Design of Seamless Multi-modal Interaction Framework for Intelligent Virtual Agents in Wearable Mixed Reality Environment

Ghazanfar Ali
Division of NT-IT
University of Science and Technology
South Korea
aliust@ust.ac.kr

Hong-Quan Le
Letsee
South Korea
quan.le1926@gmail.com

Junho Kim
College of Computer Science
Kookmin University
South Korea
junho@kookmin.ac.kr

Seung-Won Hwang
Department of Computer Science
Yonsei University
South Korea
seungwonh@yonsei.ac.kr

Jae-In Hwang*
Imaging Media Research Center
Korea Institute of Science and Technology
South Korea
hji@kist.re.kr

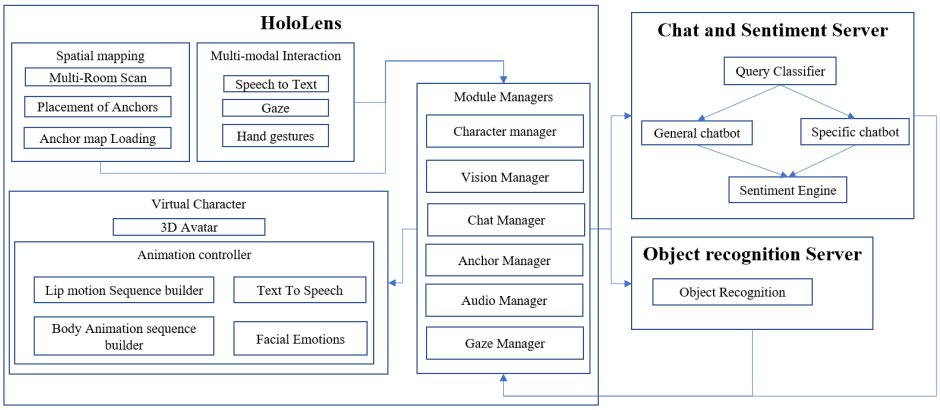

Figure 1: Overall framework for an intelligent virtual agent in wearable mixed reality.

ABSTRACT

In this paper, we present the design of a multimodal interaction framework for intelligent virtual agents in wearable mixed reality environments, especially for interactive applications at museums, botanical gardens, and similar places. These places need engaging and no-repetitive digital content delivery to maximize user involvement. An intelligent virtual agent is a promising mode for both purposes. Premises of framework is wearable mixed reality provided by MR devices supporting spatial mapping. We envisioned a seamless interaction framework by integrating potential features of spatial mapping, virtual character animations, speech recognition, gazing, domain-specific chatbot and object recognition to enhance virtual experiences and communication between users and virtual agents. By applying a modular approach and deploying computationally intensive modules on cloud-platform, we achieved a seamless virtual experience in a device with limited resources. Human-like gaze and speech interaction with a virtual agent made it more interactive. Automated mapping of body animations with the content of a speech made it more engaging. In our tests, the virtual agents responded within 2-4 seconds after the user query. The strength of the framework is flexibility and adaptability. It can be adapted to any wearable MR device supporting spatial mapping.

CCS CONCEPTS

• **Computing methodologies** → **Mixed / augmented reality**.

ACM Reference Format:
Ghazanfar Ali, Hong-Quan Le, Junho Kim, Seung-Won Hwang, and Jae-In Hwang. 2019. Design of Seamless Multi-modal Interaction Framework for Intelligent Virtual Agents in Wearable Mixed Reality Environment . In *Computer Animation and Social Agents (CASA '19), July 1–3, 2019, PARIS, France.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3328756.3328758

