Mixed Reality Methods. Preliminary Considerations for Multimodal Analysis of Human-Agent Interactions

Jonathan Harth[a], Alexandra Hofmann[a]

[a] *Universität Witten/Herdecke, Alfred-Herrhausen-Straße 50, 58448 Witten, Germany*

Abstract

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. This approach enables a more comprehensive understanding of the complex dynamics of human-agent interactions and offers the advantage of combining different types of data.

Keywords 1

Mixed Reality, Human-Agent-Interaction, Embodied Conversational Agent, Videography, Virtual Reality, Multimodal Interaction, Nonverbal Communication, Mixed Methods

1. Introduction

Intelligent, virtual voice assistance systems such as Siri, Alexa, or Google Assistant have become popular technologies for verbal interactions between humans and computers [10]. The steady increase in the quality of such systems leads to an increase in their acceptance and trustworthiness [16] and thus to a greater integration into everyday life and research [29].

In addition to solely speech-driven communication systems, the use of visually embodied assistants is also more and more popular [2]. So-called embodied conversational agents (ECA) are more or less lifelike animated characters that can engage in direct conversation with human users [5, 15]. Virtual agents can be used on screen media as well as in virtual reality (VR). ECAs are especially useful in contexts where social interactions are important, i.e. as trainers, interviewers, or therapists [9, 25].

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. 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" [29], 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 in terms of probable

Joint Proceedings of the ACM IUI 2021 Workshops,
April 13–17, 2021, College Station, USA
EMAIL: jonathan.harth@uni-wh.de (J. Harth);
alexandra.hofmann@uni-wh.de (A. Hofmann)
ORCID: 0000-0002-8433-0896 (J. Harth)

emotional 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." [29]

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 [31]. 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 [26], Mica [20] 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. Paradigms such as the "Uncanny Valley" [21], are continuously reduced by state-of-the art technology by providing simulated emotional states, animated micro-expressions of the face and body, and a further increased level of detail all together [19]. 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 further developments lead to an increasing focus on nonverbal communication as an option for interaction in the context of science as well [16]. At the same time, the aspect of nonverbal relationship management between agent and user, which is not yet fully developed, gains greater significance. Recent studies show that users prefer embodied, realistic, and human-like visualizations of assistants to assistants without visualization [24, 28]. The visual animation of a conversational agent contains important communicative cues, such as eye contact, and thus enables easier use of the system [24]. An embodied virtual agent also promotes intuitive understanding while leading to greater connectedness [15] and trust in the system [16]. In human-agent interaction, the embodiment of the agent can thus change the social psychological dynamics in the interaction [6]. Further, multimodal communication would be able to strengthen the coordination between user and virtual agent [13]. In addition, Gong [14] shows that high human resemblance of ECAs leads to more positive evaluations of the system overall and more human characteristics are attributed to the embodied agent as well. An early meta-analysis supports these findings [32]. Here, it becomes clear that an embodied agent leads to more positive social interactions compared to a solely language-based system [16].

Thus, while ECAs 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 [11]. 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 [12].

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 "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. 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

This methodological approach is a response to the question of how a suitable theoretical and methodological approach can be found for designing and understanding conversational user interfaces. In the following, we present a methodological approach that allows observing not only the user's perceptions of interaction but the interactive behavior as well. 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 (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

The methodological approach uses three different survey procedures: First, we employ quantitative instruments for capturing the users' experiential perceptions. These tools quantify co-presence with virtual agents and attributions of personality towards the agent. Second, we use videographic methods for observing the user's and agent's bodily behavior during the VR intervention (see Figure 1 and Figure 2). These nonverbal interactions are recorded by mixed reality methods. Third, we conduct guideline-based qualitative interviews after the VR intervention.

This way, both the experiential and social aspects of the interaction should come into view: While the data obtained from the interviews and the questionnaires refer to the individual perception of the interaction, the videographic method captures the joint behavior of the user and the agent.

