Motion Capturing Empowered Interaction with a Virtual Agent in an Augmented Reality Environment

Ionut Damian
Human Centered Multimedia
Augsburg University
damian@hcm-lab.de

René Bühling
Human Centered Multimedia
Augsburg University
buehling@hcm-lab.de

Felix Kistler
Human Centered Multimedia
Augsburg University
kistler@hcm-lab.de

Mark Billinghurst
The Human Interface
Technology Lab New Zealand
Christchurch, New Zealand
mark.billinghurst@canterbury.ac.nz

Mohammad Obaid
The Human Interface
Technology Lab New Zealand
Christchurch, New Zealand
mohammad.obaid@hitlabnz.org

Elisabeth André
Human Centered Multimedia
Augsburg University
andre@hcm-lab.de

Extended Abstracts of the
IEEE International Symposium on Mixed and Augmented Reality 2013
Science and Technology Proceedings
1 - 4 October 2013, Adelaide, SA, Australia
978-1-4799-2869-9/13/$31.00 ©2013 IEEE

Abstract

We present an Augmented Reality (AR) system where we immerse the user's whole body in the virtual scene using a motion capturing (MoCap) suit. The goal is to allow for seamless interaction with the virtual content within the AR environment. We describe an evaluation study of a prototype application featuring an interactive scenario with a virtual agent. The scenario contains two conditions: in one, the agent has access to the full tracking data of the MoCap suit and therefore is aware of the exact actions of the user, while in the second condition, the agent does not get this information. We then report and discuss the differences we were able to detect regarding the users' perception of the interaction with the agent and give future research directions.

Author Keywords

Augmented Reality, Motion Capturing, Virtual Agent, Full Body Interaction, Natural Interaction

ACM Classification Keywords

H.5.1 [Information interfaces and presentation (e.g., HCI)]: Multimedia Information Systems.

Introduction

Virtual agents have been widely used in various domains (e.g. training, marketing, video games) to bridge the

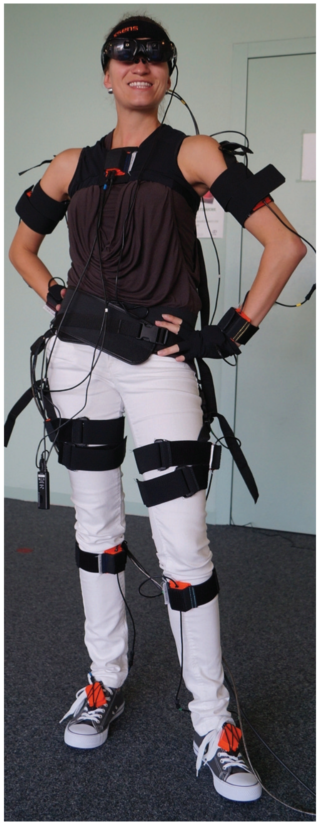

Figure 1: User wearing the proposed AR setup consisting of an inertial motion capturing suit and see-through HMD.

communication gap between users and computers. One key issue in this context is the credibility of the virtual agents as real persons. Researchers have investigated various solutions to this issue including high fidelity graphics [11, 8], human-like behaviors [4] and natural interaction between the agent and the user. However, whereas this issue has been thoroughly studied in the field of Virtual Reality (VR), virtual agents are rather new to AR environments [2, 6].

In this paper we argue that one way of enhancing the believability of virtual agents in an AR environment is by empowering their ability to sense the user, and thus increasing the realism of the human-agent interaction. To this end we present an AR system based on our previously developed approach [5] that immerses the user's whole body in the AR environment and allows for full-body natural interaction. To achieve this, the user wears a MoCap suit (Figure 1). In our case, we chose an inertial MoCap system that does not suffer from occlusion related tracking problems and also offers a higher freedom of movement thanks to an increased tracking range. This system not only handles the AR tracking but it also gives us access to a vast amount of information regarding the user's movements. The rendering of the virtual content is projected into the user's view using a see-through head mounted display (HMD).

The developed system allows the user to collaborate with a virtual agent within an AR environment to solve a task. Based on this system, we conducted a user study with 16 participants to measure the impact of our approach on the user experience. In particular, we were interested in whether the agent's ability to perceive the user's physical actions and respond with accurate social behaviors enabled by the enhanced MoCap tracking impacts the user's sense of spatial presence (being in the same space as the virtual agent), social presence (interaction similar to that with another person), social awareness (agent is able to perceive and respond to the user) and the believability of the virtual agent as a real person.

