Mixed Reality Methods for Analysis of Multimodal Human-Agent Interactions

JONATHAN HARTH*

Universität Witten/Herdecke, jonathan.harth@uni-wh.de

ALEXANDRA HOFMANN

Universität Witten/Herdecke, alexandra.hofmann@uni-wh.de

The ongoing development of embodied conversational agents requires a precise analysis of human-agent interaction. Currently, however, there are still only few approaches that investigate interactions by means of multimodal methods and both the individual reflection of experience and the interactive behavior. In this paper, we present a methodological approach that allows collecting data on individual perceptions of interacting with virtual agents as well as on the interaction itself. By means of mixed reality, the jointly coordinated behavior of users and agents in virtual spaces can be captured which enables a more comprehensive understanding of the complex dynamics of human-agent interactions.

Additional Keywords and Phrases: Mixed Reality, Mixed Methods, Virtual Reality, Multimodal Interaction, Videography, Embodied Conversational Agent

ACM Reference Format:

Jonathan Harth, Alexandra Hofmann. 2021. Mixed Reality Methods for Analysis of Multimodal Human-Agent Interactions. In CHI '18: ACM CHI Virtual Conference on Human Factors in Computing Systems, May 8-13, 2021, Yokohama, Japan. ACM, New York, NY, USA, 10 pages. NOTE: This block will be automatically generated when manuscripts are processed after acceptance.

1 INTRODUCTION

Even though conversational agents already reached a quite high level in processing natural language [1], humans are still far superior to virtual agents in processing multimodal information. In addition to using speech, humans use gestures, facial expressions, and more or less expressive body postures for communicating emotions, mental states, or relationships [2]. Nevertheless, speech recognition and language generation capabilities have made enormous advancements in recent years. Today, they deliver good results in many languages [4]. However, this is quite different when considering the aspects of "analogical communication" [17], which is based on gestures, glances, body movements, etc. In the domain of nonverbal expressions, current virtual agents only are able to express themselves particularly on a basic level. Moreover, virtual agents usually lack the competence to process those nonverbal messages on the part of human users. Even though specialized algorithms can already identify facial expressions, the situation is quite different when it comes to observing a user's hand movements, snorting, intonation, or body posture with regard to possible messages. Here, the problem is that analogical communication is usually ambiguous: "There are tears of sorrow and tears of joy, the clenched fist may signal aggression or constraint, a smile may convey sympathy or contempt, reticence can be interpreted as tactfulness or indifference, and we wonder if perhaps all analogic messages have this curiously ambiguous quality." [17]

* Place the footnote text for the author (if applicable) here.

Current virtual agents usually process only spoken language and thus miss out on further contextual information that would help to fully understand the communication. This leads to an increase of error-proneness in understanding conversations because contextualization is not processed [18]. However, it is the social context of a message that significantly contributes to the meaning of spoken words. So while ECAs are increasingly excelling in the area of processing spoken language their deficit in the area of nonverbal communication is becoming more and more apparent. For successful interactions, though, it is imperative that both digital and analogical communication is used successfully.

2 RELATED WORK

Recent development of (embodied) conversational agents shows that they are evolving from purely linguistic systems to actors that combine verbal and nonverbal communication. Vivid examples such as Virtual Mike [15], Mica [13] or Digital Douglas [7] illustrate that not only the visual approximation to human-likeness is getting closer, but also the interaction itself shows more and more similarities to human-human interaction. Even though users define the capabilities of speech interfaces differently than humans [8], it is clear that the concept of humanness is still the leading framework for evaluating (embodied) conversational agents.

These developments lead to an increasing focus on nonverbal communication as an option for interaction in the context of science as well [12]. At the same time, the aspect of nonverbal relationship management between agent and user, which is not yet fully developed, gains greater significance. Further, multimodal communication would be able to strengthen the coordination between user and virtual agent [11].