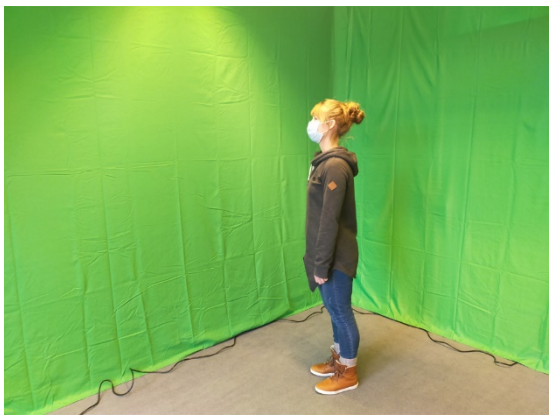

Figure 1: User standing in front of the green screen

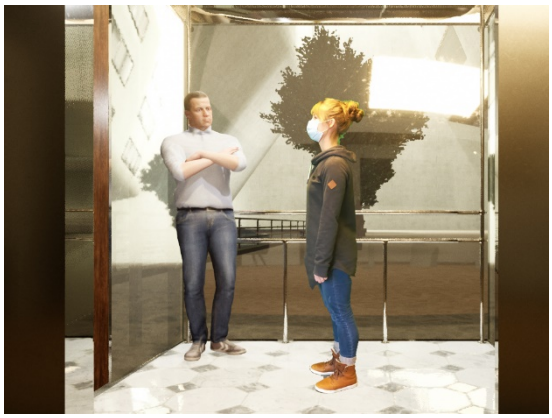

Figure 2: Mixed reality representation of user and virtual agent

3.1. Material

As technical equipment, we use a high-end VR-compatible PC (Geforce RTX 3080), a modern, high-resolution HMD (HP Reverb G2), as well as a green screen studio with a total area of about 16 m² and appropriate lighting. The ECA (see Figure 3) to be tested is currently being 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.

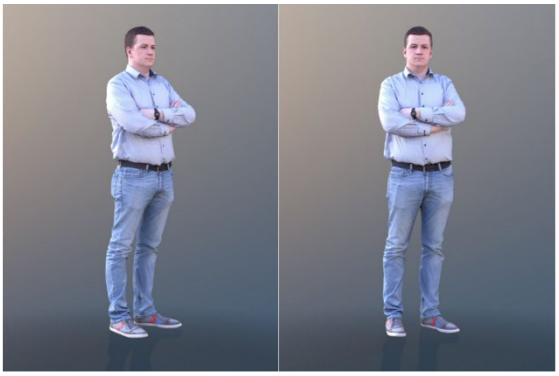

Figure 3: Static rendering of the virtual agent

The agent combines Natural Language Processing via Google DialogFlow with a graphically realistic appearance (photo-grammetric scans of a real person). 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.

3.2. Procedures

In a pilot study, we plan to test our agent following a three-step procedure. After clarifying formalities and collecting demographic data, the participants get used to the VR system. In VR, the agent conducts a questionnaire about the user's travel behavior and travel preferences. After this short survey, the participants should stay in VR for a few more minutes. During this time, they do not have an explicit task, but have the opportunity to communicate with the virtual agent at their own discretion. During all interactions, the participants are filmed by a video camera. After the VR experience, participants complete two questionnaires (regarding co-presence and attributions of humanness) and one interview about their experience, which is recorded with an audio recorder.

3.3. Analysis

The mixed methods approach of this mixed reality study allows the analysis of several data sets. On the one hand, the verbal interaction with the agent creates conversation protocols, which can be examined in more detail by means of conversational analysis. Furthermore, the videographic approach allows the examination of facial expressions and gestures and provides qualitative data on individual behavior. In the end, all types of qualitative data can be put into new perspective by comparing it with the quantitative data sets.

