2013

Real-time Embodied Agent Adaptation

Ryan M. Schuetzler
University of Nebraska at Omaha, rschuetzler@unomaha.edu

David W. Wilson
University of Arizona

Real-time Embodied Agent Adaptation

Ryan Schuetzler
University of Arizona
rschuetzler@cmi.arizona.edu

David W. Wilson
University of Arizona
dwilson@cmi.arizona.edu

***Abstract*—This paper reports on initial investigation of two emerging technologies, FaceFX and Smartbody, capable of creating life-like animations for embodied conversational agents (ECAs) such as the AVATAR agent. Real-time rendering and animation generation technologies can enable rapid adaptation of ECAs to changing circumstances. The benefits of each package are discussed.**

I. Introduction

Decision support systems provide key information to aid humans in complex decision-making and problem solving tasks [1]. Effective decision support can provide substantial gains in productivity in a number of different contexts [2], [3]. One area in which a need for efficient decision support has been demonstrated is that of processing border crossers entering a country [4]. In this context, human border agents must quickly and accurately assess many crossers' intentions, screening out those crossers with deceptive or harmful intentions. Even ignoring the fact that humans are generally very poor detectors of deception [5], the cognitive effort required for border agents to effectively process the volume of border crossers can be very demanding. These factors suggest the border crossing context to be a prime candidate for the productivity gains provided by effective decision support.

Recent research has focused on providing a useful decision support solution for use in the border crossing context [4], [6]. The initial stages of this program of research focused on identifying sensors that could effectively measure cues of deception or concealed information in a rapid, non-invasive way, as required by the border screening context. The product of this applied research initiative is an automated kiosk that uses embodied conversational agents (ECAs) to perform automated interviews of individuals, using a combination of sensors to detect behaviors often indicative of deceptive or otherwise malicious intentions [6]. We refer to this kiosk hereafter as the Automated Virtual Agent for Truth Assessments in Real Time (AVATAR). The AVATAR program of research is ongoing, and improvements are continually being developed and tested in both laboratory and field settings.

In today's complex, rapidly changing decision environment, decision-making activities are more difficult than ever before [7]. In response to this trend, modern decision support systems are becoming increasingly intelligent and adaptive in order to provide adequate support in dynamic situations [7]–[10]. The border context certainly qualifies as one in which evaluations and decisions are dynamic, with each assessment interview potentially raising unique issues or requiring unique information. A system designed to support such a complex environment must also be adaptable, providing useful and accurate information in many different situations.

The purpose of this research-in-progress paper is to report on recent efforts to enable real-time adaptability of the AVATAR. We first summarize the current framework used in the AVATAR. We then discuss several limitations of this framework in the context of dynamic decision support. A framework is then described that leverages the many strengths of the current framework, but which allows for dynamic adjustment of the interview process as needed. We then present a discussion of the potential benefits of this framework, and conclude with a summary of the future directions for research in this area.

II. Background on the AVATAR System

Following the design science research paradigm [11], the AVATAR research team repeatedly modifies and evaluates different aspects of the AVATAR system, with a particular focus on pragmatic value within the border crossing context. This results in an evolving product that continually benefits from new understandings of requirements and capabilities derived from experimental studies and field testing.

A. The Current AVATAR Architecture

The AVATAR system is currently comprised of several general conceptual components, each of which serves a separate purpose. For the purposes of this paper, we will discuss them in terms of three broad areas: data acquisition and fusion, intelligent agent system, and the embodied agent interface. The first of these–data acquisition and fusion–includes the acquisition of input data via attached sensors, and the fusion and analysis of that data into contextually meaningful information. Sensors attached to the AVATAR, such as microphones, a high-definition video camera, or an eye-tracking camera, collect behavioral data from interviewees. This collected data can include linguistics, vocalics, gaze behavior, pupil dilation, and others [6], [12]. The system must then interpret or fuse this data into information that is relevant to the rest of the system. For example, an attached microphone recording the spoken answers of an interviewee might capture a higher-pitched response to a given answer. This change in pitch, if significant, may indicate arousal and possible malintent [12]. If, when combined with information from other streams of input (e.g., pupil dilation or increased rigidity), there is a confluence of evidence suggesting possible deception, the fusion engine would then pass an indication to the interview logic, which could then cause the ECA agent to respond accordingly.

The second component of the AVATAR system is the intelligent agent system [6], which contains the decision logic

that drives the interview process. The interview logic specifies which questions are asked and in what order. It is the intelligence behind the ECA interacting with the interviewee. This logic is customized for different contexts in which the AVATAR is field-tested–an interview in an airport will require different interview logic than an interview at a pedestrian border crossing.

The final component of the AVATAR system is the ECA user interface. This component consists of a virtual, three-dimensional human-like agent, presented to the user on a display screen mounted on the kiosk [6]. The ECA responds to commands from the intelligent agent system, and communicates with the interviewee both verbally and nonverbally. The AVATAR research team has expended considerable empirical effort to better understand how changes in the attractiveness, demeanor, voice quality, or other aspects of the ECA influence perceptions and persuasive capability of system users [6], [13].