Related Work

Various attempts have been made to populate Augmented Realities with virtual agents. However, the perceptive capabilities of these agents are rather limited. One of the first AR application to use virtual agents was the ALIVE system [10] that allowed the user to interact with a virtual dog using gestures. Anabuki and colleagues [1] presented a virtual agent named Welbo which is able to perceive the user's position in the environment and react accordingly. Another example was presented by Wiendl and colleagues [13] in form of a Virtual Anatomy Assistant called Ritchie which teaches anatomy of the human body using a real skeleton. While the user is positioning virtual organs using a pointing device, the virtual agent provides verbal feedback on the correctness of the user actions. One key difference between these applications and the system proposed in this paper comes from the limited sensing abilities of the other systems. Our approach enables the virtual agent to know the exact position of the user and her/his joints at any point in time.

The System

Augmented Reality Setup

In order to immerse the user's full body in the AR environment, we use the Xsens inertial motion capturing suit[1]. The suit fulfills two roles. First, it handles the synchronization of the real and virtual environments. While this usually happens with the help of tracking

[1] http://www.xsens.com

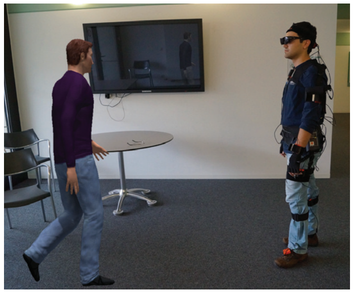

Figure 2: Illustration of what a user sees when immersed in an AR environment together with a virtual agent.

Feedback Class	Condition
smaller	$d < \frac{d_r}{2}$
slightly smaller	$d < d_r - 3cm \wedge$ $d \geq \frac{d_r}{2}$
larger	$d > d_r + 3cm$
equal	$d \geq d_r - 3cm \wedge$ $d \leq d_r + 3cm$

Table 1: Feedback classes. d is the current distance between the user's hands and d_r the requested distance.

markers (e.g. [12]), in our system the user acts as the synchronization point between the two environments. More precisely, the MoCap system computes the exact position and orientation of the user's head in the real world. This is possible because the MoCap suit not only tracks the skeleton configuration but also its position in the real world relative to a predetermined starting point (translation error $\sim 2\%$). Given the tracking of the user's head, we synchronize the real and virtual environments by continuously updating the virtual scene's camera position and orientation with the position and orientation of the user's head. This allows us to place any object in the virtual world and its position and orientation will be automatically updated to match the user's perspective, thus generating the AR effect without the need of markers. Figure 2 illustrates how the user perceives an AR environment with a virtual agent.

Secondly, the MoCap suit also computes accurate positions and orientations (orientation error $< 0.5\ deg$) of 23 joints in the user's body in real time at 120 Hz. This data can be used to create intuitive interaction modalities with virtual entities within the AR environment.

To simulate binocular vision we render the scene stereoscopically on the HMD, a see-through Vuzix Star 1200[2], using two different camera positions and frustums, one for each eye. The chosen HMD offers a resolution of 1280 x 720 with a diagonal field of view of 23 degrees.

Prototype Application

To test the impact of our approach on users, we developed a prototype application in which the user collaborates with a virtual agent to solve a predefined task. The virtual agent is implemented using the Advanced Agent Animation framework [4] and it is capable of executing both verbal and non-verbal behaviors.

First, the virtual agent instructs the user to position her/his hands at a certain distance apart. After the user repositioned her/his hands, the system computes the distance between them and provides feedback accordingly. For example, if the hands are less than what was requested, the virtual agent will instruct the user to move the hands further apart by using both synthesized speech and non-verbal behavior. This is repeated until the user reaches the requested distance.

In this context, two factors are crucial to generating credible interaction: accurate feedback timing and adequate feedback content. In order to compute when the virtual agent should give the feedback, the system continuously monitors the position of the user's hands as provided by the MoCap system. More precisely, it computes the deviations of the hand distances measured over the last 200 ms from the average hand distance of the last 1 second. If the average of these deviations exceeds 1 cm, the user is most likely repositioning her/his hands, otherwise, the hands are still. Using this algorithm, we can time the virtual agent's feedback to occur after the user finishes repositioning the hands (once the hands are still). Small scale pretests suggest that the algorithm has a near perfect accuracy in detecting when the users are repositioning their hands. The second crucial factor is deciding on the feedback content. Table 1 presents the different classes of feedback and the respective triggering conditions regarding the actual distance between the user's hands d and the requested distance d_r. In order to make the interaction less monotone, each feedback class contains multiple predefined utterances from which the system chooses at runtime. Additionally, the agent gazes

[2] http://www.vuzix.com

at the user's hands before performing the chosen utterance and also executes a gesture while the utterance is being spoken.

