Multimodal Data Processing and Mining.*** 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 [7]. This system has played a critical role in our previous studies on human-human interaction and human-robot interaction [8, 9]. Our interactive



**Fig. 2.** Examples of multimodal data streams and data analysis. The 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 (e.g. gazing at one of the three virtual objects or looking straight toward the human user). 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. 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 red); 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. The goal of building the present framework is to study those moment-by-moment micro-level multimodal behaviors in human-avatar interactions.

visualization system provides an integrated solution with automatic and supervised data mining techniques that allow researchers to control (based on the visualization results) the focus of the automatic search for interesting patterns in multiple data streams. In particular, we use this system to not only detect the changes in each information channel (from either the human and the virtual agent data) but also consequential interactive patterns across two agents in real-time interaction. The insights from such active data exploration can subsequently be used for quantifying those interactive patterns as well as developing computational models of better human-agent interactions.

## 4 Preliminary Experiment

The overall goal of this research platform is to build a better human-agent communication system and understand multimodal agent-agent interaction. Joint visual attention has been well documented as an important indicator in smooth human-human communication. In light of this, our first pilot study focuses on joint attention between a human user and a virtual agent. More specifically, given the real-time control mechanism implemented in our platform, we ask how a human agent reacts to different situations wherein the virtual agent may or may

not pay attention to and follow the human agent’s visual attention. 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 follow the human teacher’s attention inferred from the teacher’s manual action and gaze, and then look toward the object that the real person is attending and meanwhile show interests by generating positive facial expressions. 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 pilot 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. Figure 2 shows an example to illustrate what kinds of multimodal behavioral patterns can be extracted from human-agent interaction and what kinds of research questions can be investigated using our platform. With this temporal window of only 30 seconds, the behaviors of both agents dynamically change moment by moment, creating various interactive patterns. Four joint activities are highlighted in Figure 2 (from left to right, see details in the caption): 1) joint attention; 2) eye-hand coordination; 3) adaptive following; and 4) face-to-face. We argue that multimodal human-avatar integration is made of those interactive patterns dynamically mixed with other joint activities between two interacting agents. Our ongoing research focuses on analyzing and comparing various joint action patterns across three engagement conditions to measure how the virtual agent’s attentional state influences human users’ moment-by-moment actions, and how human users adaptively adjust their reactive behavior based on their perception of the virtual agent’s actions.

## 5 Conclusion and On-Going Work

In multimodal human-avatar interaction, dependencies and interaction patterns between two interacting agents are bi-directional, i.e., the human user shapes the experiences and behaviors of the virtual agent through his own bodily actions and sensory-motor experiences, and the virtual agent likewise directly influences the sensorimotor experiences and actions of the human user. The present paper describes a multimodal real-time human-avatar platform and demonstrates the

potential of this platform for discovering novel and interesting results that can significantly advance the field of human-agent collaborative research. The learning task example is only one of many possible studies that could be conducted using the proposed framework. For example, another application of this framework is to determine the timing of back-channel feedback (from eye gaze, to gestures, to body postures, to verbal acknowledgments), which is critical for establishing common ground in conversations. This could include questions about how head movements, gestures and bodily postures are related to natural language comprehension, as well as general questions about the functional role of non-linguistic aspects of communication contributing to natural language understanding. Thus, with this real-time interactive system, we can collect multimodal behavioral data in different contexts, allowing us to systematically study the time-course of multimodal behaviors. The results from such research will provide insightful principles to guide the design of human-computer interaction. Moreover, those fine-grained patterns and behaviors can also be directly implemented in an intelligent virtual agent who will demonstrate human-like sensitivities to various non-verbal bodily cues in natural interactions.

## References

1. Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., Sedivy, J.C.: Integration of visual and linguistic information in spoken language comprehension. *Science* **268** (1995) 1632–1634
2. Shockley, K., Santana, M.V., Fowler, C.: Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance* **29** (2003) 326–332
3. Kipp, M., Gebhard, P.: Igaze: Studying reactive gaze behavior in semi-immersive human-avatar interactions. In: IVA '08: Proceedings of the 8th international conference on Intelligent Virtual Agents, Berlin, Heidelberg, Springer-Verlag (2008) 191–199
4. Morency, L.P., Kok, I., Gratch, J.: Predicting listener backchannels: A probabilistic multimodal approach. In: IVA '08: Proceedings of the 8th international conference on Intelligent Virtual Agents, Berlin, Heidelberg, Springer-Verlag (2008) 176–190
5. Ishii, R., Nakano, Y.I.: Estimating user's conversational engagement based on gaze behaviors. In: IVA '08: Proceedings of the 8th international conference on Intelligent Virtual Agents, Berlin, Heidelberg, Springer-Verlag (2008) 200–207
6. Lee, J., Marsella, S., Traum, D., Gratch, J., Lance, B.: The rickel gaze model: A window on the mind of a virtual human. In: IVA '07: Proceedings of the 7th international conference on Intelligent Virtual Agents, Berlin, Heidelberg, Springer-Verlag (2007) 296–303
7. Yu, C., Zhong, Y., Smith, T., Park, I., Huang, W.: Visual data mining of multimedia data for social and behavioral studies. *Information Visualization* **8** (2009) 56–70
8. Yu, C., Smith, L., Shen, H., Pereira, A.F., Smith, T.: Active information selection: Visual attention through the hands. *IEEE Transactions on Autonomous Mental Development* **2** (2009) 141–151
9. Yu, C., Scheutz, M., Schermerhorn, P.: Investigating multimodal real-time patterns of joint attention in an hri word learning task. In: HRI '10: Proceeding of the 5th ACM/IEEE international conference on Human-robot interaction, New York, NY, USA, ACM (2010) 309–316