We note that the AVATAR system has been designed to be modular so that improvements to one component can be implemented without negatively affecting the other components. For example, recent research has investigated the utility of additional input sensors [14]. These additional streams of input can be added to the data acquisition and fusion component, augmenting the effectiveness of the intelligent agent system and the subsequent output of the ECA user interface.

B. Limitations of the Current Framework

The first two components of the AVATAR can potentially support an adaptive, dynamic interview process. The data acquisition and fusion component provides input to the system, which input will serve as the primary indicator of a need to adapt the interview process. The intelligent agent system currently uses a scripted set of logic, which does not adapt very extensively in response to input from the interviewee. One can clearly see, however, that with advances in artificial intelligence and with increasing richness of the input data provided by the data acquisition and fusion component, the interview logic will become increasingly intelligent in the future, demanding the ability for adaptive output to the user.

We note that the improvement of the first two components of the AVATAR falls outside the scope of this paper. We are solely concerned with the limitations associated with the third component, the ECA interface, which limitations are described next.

The current implementation of the ECA user interface greatly limits the extent to which the AVATAR can dynamically adjust the course of the interview. The ECA interface currently consists of a number of pre-rendered videos that depict the human-like agent speaking, synced with a pre-recorded human voice asking a question. These video recordings are played according to decision-tree logic embedded within the intelligent agent component of the system. As such, there is limited opportunity for the AVATAR to ask questions on-the-fly, as would be required if, for example, the AVATAR's sensors indicated a deceptive-sounding response about which the decision logic would like more related information. This is one of the key factors distinguishing human interviews from automated interactions–the human interviewer has the ability to ask probing questions when he or she detects suspicious responses regarding a particular question.

In summary, the AVATAR needs to effectively engage interviewees and provide useful, complete information to decision-makers in the complex, dynamic border crossing context. In order to provide such complete information, the AVATAR must be able to dynamically adjust the interview process in a way similar to that of a human interviewer, able to probe into suspicious responses and extract more, and more relevant, information. The current AVATAR framework is limited in its ability to dynamically interact with interviewees, most notably in the capabilities of the ECA interface. The framework we propose in the following sections attempts to remove these limitations from the ECA interface, providing a process through which an improved intelligent agent system can dynamically direct the ECA conducting the interview.

III. The Adaptive AVATAR

The focus of the framework proposed in this section is the ability to adapt the animation of the ECA to changing conversational requirements. Two primary routes can be taken to enable this adaptiveness. As the current process involves pre-rendered, animated video clips, the process is time-consuming and requires extensive computer processing. Creating a set of question for a simple interview involving yes/no questions can take hundreds of hours of computing time. To create an adaptable interviewer, we advocate the migration from pre-rendered video to the real-time rendering capabilities of a game system.

The proposed migration to real-time rendering will be a two-phase process discussed below. In the first phase, predefined animations can be created using animation software such as FaceFX [15]. In the second phase, animations and actions can be defined in real-time using SmartBody [16].

A. Phase 1 - FaceFX

In the first phase of creating a more adaptive ECA for the AVATAR, animations will continue to be defined in advance. The other components of the AVATAR do not currently support the artificial intelligence needed for a fully automated, intelligent interview. Thus, little will be done to modify the underlying logic and structure of the interview process. However, rather than creating and rendering the animations as video, software will be used to define the motion of a 3D character. Those predefine motions will be rendered only when needed, and in real-time using a game engine such as Unity or Unreal.

FaceFX [15] is a software package created by OC3 Entertainment that creates lip-synced 3D animations from an animated model and the audio clip to be animated. The audio clips can be either human voices or text-to-speech creations. In either case, FaceFX uses key frame animation to generate very realistic speaking animations. FaceFX automatically adds small nonverbal motions such as blinking, gaze, and head movements. These behaviors can be manually adjusted after the initial animation is created.

FaceFX is still limited, however, in that animations must be created, though not rendered, in advance. To adjust an

interview and, for example, add more questions, the animations would need to be created in FaceFX and then passed along to the computers used to conduct the interviews. This process follows the current video-based interview framework, but it offers one tremendous advantage. Unlike the current framework, which requires hours of processing time to render the speaking animations, the creation and updating of the interview sequence will be done in minutes.

B. Phase 2 - SmartBody

To further increase the adaptability of the AVATAR system, the next step will be to adopt the Smartbody behavior realization system [16]. Smartbody is a modular system designed for real-time animation of human characters. It incorporates game engine functionality for real-time rendering, along with modules for the real-time generation of animations. The Smartbody system uses the Behavioral Markup Language (BML) to represent actions to be performed by an animated actor. Instructions formatted as BML can include speech, gestures, and body movement.

Smartbody would increase the adaptability of the AVATAR beyond even FaceFX by enabling real-time generation of animations. Smartbody uses the same game engine rendering technology as FaceFX, but rather than requiring animations to be created ahead of time, Smartbody can use the BML signals to drive an ECA's animation on-the-fly. By replacing the animation system of the AVATAR with Smartbody, it would be possible to create and animate new questions using a dialogue manager.