Evaluation

In order to evaluate the effect of the agent's perceptive capabilities enabled by the motion capturing system, we performed a user study where we confronted users with two versions of our system. The first version ($C1$) corresponds to the prototype application presented in the previous section in which the virtual agent is able to perceive the users physical behaviors and to generate corrective multimodal feedback using speech and non-verbal behavior. In the second version ($C2$) we do not use the data coming from the MoCap suit the user is equipped with. Instead, we generate randomized corrective feedback at predefined time intervals. Additionally, $C2$ is also limited in terms of the non-verbal behavior shown by the virtual agent. Whereas in $C1$ the virtual agent would gaze at the hands of the user before performing an utterance and gaze at the user's head while talking, in $C2$ the information on the position of the hands and head is not available. This means the virtual agent always looks straight ahead.

Considering how the MoCap component's enhanced tracking enables more natural behaviors in $C1$, we expect that the agent comes across as more believable in this condition. The more elaborated gaze behaviors in $C1$ should also contribute to the users' impression of interacting with a real person rather than with a computer. Furthermore, we expect the users to feel the agent is more aware of them in $C1$ than in $C2$ due to the agent's attentive gaze behaviors and accurate feedback. Finally, we anticipate an effect on the users' sense of spatial presence, i.e. they would rather have the impression of sharing the same physical environment with the agent in condition $C1$ than in condition $C2$. Based on these considerations, we formulated the following hypotheses:

- (H1) The believability of the virtual agent as a real person is higher in $C1$ than in $C2$
- (H2) The interaction with the virtual agent is more similar to an interaction with a human (rather than with a computer) in $C1$ than in $C2$
- (H3) Participants will have a stronger impression that the agent is aware of them in $C1$ than in $C2$.
- (H4) Participants will experience a greater sense of spatial presence in $C1$ than in $C2$.

Procedure and Participants

16 persons, 13 males and 3 females, with an average age of 28.75 took part in the evaluation of our system. Each person participated in both conditions and the order of the conditions was balanced between users. In each condition, the virtual agent asked the participant to perform 3 tasks: position hands 20 cm apart, 60 cm apart and 40 cm apart. After a task has been completed, event marked by the virtual agent uttering the "well done" message, the experimenter switched to the next task. At the same time, the virtual agent changed orientation and the participant was instructed to reposition so as to always face the agent directly. This was done in both conditions to ensure that the participants see the virtual agent from multiple angles, and thus experience the AR effect. Additionally, in order to increase the participants' sensation of interacting both with virtual and real entities, during the whole interaction, they were instructed to hold a hollow, 120 cm long rod and perform all tasks while holding onto it. This resulted

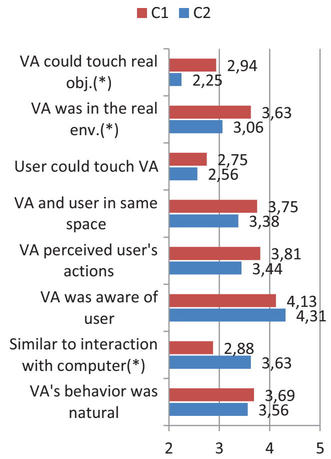

Figure 3: Questionnaire results. "VA" stands for virtual agent. Questions marked with (*) yielded significant differences.

in them simply sliding their hands on the rod when repositioning them to reach the requested distances.

After each condition, the participants were asked to fill out a questionnaire targeted at the aforementioned hypotheses. Answers to all questions should be given on a 5-point Likert scale ranging from strongly disagree to strongly agree. The questionnaire included items related to believability, social presence, social awareness and spatial presence.

Results and Discussion
A Kolmogorov-Smirnov test revealed that parts of the data extracted from the questionnaires was non-normally distributed. Therefore, we used Wilcoxon signed-rank tests to investigate differences between the answers to our questionnaires from the accurate condition ($C1$) and the random condition ($C2$).

We did not find any significant differences for the agent's believability. Despite the more sophisticated gaze behaviors, the agent's behavior was not perceived as more natural in $C1$ than in $C2$. Thus, $H1$ could not be confirmed. However, we got evidence for the validity of $H2$. Users had a stronger feeling of interacting with a computer (rather than with a real person) in $C2$ ($M = 3.62$) than in $C1$ ($M = 2.88$), $T = 4$, $p < .05$, $r = -.44$. These results are also in line with Garau and colleagues [7] who found that the eye gaze of an avatar that follows the flow of a conversation leads to a higher amount of co-presence. Surprisingly, we did not find any significant differences when asking the participants whether they had the impression the virtual agent was aware of their presence and observing them. Furthermore, the participants did not rate the agent's perceptive capabilities in $C1$ significantly different than in $C2$. As a reason, we assume that participants were not always able to validate whether the agent's instructions were correct. Indeed, some participants stated during short post-hoc interviews that even when they felt the feedback was odd, their personal insecurity in this situation caused them to accept the statement of the virtual agent and drop their own assessment of the distance. $H4$ has been partially confirmed. The participants' sensation of being in the same space did not significantly differ in the two conditions. Also they did not have a stronger impression they could touch the agent in $C1$ than in $C2$. However, they felt that the agent was more connected to the physical space in $C1$ than it was in $C2$. They indicated that the virtual agent was more in the same environment as the real objects in $C1$ ($M = 3.63$) than in $C2$ ($M = 3.06$), $T = 5$, $p < .05$, $r = -.44$. Further, the tests yielded that the virtual agent was more able to touch the real object in $C1$ ($M = 2.94$) than in $C2$ ($M = 2.25$), $T = 2.5$, $p < .05$, $r = -.39$. The results are illustrated in Figure 3.