Thus, while ECAs [5] are becoming more technologically sophisticated and introduce more complex multimodal information into the interaction, existing methodological approaches have so far lacked the necessary tools to deal with this. In a recent meta-analysis of instruments used in human-agent interaction, it becomes clear that the question of “relationship” between user and agent usually remains untouched [9]. It seems as if the *social dimension* of interaction is overlooked by the predominantly psychologically motivated research on human-agent interaction. Here it becomes clear that interaction often is reduced to the perspective of only one participant: the human user and his/her impression of the agent [10]. From a sociological perspective, however, it can be stated that interaction is an emergent outcome that occurs whenever at least two actors have an effect on each other. For sociology, interaction is something third that occurs when at least two actors meet.

In most studies on human-agent interaction, this social aspect gets lost out of sight. Usually, neither the agent’s "impression" towards the user is taken into account, nor the interaction of both actors is analyzed in terms of jointly coordinated behavior. As a result, there are currently no standardized instruments on how to methodically control *possible discrepancies* between a user’s view on interaction and his/her behavior in the interaction [11]. It is exactly this blind spot, where our methodological approach comes into play: We suggest a sociological turn in studying interactions with embodied conversational agents by actually looking at the interaction itself!

3 METHODS AND MATERIAL

By using mixed reality methods, we are able to capture the jointly coordinated behavior of human users and virtual agents in virtual spaces. This allows a more comprehensive understanding of the complex dynamics in human-agent interactions and provides different types of data (see Table 1).

Table 1: Possible types of data

Time	Methods	Data
Pre-VR	Quantitative	Demographics
In-VR	Mixed	Behavior
Post-VR	Quantitative	Questionnaires
Post-VR	Qualitative	Interview

As technical equipment, we use a high-end VR-compatible PC (Geforce RTX 3080), a modern, high-resolution HMD (Valve Index), as well as a green screen studio with a total area of about 16 m² and appropriate lighting (Figure 1). The ECA to be tested is currently developed as part of the research project "Ai.vatar - the virtual intelligent assistant" (EFRE). The VR application is built in Unreal Engine and controlled by an individually designed Bot Management System. Project partners HHVision and IOX realized both features.The next subsections provide instructions on how to insert figures, tables, and equations in your document.

The agent combines Natural Language Processing via Google DialogFlow with a graphically realistic appearance (photogrammetric scans of a real person, see Figure 2). In this way, users can communicate with the agent in VR by using spoken language. Mixed reality rendering is enabled by implementing the LIV Suite directly into the application.

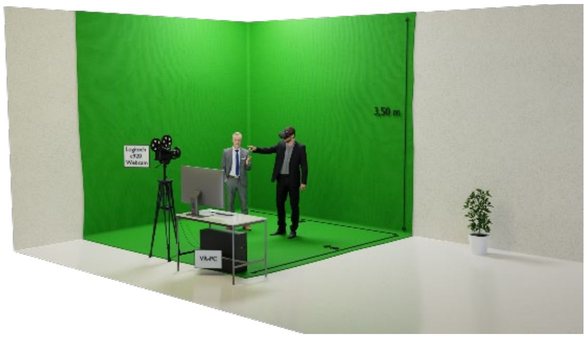

Figure 1: Schematic of the Mixed Reality Lab

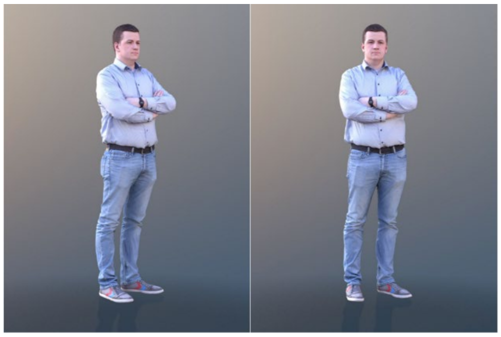

Figure 2: Static rendering of the virtual agent

Because we have not yet been able to conduct the planned study (due to COVID-19), the main focus of this paper is to outline the methodological approach. The main advantage of this methodology is that it allows capturing interactions in virtual reality that take place verbally and nonverbally. By using mixed reality and conversational analysis [16], we are able to capture the jointly coordinated behavior of human users and virtual agents in virtual spaces. In contrast to just verbalizing what is happening, this approach provides a more complete picture of human-agent interactions.