The modular architecture of Smartbody allows for different components to be replaced without affecting others. For example, if a new text-to-speech engine is created that enables more human-like speech to be generated, the text-to-speech module can be replaced without requiring massive changes to the entire system. Lip sync and gesture animations would remain the same. If a certain project requires pre-recorded voices rather than text-to-speech, that module can be removed, and other animations will be unaffected.

C. Advantages

Several advantages will be realized when the AVATAR system is moved from pre-rendered video to real-time rendering with FaceFX or Smartbody.

1) Speed: The speed of development of new animations for new questions will dramatically increase. As mentioned above, the current animation process involves rendering dozens of high-quality videos requiring hundreds of hours of computing time. With real-time rendering, the creation of the lip-synced speaking animations takes seconds rather than hours. Using Ogre3D, an open source graphics rendering engine, with FaceFX, we were able to create high-quality speaking animations in minutes.

Using Smartbody, the speaking animations are created on-the-fly. Using text-to-speech, not even the audio has to be created in advance. The trade-off for this level of real-time adaptation is animation quality. The lip sync capabilities of FaceFX generate very lifelike movements. The lip sync in Smartbody is not as closely matched.

2) Adaptability: Along with the increases in speed come improvements in system adaptability. These improvements will be realized more fully with the use of Smartbody (Phase 2) than of FaceFX (Phase 1). The ability of the Smartbody system to generate lip-synced animations in real-time, along with the ability to use text-to-speech to speak, will allow the creation of a system that can quickly adapt to a wide variety of situations. For example, new questions could be added in a deployed system without requiring the creation of new animations, again requiring hundreds of hours to generate.

IV. Conclusion

The AVATAR research team is continually working to improve and expand the capabilities of the AVATAR. As a research in progress, this paper reports the result of initial feasibility testing that we have performed to investigate the usefulness and effectiveness of converting the interacting component of the AVATAR to a more adaptive framework. Initial testing indicates that the proposed migration could very quickly produce benefits in reduced rendering time and a more easily modified interview structure. In addition, moving from pre-rendered videos to real-time animations via a game engine lays the groundwork for a more adaptable ECA interaction as other technologies embedded within the AVATAR continue to improve, eventually giving rise to a more intelligent, dynamic interviewing agent.

References

[1] J. Shim, M. Warkentin, J. F. Courtney, D. J. Power, R. Sharda, and C. Carlsson, "Past, present, and future of decision support technology," *Decision Support Systems*, vol. 33, no. 2, pp. 111 – 126, 2002.

[2] G. DeSanctis and R. B. Gallupe, "A foundation for the study of group decision support systems," *Management Science*, vol. 33, no. 5, pp. 589–609, 1987.

[3] P. G. W. Keen, "Value analysis: justifying decision support systems," *MIS Quarterly*, vol. 5, no. 1, pp. 1–15, Mar. 1981.

[4] M. W. Patton, "Decision support for rapid assessment of truth and deception using automated assessment technologies and kiosk-based embodied conversational agents," Ph.D. dissertation, University of Arizona, 2009.

[5] B. DePaulo, J. Lindsay, B. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to deception," *Psychological Bulletin*, vol. 129, no. 1, p. 74, 2003.

[6] J. F. Nunamaker Jr., D. C. Derrick, A. C. Elkins, J. K. Burgoon, and M. W. Patton, "Embodied conversational agent–based kiosk for automated interviewing." *Journal of Management Information Systems*, vol. 28, no. 1, pp. 17–48, Summer2011 2011.

[7] G. Zhang, Y. Xu, and T. Li, "A special issue on new trends in intelligent decision support systems," *Knowledge-Based Systems*, vol. 32, no. 0, pp. 1 – 2, 2012.

[8] S. Chan and W. Ip, "A dynamic decision support system to predict the value of customer for new product development," *Decision Support Systems*, vol. 52, no. 1, pp. 178 – 188, 2011.

[9] C.-S. J. Dong and A. Srinivasan, "Agent-enabled service-oriented decision support systems," *Decision Support Systems*, no. 0, pp. –, 2012.

[10] E. Ngai, T. Leung, Y. Wong, M. Lee, P. Chai, and Y. Choi, "Design and development of a context-aware decision support system for real-time accident handling in logistics," *Decision Support Systems*, vol. 52, no. 4, pp. 816 – 827, 2012.

[11] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, Mar. 2004.

[12] D. Derrick, A. Elkins, J. Burgoon, J. Nunamaker Jr, and D. Zeng, "Border security credibility assessments via heterogeneous sensor fusion," *IEEE Intelligent Systems*, pp. 41–49, 2010.

[13] M. D. Pickard, "Persuasive embodied agents: Using embodied agents to change people's behavior, beliefs, and assessments," Ph.D. dissertation, University of Arizona, 2012.

[14] N. W. Twyman, "Automated human screening for detecting concealed knowledge," Ph.D. dissertation, University of Arizona, 2012.

[15] OC3 Entertainment. (2012) FaceFX. [Online]. Available: http://www.facefx.com

[16] M. Thiebaux, S. Marsella, A. Marshall, and M. Kallmann, "Smartbody: Behavior realization for embodied conversational agents," in *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 2008, pp. 151–158.