*Corresponding author

CASA '19, July 1–3, 2019, PARIS, France

ACM ISBN 978-1-4503-7159-9/19/07…$15.00
https://doi.org/10.1145/3328756.3328758

1 INTRODUCTION

Mixed reality (MR) is a technique that blends the real and virtual worlds to create new environments where users can interact in real time. The ability to overlay the computer graphics images onto the real world with seamless interaction between virtual and physical worlds make MR technology suitable for wider ranges of the application as defined by Liu et al. [12]. Presently, current-generation commercial mixed reality devices, such as Microsoft HoloLens, Magic Leap, and Meta 2, drive new methods of user interactions via hand gestures, voice, and gazing argues Xiao et al. [20]. For decades, Agent-based systems have gained researcher's attention because of their potential for diverse applications as described by McArthur et al. [14], McArthur et al. [15], Catterson et al. [8], Castano et al. [7] and Alencar et al. [1]. In the paper of Xie et al. [21], significant features of agent-based systems were defined to emphasize its advantages. Anabuki et al. [2] introduced a new type of anthropomorphic agent that lives in a 3D space where the real and virtual worlds are seamlessly merged. In research of Kopp et al. [11], a multimodal virtual humanoid agent assistant was created to have face-to-face discussions with users in a virtual environment. Ideally, the agent should mimic human behavior. Work of Cai et al. [6] models human behavior for virtual agents. Machidon et al. [13] provide a comprehensive review of the current state of the art in the use of virtual humans in cultural heritage ICT applications. Vosinakis et al. [19] argue that virtual humans are more useful in the dissemination of intangible cultural heritage in virtual environments. Today, many studies like Arroyo-Palacios et al. [3] and Bogdanovych et al. [5] focus on making a believable character or agent partially show emotions and behavioral abilities when interacting with the environment or users. Unfortunately, this does not seem to be enough to enhance the interactive user experience. Richards [17] presented a very detailed current study regarding agent-based architecture and intelligent virtual agent issues. However, most of these solutions are toward PC based systems and suffer from repetitive content delivery. The current generation of Wearable MR devices provides a more comfortable and engaging experience. It can further be enhanced by artistically crafted 3D virtual human, intelligent chatbot, object recognition, and intuitive interaction technique. Accordingly, our objective is the multimodal interaction of more than one agent in an MR environment, especially for museums and botanical gardens. We propose a framework which includes virtual character, multi-modal interaction, domain-specific chatbot, and object recognition system. The primary goal of this framework is to provide a seamless, interactive, engaging, and non-repetitive experience. The major strengths of our framework are its flexibility and adaptability as we utilized and modular approach. This framework gives the user the ability to experience the real world with digital information being delivered through a friendly virtual human. This gives the feeling of having a personal tour guide to enrich one's experience. Our contributions are:

- Design and implementation of the virtual agent framework for wearable MR.
- Interaction design and implementation on the intelligent virtual agent
- Expression and body gesture mapping of intelligent virtual agent
- Scalable anchor system for wearable MR

The framework is explained in section 2. In section 3, we briefly introduce the development platform. In section 4, a scenario is presented and discussed.

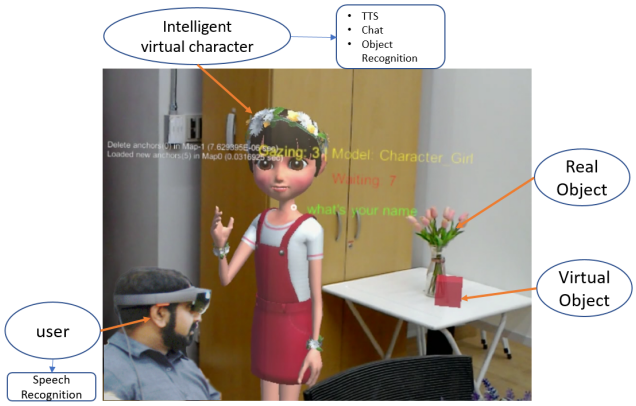

Figure 2: Scenario usage of the framework.

2 SEAMLESS MR SYSTEM ARCHITECTURE FOR INTELLIGENT VIRTUAL AGENT

To build a useful and meaningful system by integrating all required components, requirements in the wearable mixed reality system are:

- **Interaction:** Virtual agent which recognize user gazing objects and converse about it.
- **Expression:** Virtual agent which shows emotion and body gesture which is appropriate to the conversation.
- **Scalability:** Virtual agent who shows up at an infinite working area.

The overall architecture is shown in Figure 3.

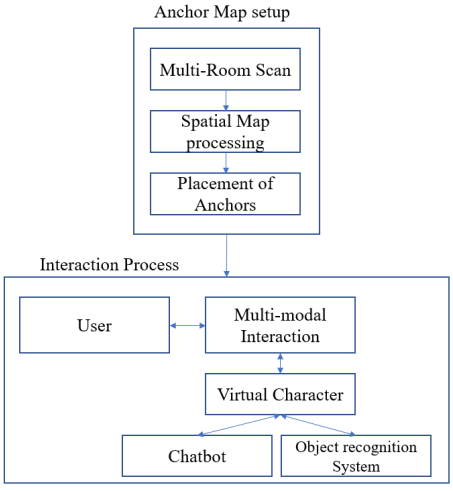

Figure 3: Overall architecture.

