Record, Transform & Reproduce Social Encounters in Immersive VR - An Iterative Approach

Jan Kolkmeier
University of Twente
P.O. Box 217, 7500 AE, Enschede, The Netherlands
j.kolkmeier@utwente.nl

ABSTRACT

Immersive Virtual Reality Environments that can be accessed through multimodal natural interfaces will bring new affordances to mediated interaction with virtual embodied agents and avatars. Such interfaces will measure, amongst others, users' poses and motion which can be copied to an embodied avatar representation of the user that is situated in a virtual or augmented reality space shared with autonomous virtual agents and human controlled or semi-autonomous avatars. Designers of such environments will be challenged to facilitate believable social interactions by creating agents or semi-autonomous avatars that can respond meaningfully to users' natural behaviors, as captured by these interfaces. In our future research, we aim to realize such interactions to create rich social encounters in immersive Virtual Reality. In this current work, we present the approach we envisage to analyze and learn agent behavior from human-agent interaction in an iterative fashion. We specifically look at small-scale, 'regulative' nonverbal behaviors. Agents inform their behavior on previous observations, observing responses that these behaviors elicit in new users, thus iteratively generating corpora of short, situated human-agent interaction sequences that are to be analyzed, annotated and processed to generate socially intelligent agent behavior. Some choices and challenges of this approach are discussed.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]: Artificial, augmented, and virtual realities.

Keywords

Non-verbal Behaviors; Multimodal Interactive Systems; Virtual Reality

1. INTRODUCTION

In the Smart Social Systems and Spaces for Living Well project (S4Living, [1]), we are interested in employing smart mixed and virtual reality environments to improve the lifestyle of people. In such environments, artificial agents, avatars and their hybrids [6] could facilitate social encounters with remote visitors, as simulated companions or as intelligent actors in serious games as part of virtual reality training. Moreover, through natural, multimodal interfaces, users can be present in shared virtual spaces to collaborate with other remotely connected users.

The effectiveness of agents and avatars employed in such environments depends on more than their capabilities to engage in and facilitate conversation. They also need to understand and employ non-verbal behaviors during social interaction, such as facial expressions, gaze, gestures, posture and proxemics, but also touch and prosodics of speech. These behaviors can be employed to signal the agents' attention and intentions, to express agents' emotions and to affect how the agents' personalities are perceived - and of course, if the agent acts as an avatar, to extend these states and attributes from their controller to other agents and avatars.

How elaborate and interactive the implementations of such behaviors for artificial agents need to be depends on the affordances of the user interface through which one interacts with these agents. To illustrate this, consider a serious game where the user controls a virtual policeman that has to interrogate a virtual agent (VA) that acts as a suspect in a crime. The objective of the game is to teach the user that different non-verbal behaviors can be used to affect interpersonal stances between suspect and interrogator, which can be used as part of the interrogation strategy[1]. Now, if the user presses the 'intimidate'-key, the user's avatar may perform an 'intimidating' animation, signalizing a change in stance. This animation is metaphoric for a dominant stance. The keystroke does not allow the user to add or change qualities of this action. For the suspect agent to respond, only a single, predetermined animation that signals 'being intimidated' suffices and the interaction unfolding on the screen is understood by the user. We find that non-verbal behaviors of other agents need to respond appropriately to the capabilities that the interface translates to the user's avatar on the screen. In this case, they need to respond appropriately to the predetermined animation triggered by a keystroke.

In immersive Virtual Reality Environments (iVREs) however, the user is situated in a virtual space shared equally with those agents, taking a true first-person perspective. One ambition of iVR is the attempt to simulate reality [7]. To this end, natural interfaces such as those using full-body tracking will increase the user's capabilities to manipulate the environment and express him- or herself in a natural fashion using multiple channels of communication. Given the perspective, actions directed towards virtual interaction partners are more immediate, and VAs have higher immediacy when directing their actions towards the user. With sensors that can accurately estimate behavioral cues such as pose, gaze, or proxemics, there is

[1] This scenario serves as an example and is taken from a related project. It is not (yet) a definitive scenario in our own work.