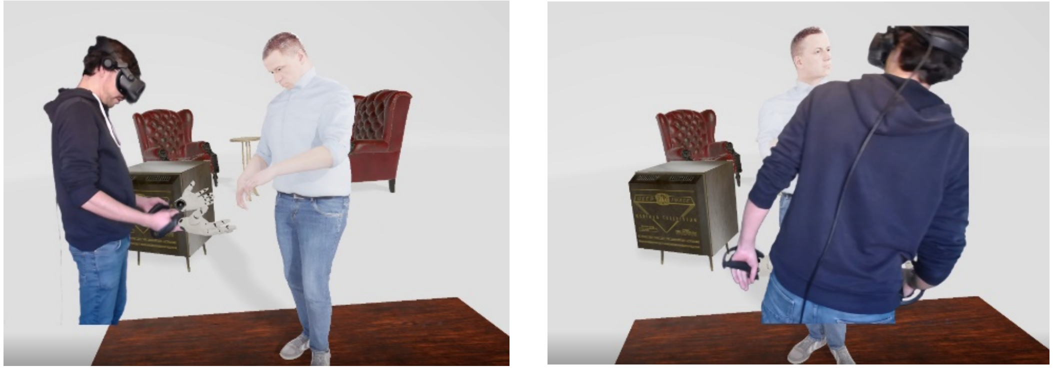

Figure 3: Mixed reality capturing of user-agent interaction

This multi-method approach enables the evaluation of user-agent interactions from multiple perspectives and allows a more comprehensive understanding of the complex dynamics in human-agent interactions. In summary, this paper describes the theoretical background and methodological procedures of analyzing interactions between users and virtual agents by means of mixed reality. This approach aims at using mixed reality videography to create an additional data set that can be compared with data on the user's experiential perception of the interaction. In this way, the videographic turn of common conversational approaches enables a more complete observation of the emergent social behavior between user and agent.

Current technological research and development is heading in precisely this direction: Recently, DeepMind announced their new "research program whose goal is to build embodied artificial agents that can perceive and manipulate the world, understand and produce language, and react capably when given general requests and instructions by humans" [6]. Facebook as well is fully committed to develop both situated and multimodal agents for rich human-agent interactions [14].

Apart from the multi-method nature of this methodology, our approach manages to leave behind the boundary between the virtual and the real world. Whereas in most studies users typically sit in front of a computer screen to interact with the virtual agent, in this study the user actually visits the agent in "its" habitat - the virtual reality of the application. The VR medium enables the creation of a virtual face-to-face situation instead of a (distanced) screen-mediated situation. This is very beneficial to studies dedicated to interaction, because face-to-face interaction is still considered the gold standard for conversations and enables the richest and most natural interaction [3].

ACKNOWLEDGMENTS

The European Regional Development Fund (ERDF) funded this research. We would like to thank our project partners HHVision for the visual representation of the agent's body, as well as IOX for the creation of the Bot Management System.

4 REFERENCES

[1] D. Adiwardana M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, Q. V. Le, Towards a human-like open-domain chatbot, arXiv preprint (2020). Doi: arxiv-2001.09977

[2] H. M. Aljaroodi, M. T. Adam, R.Chiong, T. Teubner, Avatars and embodied agents in experimental information systems research: A systematic review and conceptual framework, Australasian Journal of Information Systems 23 (2019). doi: 10.3127/ajis.v23i0.1841.