For the quantitative assessments, the first step is to carry out descriptive statistics, both with regard to the sample characteristics (age, gender, etc.) and the test values. Here, common distribution measures, such as central tendency, skewness and kurtosis coefficients, as well as descriptive correlations are identified in order to ensure representativeness. Subsequently, the actual data evaluation of the test values is carried out using inferential statistical tests. From the analysis of the interviews, we hope to gain more information on the users' orientations towards virtual agents: Which basic expectations towards conversational agents do the users show and which beliefs do they have about our prototype in particular? The videographic and interview data will be analyzed with the Documentary Method [3].

Central to the analysis of social behavior is the definition of evaluation categories. Thus, for analysis, the mixed reality recordings are transferred into a notation common for videography [18]. Here, we align with communication theory and conversational analysis and define various behavioral actions such as pauses, turn taking, overlapping, addressing, and other social mechanisms of face-to-face interactions like repairing failed sense-making [27].

4. Discussion

The main aspect of this methodology allows capturing interactions that happen nonverbally. In contrast to just verbalizing what is happening, this approach provides a more complete picture of human-agent interactions. With the help of mixed reality methods, the methodology is able to analyze the joint behavior of both user and agent. This multi-method approach enables the evaluation of user-agent interactions from multiple perspectives (see Table 2). First, users share their individual perceptions through quantitative and qualitative assessment. Second, we can observe the reaction parameters of the agent in regard to the actions of the user.

Third, we are able to observe the interaction itself in its mutual social performance.

Table 2
Methodological Advantages

Method	Advantage
Mixed Methods	Combination of questionnaires, interviews and videography.
Mixed Reality	Recording of the joint behavior of user and virtual agent.
Sociological extension	Social dimension of interaction as an emergent result comes to the fore.
Videography	Nonverbal communication can be analyzed.

This way, potential discrepancies between the actual behavior and the user's evaluation of the interaction can be brought into view, which could not be found with other methods. For example, it would be possible to find strongly rationalizing users, who describe the agent as not human-like, but behave towards it in a very human way [22]. Other users might provide socially desirable responses, but at the same time show pejorative, unfriendly, or dominant behavior towards the agent.

4.1. Limitations

Currently, the biggest limitation to this approach is the lack of testing experience and validation, as our studies have not yet been conducted (because of COVID-19). However, the method will be tested during a two-year empirical phase, which will also yield a large amount of quantifiable data.

However, concerning the evaluation of the methodology, the aspects of social control and social desirability come to the fore. Users will be observed in their full behavior by the experimenters and the camera, which will likely influence the behavior. Therefore, the question arises, how interaction would look like, when users are alone with the agent.

Beyond that, the further technical elaboration of virtual agent is much needed. Virtual agents need to include the ability to process nonverbal inputs from users. Only then, we can speak of an equally structured two-way interaction that does not pause at the illegibility of human nonverbal signals. Until then, virtual agents always refer only to the users' verbal expressions, while the user, on the other hand, processes both the agent's verbal and physical expressions in communication.

5. Conclusions

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.

6. Acknowledgements

The European Regional Development Fund (EFRE) funded this research. The study received a positive vote from the Ethics Committee of Witten/Herdecke University. We would like to thank our project partners HHVision for the visual animation of the virtual body, as well as IOX for the creation of the Bot Management System. Furthermore, we would like to thank Sophia Bermond and Werner Vogd for their valuable support.

7. 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] R. Bohnsack, Documentary method and group discussions, in: R. Bohnsack, N.

Pfaff, and W. Weller (Eds.), Qualitative Analysis and Documentary Method in International Educational Research, Opladen: Barbara Budrich, 2010, pp-99-124. doi: 10.3224/86649236

[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, Embodied conversational agents, MIT press, 2000.

[6] K. Corti, A. Gillespie, A truly human interface: interacting face-to-face with someone whose words are determined by a computer program, Frontiers in psychology 634 (2015). doi: 10.3389/fpsyg.2015.00634

[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] J. Ehret, J. Stienen, C. Brozdowski, A. Bönsch, I. Mittelberg, M. Vorländer, T. Kuhlen, Evaluating the Influence of Phoneme-Dependent Dynamic Speaker Directivity of Embodied Conversational Agents' Speech, in: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents, ACM Press, New York, NY, 2020, pp 1-8. doi: 10.1145/3383652.3423863.