2.1 Interaction: Gazing object recognition and conversation.

The interaction process is aimed at providing a more natural method to interact with the virtual character. The major problem in the process was to figure out the way to ask the virtual character about any object for the first time. The problem arises when the user needs to ask about some object. If the user gazes at the character and asks about the object, the user must quickly look at the object to capture it. This sometimes does not give enough time to react. This problem was solved by including limited voice commands, i.e., "What is this," "Tell me about this" etc. in our framework. Figure 4 shows the interaction process with and without anchors in sight.

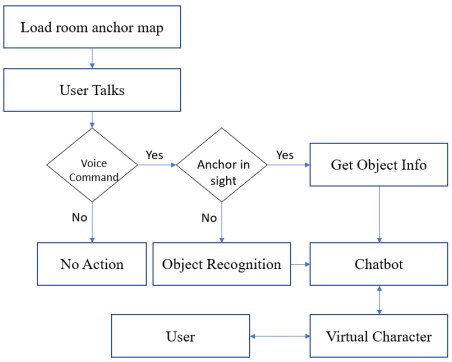

Figure 4: Process to get recognized object information

2.1.1 Gaze and speech based multimodal interaction: There has been plenty of research about creating natural methods for human-agent interaction. For example, this study by Freigang et al. [9] attempted to emphasize the role of speech and hand gestures in conveying pragmatic information between users and virtual humans. Arroyo-Palacios et al. [3] proposed an approach for interactive virtual characters for MR applications. Gaze and gestures were applied by users so that they could feel comfortable with a natural way to interact with agents. According to Kolkmeier et al. [10], gaze and proxemics behaviors can establish and maintain a level of intimacy, not only for human-human interaction but also for human-agent interaction. Results indicated that agents exhibiting higher proximity caused participants to step away from more than agents with low proximity. In a study by Yumak [22], a gaze behavior model was presented for an interactive virtual character that was situated in the real world. Uchino et al. [18] and Rebman et al. [16] showed the importance of speech recognition in interaction. The study of Avramova et al. [4] presented a toolkit with full-body animation for the virtual character, which offered MR interaction. We utilized gaze and speech recognition to interact with the virtual character. HoloLens provides the API for gaze and speech recognition. HoloLens Toolkit provides high-level functions of speech recognition like dictation manager, which enables us to say long sentences. Our method, as shown in Figure 5, does not require any trigger keyword. When the user gaze on the virtual character for 4 seconds, speech recognizer starts. If the user talks, then it will wait to finish user talking and then send the query to a chatbot. If the user does not speak then, the virtual character will try to start

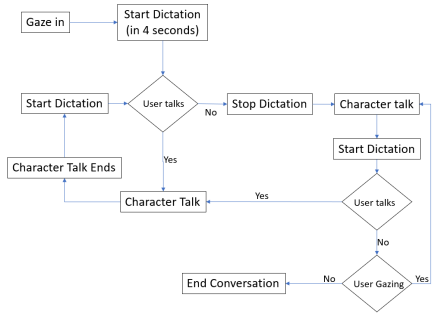

Figure 5: Gaze and Speech interaction architecture

the conversation by saying "Hello, do you need help?" or similar sentences. If the user gaze out while speech recognizer is running, and the user is silent, then the conversation will end.

2.1.2 Gazed object recognition: Object recognition system works as the eyes of the virtual character. It enables the virtual character to be aware of its soundings. Any state of the art or custom trained object recognition or image classification system can be used if it can handle web requests. Vision manager module handles object recognition related task. It only needs a valid API Endpoint. The user can make their API Endpoint using any available system or use existing API Endpoint. Vision manager is flexible to handle as it can also be connected to the on-device system. By design, this can cover any number of objects if chatbot has the required knowledge base.

2.1.3 Conversation of intelligent virtual agent: For a virtual character to interact with the user, it needs a knowledge base and associated program to give out the required information. Those programs are called chatbots. Chatbots are playing a vital role in Human-Computer Interactions. Chatbots are also being used to automate customer services. In our system, chatbot serves two purposes, i.e., 1) It provides domain-specific information. 2) General conversation options. Our chatbot is equipped with a sentiment engine which analyses the response and attaches a sentiment class and level of sentiment to it as shown in Figure 6. This chatbot is deployed on the cloud and accessed through web requests. Input parameter to this chatbot are "(*Query, Object*)." *Query* can be a question or comment or an answer to virtual character's question. *Object* is any virtual or real object detected through object recognition about which we want to ask a question. *Query* can be a general query like asking about the character itself or can be a specific query about the objects around you. Prerequisite for a specific query about any object is that the user must look at the object for the first time and ask about it. This will trigger the object recognition component and will pass the information to a chatbot. Follow up queries don't require to look at the object. Reply to the *Query* is "(*Reply, Sentiment Class, Sentiment Level*)." Sentiment class and the level are used for appropriately animating the character. The user can ask both general or specific query in one conversation and any order. This style of conversation mimics human-human interaction. This increases engagement and improves the quality of experience.