ICMI '15, November 09-13, 2015, Seattle, W A, USA
© 2015 ACM. ISBN 978-1-4503-3912-4/15/11...$15.00.
DOI: http://dx.doi.org/10.1145/2818346.2823314

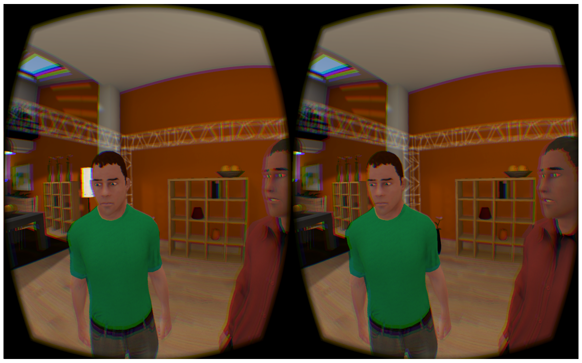

Figure 1: Two virtual agents manipulating their interpersonal distance and gaze behaviors in our previous study in iVR.

an opportunity to have agents interactively respond to these cues using their own channels of communication.

If the interrogation game was realized in iVR, a user should be able to take an intimidating pose by changing his physical pose only, not needing to make this explicit otherwise. The virtual suspect should respond by adjusting his or her pose appropriately, not by playing an animation metaphoric for 'being intimidated'. Effects of nuances in the user's behavior must be preserved. If a user modulates his interpersonal gaze while taking the intimidating pose, the agents gaze behavior and possible behaviors in other channels should adjust accordingly, for example to seek or avoid eye-contact.

The challenge is to design intelligent agents that can meaningfully employ such multimodal behaviors, that are interactive and continuous rather than metaphoric and static. To realize interaction, we must be aware what channels we can use in agent behavior, the channels used by users to response to such behavior, and their relationship. To some extend, we have addressed this in our previous work, which we present in the following section.

2. PREVIOUS WORK

In our previous work we examined the effects and relationships of agent behavior manipulations on avatar perception and behavioral responses of users. In [4], we situated virtual humanoid agents in augmented reality. Agents could exhibit social touch through the congruent visual and haptic cues of the VA touching the user's arm and the actuation of a haptic display on the user's arm. In a collaborative setting, touching agents were rated higher on affective attributes than non-touching agents. In our study in iVR [5], we examined the single and joint effects of gaze and proxemic behaviors during social interaction. Gaze and proxemic behaviors of two virtual humanoid agents (see Figure 1) were manipulated dynamically in terms of the perceived intimacy they elicit. Participants' gaze and proxemic responses were measured on-line. Participants showed strongest gaze and proxemic responses when agents manipulated both proxemic and gaze manipulations at the same time. More intimate manipulations such as standing closer and seeking more mutual gaze elicited gaze aversion and increase of personal distance from the participants. Less intimate manipulations such as increasing distance and averting gaze elicited more mutual gaze and reduction of personal distance from the participants. Agents that only manipulated gaze elicited weaker responses compared to agents that only manipulated proxemics. Agents with more intimate behaviors were rated higher in intimacy-related attributes, but lower on those related to warmth and trustworthiness.

Concluding from this previous work, there exists cross-channel interaction between agent and user behaviors in mixed and virtual reality settings, and these behaviors affect how agents' personalities are perceived. To build on this, our future research could contribute by disentangling this relationship even further for multiple modalities to allow agents to formulate and realize responses to detected cues in the appropriate channels during human-agent interaction. In the following section, we outline the envisaged method to approach our research.

3. MULTIMODAL SOCIAL INTERACTION ANALYSIS

We are interested in creating virtual agents and avatars that regulate non-verbal behaviors in an interactive fashion during social encounters in iVREs. With the increased equality of users' and agents' capabilities in such environments, agents could learn directly from the behaviors expressed by human users. The envisaged approach is to reproduce previously observed human behavior towards new users in an iVRE, and then learn from the response behaviors by these new users. Contingencies of these responses can further be examined by procedurally creating variations of behaviors as well as by comparing different responses between users.

3.1 Recording & Reproducing

'Equality of capabilities' means that whatever a user can do in VR, the agent can do as well. Given our approach of agents reproducing observed behaviors, we formulate an 'equality requirement', satisfied only by those communication channels in which users' activities can be recorded and later reproduced in VAs without significant loss in quality of the signal. Two examples of technologies that could satisfy the equality requirement for some modalities are motion capture and facial expression recognition to animate pose and facial expressions in the agent. Combinations of haptic interfaces and displays may also qualify here, measuring and displaying aspects of social touch. Emotions detected in speech prosody could be reproduced in agents speech using speech synthesizers.