[10] M. Eskenazi, S. Mehri, E.Razumovskaiam T. Zhao, Beyond turing: Intelligent agents centered on the user, arXiv preprint (2019). doi: arXiv:1901.06613.

[11] 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

[12] 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

[13] 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

[14] L. Gong, How social is social responses to computers? The function of the degree of anthropomorphism in computer representations. Computers in Human Behavior 24.4 (2008) 1494-1509. doi: 10.1016/j.chb.2007.05.007

[15] H.-H. Huang, Embodied conversational agents, in: K. L. Norman & J. Kirakowski (Eds.), The Wiley handbook of human computer interaction, Wiley Blackwell, 2018, pp. 601–614. doi: 10.1002/9781118976005.ch26.

[16] 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

[17] M. L. Knapp, J. A. Hall and T. G. Horgan. Nonverbal communication in human interaction. Cengage Learning, 2013.

[18] H. Knoblauch, R. Tuma. Videography: an interpretive approach to video-recorded micro-social interaction. In: Eric Margolis and Luc Pauwels (Eds.). The Sage Handbook of Visual Methods. Thousand Oaks, CA: Sage, 2011, pp. 414–430.

[19] M. Koschate, R. Potter, P. Bremner, M. Levine, Overcoming the uncanny valley: Displays of emotions reduce the

uncanniness of humanlike robots. in: 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE, 2016, pp. 359-366. doi: 10.1109/HRI.2016.7451773.
[20] MICA. Magic Leap's Mica at GDC. *2019*. URL: https://www.youtube.com/watch?v=-PzeWxtOGzQ.
[21] M. Mori, K. F. MacDorman, N. Kageki, The Uncanny Valley, IEEE Robotics & Automation Magazine, 19 (2012) 98–100. doi: 10.1109/MRA.2012.2192811.
[22] C. Nass, Y. Moon, Machines and mindlessness: Social responses to computers, Journal of social issues 56. 1 (2000) 81-103. doi: 10.1111/0022-4537.00153.
[23] C. Regenbogen, D. A. Schneider, R. E. Gur, F. Schneider, U. Habel, T. Kellermann, Multimodal human communication—targeting facial expressions, speech content and prosody. Neuroimage, 60.4 (2012) 2346-2356. doi: 10.1016/j.neuroimage.2012.02.043.
[24] J. Reinhardt, L. Hillen, K. Wolf. Embedding Conversational Agents into AR: Invisible or with a Realistic Human Body? in: Proceedings of the Fourteenth International Conference on Tangible, Embedded, and Embodied Interaction, ACM Press, New York, NY, 2020, pp. 299-310. doi: 10.1145/3374920.3374956.
[25] P. Sajjadi, L. Hoffmann, P. Cimiano, S. Kopp, A personality-based emotional model for embodied conversational agents: Effects on perceived social presence and game experience of users, Entertainment Computing, 32 (2019) 100-113. doi: 10.1016/j.entcom.2019.100313.
[26] 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.
[27] J. Sidnell, Conversation analysis. An introduction. Chichester: Wiley-Blackwell, 2010.
[28] I. Wang, J. Smith, J. Ruiz, Exploring virtual agents for augmented reality. in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, ACM Press, New York, NY, 2019, pp. 1-12, doi:10.1145/3290605.3300511.
[29] 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.
[30] J. N. Weinstein, Artificial Intelligence: Have You Met Your New Friends; Siri, Cortana, Alexa, Dot, Spot, and Puck, Spine, 44.1 (2019), 1-4. doi: 10.1097/BRS.0000000000002913.
[31] 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.
[32] N. Yee, J. N. Bailenson, K. Rickertsen, A meta-analysis of the impact of the inclusion and realism of human-like faces on user experiences in interfaces. in: Proceedings of the SIGCHI conference on Human factors in computing systems, ACM Press, New York, NY, 2007, pp. 1-10. doi: 10.1145/1240624.1240626