Conclusion

In this paper we presented a system which immerses the user's whole body in an AR environment enabling intuitive interaction with a virtual agent. Using an evaluation with 16 users, we found that the virtual agent's increased awareness of the user's body enabled by the MoCap component does impact the user's sense of spatial presence, in particular, the perception that the agent had access to the real environment. Additionally, when using the more accurate gaze behaviors, the users also rated the interaction with the virtual agent as more human-like. Surprisingly, we were not able to find significant differences regarding the perceived awareness of the agent nor did we measure any impact on the believability of the agent as a real person. Overall, we were able to confirm two (one fully and one partially) out of our four initial hypotheses.

As part of our future work we plan to extend the complexity of the scenario to include additional virtual agents and objects. Furthermore, we are looking into developing new full body interaction modalities and measure their effect on the AR experience. For example, the MoCap data can be fed into a gesture or posture recognizer [9] to react to specific user behaviors or it can be used directly for precise object manipulation. Various expressivity features of the user's movements, such as fluidity, energy, spatial extent or overall activation [3], can also be computed in real time to give an insight into the user's affective state. A future vision of such interaction modalities is presented in the annexed video.

Acknowledgment. This work was partially funded by the European Comission within FP7-ICT-2011-7 (Project TARDIS, grant agreement no. 288578).

References

[1] Anabuki, M., Kakuta, H., Yamamoto, H., and Tamura, H. Welbo: an embodied conversational agent living in mixed reality space. In *CHI EA '00*, ACM (2000), 10–11.

[2] Barakonyi, I., and Schmalstieg, D. Augmented reality agents in the development pipeline of computer entertainment. In *Entertainment Computing - ICEC 2005*, F. Kishino, Y. Kitamura, H. Kato, and N. Nagata, Eds., vol. 3711 of *LNCS*. Springer Berlin Heidelberg, 2005, 345–356.

[3] Caridakis, G., Raouzaiou, A., Karapouzis, K., and Kollias, S. Synthesizing gesture expressivity based on real sequences. *Workshop on multimodal corpora: from multimodal behaviour theories to usable models, LREC Conference Genoa, Italy* (Mai 2006).

[4] Damian, I., Endrass, B., Huber, P., Bee, N., and André, E. Individualizing Agent Interactions. In *MIG '11*, Springer (2011).

[5] Damian, I., Obaid, M., Kistler, F., and André, E. Augmented reality using a 3d motion capturing suit. In *AH '13*, ACM (2013), 233–234.

[6] Dow, S., Mehta, M., Lausier, A., MacIntyre, B., and Mateas, M. Initial lessons from AR Façade, an interactive augmented reality drama. In *ACE '06*, ACM (2006).

[7] Garau, M., Slater, M., Vinayagamoorthy, V., Brogni, A., Steed, A., and Sasse, M. A. The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. In *CHI '03*, ACM (2003), 529–536.

[8] Geller, T. Overcoming the uncanny valley. *Computer Graphics and Applications, IEEE 28*, 4 (2008), 11–17.

[9] Kistler, F., Endrass, B., Damian, I., Dang, C. T., and André, E. Natural interaction with culturally adaptive virtual characters. *Journal on Multimodal User Interfaces 6* (2012), 39–47.

[10] Maes, P., Darrell, T., Blumberg, B., and Pentland, A. The alive system: wireless, full-body interaction with autonomous agents. *Multimedia Systems 5*, 2 (1997), 105–112.

[11] McDonnell, R., Breidt, M., and Bülthoff, H. H. Render me real?: investigating the effect of render style on the perception of animated virtual humans. *ACM Trans. Graph. 31*, 4 (July 2012), 91:1–91:11.

[12] Obaid, M., Niewiadomski, R., and Pelachaud, C. Perception of spatial relations and of coexistence with virtual agents. In *IVA '11*, Springer (2011), 363–369.

[13] Wiendl, V., Dorfmüller-Ulhaas, K., Schulz, N., and André, E. Integrating a virtual agent into the real world: The virtual anatomy assistant ritchie. In *IVA* (2007), 211–224.