[3] J. B. Bavelas, Sarah Hutchinson, Christine Kenwood and Deborah H. Matheson.1997. Using face-to-face dialogue as a standard for other

communication systems. Canadian Journal of Communication 22, 1, 5.doi: 10.22230/cjc.1997v22n1a973

[4] J. Cahn. CHATBOT: Architecture, design, & development. University of Pennsylvania School of Engineering and Applied Science Department of Computer and Information Science. 2017. URL: https://www.academia.edu/37082899/CHATBOT_Architecture_Design_and_Development

[5] J. Cassell, J. Sullivan, E. Churchill and S. Prevost. 2000. Embodied conversational agents, MIT Press.

[6] DeepMind Interactive Agents Group. 2021. Imitating Interactive Intelligence. arXiv:2012.05672v2.

[7] Digital Domain. Introducing Douglas - Autonomous Digital Human, 2020. URL: https://www.youtube.com/watch?v=RKiGfGQxqaQ.

[8] P. R. Doyle, J. Edwards, O. Dumbleton, L. Clark, B. R. Cowan, Mapping perceptions of humanness in intelligent personal assistant interaction. in: Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services, ACM Press, New York, NY, 2019, pp. 1-12. doi: 10.1145/3338286.3340116.

[9] S. Fitrianie, M. Bruijnes, D. Richards, A. Abdulrahman, W.-P. Brinkman, What are We Measuring Anyway? -A Literature Survey of Questionnaires Used in Studies Reported in the Intelligent Virtual Agent Conferences. in: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, ACM Press, New York, NY, 2019, pp. 159-161. Doi: 10.1145/3308532.3329421

[10] S. Fitrianie, M. Bruijnes, D. Richards, A. Bönsch, W.-P. Brinkman, The 19 Unifying Questionnaire Constructs of Artificial Social Agents: An IVA Community Analysis. in: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents, ACM Press, New York, NY, 2019, pp. 1–8. doi: doi.org/10.1145/3383652.3423873

[11] M. E. Foster, Face-to-face conversation: why embodiment matters for conversational user interfaces. In: Proceedings of the 1st International Conference on Conversational User Interfaces, ACM Press, New York, NY, 2019, pp. 1-3. doi: 10.1145/3342775.3342810

[12] K. Kim, L. Boelling, S. Haesler, J. Bailenson, G. Bruder, Greg F. Welch, Does a digital assistant need a body? The influence of visual embodiment and social behavior on the perception of intelligent virtual agents in AR. In: IEEE International Symposium on Mixed and Augmented Reality (ISMAR), IEEE, 2018, pp. 105-114. doi: 10.1109/ISMAR.2018.00039

[13] MICA. Magic Leap's Mica at GDC. 2019. https://www.youtube.com/watch?v=-PzeWxtOGzQ.

[14] S. Moon, Satwik Kottur, Paul A. Crooky, Ankita Dey, Shivani Poddary, Theodore Levin, David Whitney, Daniel Difranco, Ahmad Beirami, Eunjoon Cho, Rajen Subba, Alborz Geramifard. 2020. Situated and Interactive Multimodal Conversations. Retrieved from: https://github.com/facebookresearch/simmc.

[15] M. Seymour, C. Evans, K. Libreri, Meet Mike: epic avatars. in: ACM SIGGRAPH 2017 VR Village (SIGGRAPH '17). Association for Computing Machinery, New York, NY, USA, 2017, pp .1–2. Doi: 10.1145/3089269.3089276.

[16] J. Sidnell, Conversation analysis. An introduction. Chichester: Wiley-Blackwell, 2010.

[17] P. Watzlawick, J. H. Beavin, D. D. Jackson, Pragmatics of Human Communication. A Study of Interactional Patterns, Pathologies, and Paradoxes. New York, W.W. Norton & Company, 1967.

[18] B. Weiss, I. Wechsung, C. Kühnel, S. Möller, Evaluating embodied conversational agents in multimodal interfaces, Computational Cognitive Science 1.6 (2019). doi: 10.1186/s40469-015-0006-9.