3.2 Considered Behaviors & Cues

Not all conceivable behaviors may be meaningfully captured and reproduced in our approach. Recordings of highly contextual behaviors such as manipulating objects, pointing, miming or even laughing may not be suitable for reproduction unless we are also able to consistently reproduce the relevant context. Behaviors that we expect to be feasible and meaningful to reproduce and learn from are small-scale behaviors that are continuously and subconsciously regulated by humans. Think of regulation of interpersonal distance and gaze, posture, foot placement or the harmonic properties in the overall kinetic channels such as shakiness or calmness.

Detecting these behaviors is challenging. It is necessary to constrain this problem to a given context. For example, if we aim to realize an agent such as the virtual suspect from the example in the introduction, we may want to inform our method based on theories that describe the nonverbal cues during interrogation situations. For example, reducing interpersonal distance while engaging in eye-contact may be labeled as aggressive or competing user behavior when interrogating a virtual suspect agent.

3.3 Context & Scenario

As mentioned above, constraining the context of the interaction is crucial in managing the complexity of the problem. As of yet, we have not decided on a particular context or scenario for our research or for the application in which we want to demonstrate our results.

For the time being, the police interrogation scenario with the virtual suspect agent serves as our use-case. However, we aim keep the generalization of the method in mind, so that it is not confined to a solution for a single scenario.

3.4 Iterative Corpus Creation

To create a corpus that is suitable for learning, data needs to be collected where interaction between the considered behaviors can be observed. The classical virtual reality approach to examining interaction is to create a controlled environment with a virtual agent whose behavior is the same towards each participant in a given condition. This way, the VAs function as standardized interaction partners that make it possible to observe more than one response to the exact same behavior trigger.

The specified 'equality requirement' allows us to extend on this method as we may include observed responses in our agent's repertoire of behavior triggers. Again, this is only sensible for those behaviors that can be reproduced meaningfully without also reproducing the entire context of the original recording. One consequence of this approach is that we need to treat these regulative behaviors as equal. That is, an observed response behavior is from the same set of classes as the behavior that triggered this observed behavior. For example, mirroring [3] would translate to both the trigger and the response behavior being of the same class. If we observe a behavior response, we need to decide whether it belongs to an existing class, or whether we need to assign a new class to it, and then add it to the agent's repertoire for the next iteration of data collection.

Data collection is complete when we have a broad enough repertoire of behavior classes, and when we observed sufficient samples of responses to (each of) these behaviors.

3.5 Personality

We may expect variance in the responses to behaviors. One approach to explain this variance is to consider more contextual information. For the virtual suspect agent, a measure of the user's personality may be a good predictor of the user's response. For example, introvert users may respond to aggressive behavioral cues with submissive behavior, whereas extrovert users may respond with aggressive behavior. Such information may further be used to eventually let agents portray a given personality by learning responses from users with that personality.

3.6 Learning to Respond

For our agent to be able to respond dynamically to a user's behavior regulation, we need to train a classifier that separates the classes of known behaviors. Then it must classify observed user behavior online. If a known behavior is detected, the agent's response is a realization of a behavior class with the highest response probability given the contextual information such as the personality that the agent portrays.

3.7 Cross Modality

So far, we have treated behavior as a holistic, single signal. In practice, we may expect the considered behaviors to be multimodal. Some cues might only be mediated through a single channel, whereas others are only meaningful when all channels are considered at the same time. Further, as we have found in our previous work, regulating behaviors do not need to be symmetric across channels, but can also be cross channel. For example, a reduction in interpersonal distance may cause a regulating response in gaze behavior, such as averted gaze. This concept is to be integrated more carefully into our approach, as it will certainly inform architectural and methodological choices.

4. CHALLENGES

One fundamental challenge is to determine, once we have selected a context for our study, whether or not the relevant behaviors can be captured and preserved in the reproduction. Subsequently, can we conceive a scenario for data collection in which triggers elicit meaningful response behaviors reliably, rather than just noise.

Another challenge is the annotation of behavior segments. During interaction, the start and end of a segment needs to be determined. Behaviors from different channels related to the same action may have different starting and ending times. Then, behavior segments between interaction partners need to be related to each other to determine which segment is a response to a given earlier event. This does not only require to setup labeling guidelines for annotators, but also the integration (and possibly extension) of annotation tools into the process.

In our previous work, we manipulated the behavior of agents with an explicit goal in mind, such as being perceived as more intimate. In the proposed approach however, goals or intentions behind observed (and then reproduced) behavior segments may not be known. We therefore need to evaluate and annotate these segments. Some labels, such as those related to the current task, may be automatically computed on-line. Other labels, such as those related to the internal, emotional or cognitive state of the user, require rigorous manual annotation or off-line computation.

For meaningful reproduction, we may require further transformation of the observed data. As [8] points out, the intrinsic value of some measures such as the absolute gaze direction of a users may be limited. In terms of gaze behavior, the object that the user is looking at or the angle that the user's gaze is averted from the face of an agent may be more valuable than the absolute gaze direction.

There are also several technical challenges. The amount of data generated by capturing high-dimensional motion captures at high frame-rates must be stored and retrieved. Others even have suggested (soft) real time requirements for transmitting and reproducing activities in such applications [6].

5. RESEARCH PLAN

To develop our approach further, we focus on kinetic channels, such as gaze and pose. Intensive literature reviews on previous, related approaches and specifically on the machine learning aspects are still required.

Besides the literature study, in our first year we aim for the realization of a demonstrator that provides real-time mimicry agent, much as in [2]. Such a study would give insight to the feasibility of several aspects of our approach, such as the quality to be expected from reproducing behaviors, and a first insight to the technical challenges. Follow up studies would focus on selecting a context and on labeling processes of observations inside that context.

During the second year, we plan to employ the mimicry setup in a specific scenario. Recordings are to be annotated, to create a first corpus. Next we implement capabilities to retrieve earlier observations at runtime, to have users in a second iteration interact with agents that reproduce behaviors of other (previous) users. Besides growing our corpus, studies on including contextual information such as 'personality' could be used to evaluate the success of the approach. One research question could be to what extend we can have agents portray a given personality by selecting responses by previous users with that personality.

During the third year, we plan to implement a system that can automatically select and realize appropriate responses in a given context. For evaluation, we test again whether the chosen personality is perceived by the user - now with the agent acting autonomously.

From there, possible future studies could address the possibilities of hybrid avatar/agent approaches, specifically extending the learning and generation of behaviors from observations, the transformation of observations to non-human agents.

6. CONCLUSIONS

Natural interfaces in iVR bring new affordances and opportunities to social interaction with virtual agents and avatars. Our research aims to answer these affordances by having agents learn directly from user activity.

We contribute an extension of the classical virtual reality method of controlled behavior analysis by exploiting the possibility of having equal communication capabilities that allow recording and reproduction of multimodal behaviors by the artificial agents. Through controlled reproduction towards multiple other users, corpora of elicited responses are to be created iteratively and are to be used for teaching agents to predict and respond to future cues. Several challenges of this approach have been highlighted and need to be addressed in future research. An overall proposed plan has been presented.

7. REFERENCES

[1] 3TU. 2015. Smart Social Systems and Spaces for Living Well. (30 7 2015). http://www.3tu.nl/ht/en/research/research-programme/s4living/.

[2] Jeremy N Bailenson, Nick Yee, Kayur Patel, and Andrew C Beall. 2008. Detecting digital chameleons. *Comput. Human Behav.* 24, 1 (Jan. 2008), 66–87. DOI: http://dx.doi.org/10.1016/j.chb.2007.01.015

[3] Tanya L Chartrand and John A Bargh. 1999. The chameleon effect: the perception–behavior link and social interaction. *J. Pers. Soc. Psychol.* 76, 6 (1999), 893.

[4] Gijs Huisman, Jan Kolkmeier, and Dirk Heylen. 2014. With Us or Against Us: Simulated Social Touch by Virtual Agents in a Cooperative or Competitive Setting. In *Intell. Virtual Agents*. Springer, 204–213.

[5] Jan Kolkmeier. 2015. *Intimacy is Induced and Regulated Through Proxemic & Gaze Behaviour - A Study in Immersive Virtual Reality*. Master thesis. University of Twente.

[6] Daniel Roth, Marc E Latoschik, Kai Vogeley, and Gary Bente. 2015. Hybrid Avatar-Agent Technology - A Conceptual Step Towards Mediated "Social" Virtual Reality and its Respective Challenges. *i-com* 14, 2 (2015), 107–114.

[7] Mel Slater. 2014. Grand Challenges in Virtual Environments. *Front. Robot. AI* 1, May (2014), 1–4. DOI: http://dx.doi.org/10.3389/frobt.2014.00003

[8] William Steptoe and Anthony Steed. 2012. Multimodal Data Capture and Analysis of Interaction in Immersive Collaborative Virtual Environments. *Presence-Teleoperators Virtual Environ.* 21, 4 (2012), 388–405. DOI: http://dx.doi.org/10.1162/PRES_a_00123