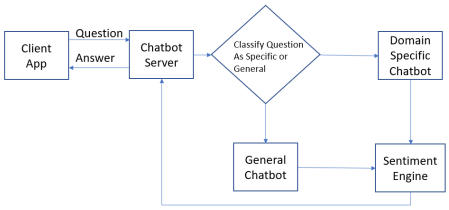

Figure 6: Chatbot Architecture

2.2 Expression: Virtual character emotion and body gesture

Our virtual human is a 3D model. To make a realistic virtual human, it needs to have rich body animations and facial emotions. Body animations are for verbal and nonverbal activities. As virtual human is required to move around and respond to verbal cues, it needs to have a large set of animations available. In our system, we have 30 different animations. Facial emotions are for the expression of different emotions. In our system, we have four emotions, i.e., Joy, Angry, Sad, Fear. Each emotion has three levels, i.e., High, Medium, Low, as shown in Figure 7.

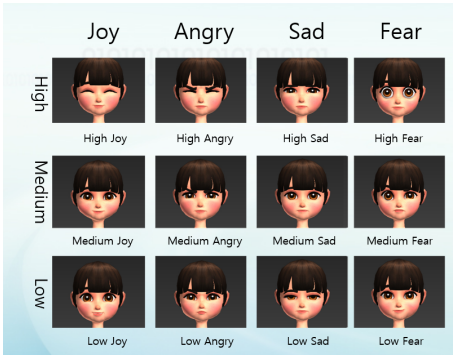

Figure 7: Facial emotions types and levels

Body animations and facial emotions are applied based on the content of the speech. Speech is generated from text using Text-to-Speech system. The architecture for building animation sequence is shown in Figure 8. This architecture requires a mapping table for mapping of text to animations. Human experts create this map. It has an animation for phrases and individual words which are utilized by the animation builder, as shown in Figure 8. Facial emotions are predicted from the text by sentiment engine integrated into chatbot and applied for the complete duration of the text. Speech from the text is lip-synced by converting text to phonemes sequence. Phonemes map is used for this purpose. Lip shape for each phoneme is shown in Figure 9. Speech, body animation, facial emotions are applied in parallel, as shown in Figure 10.

2.3 Scalability: Anchor map setup

HoloLens provides a mechanism known as anchors through which static objects can be tracked. By combining image recognition and anchors, we created a process by which we can map infinite working area. This process can save information about real objects which

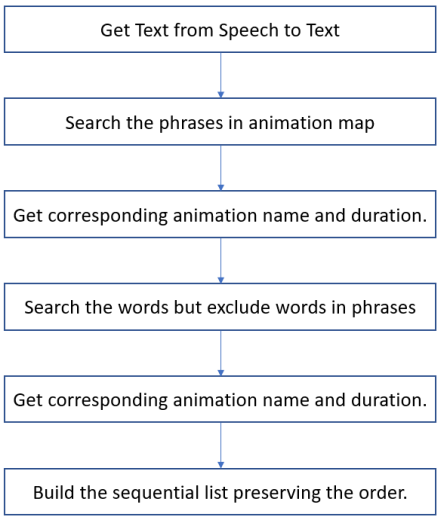

Figure 8: Architecture for building body-animation sequence

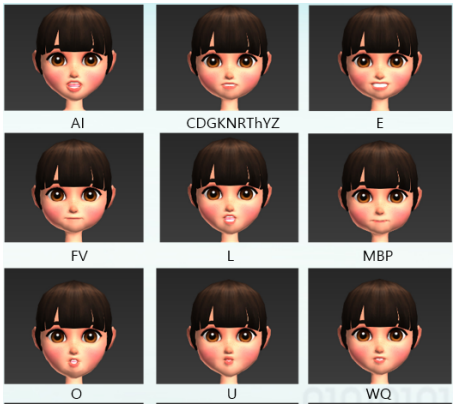

Figure 9: Lip shape for phonemes

can be accessed by the virtual agent. This process minimizes the use of object recognition for each object. The anchor placement process is shown in Figure 11 a). Initial two steps are part of the framework's anchor manager. Since there is no mechanism to load room-specific anchors in HoloLens and for identical rooms wrong anchors may be loaded. To avoid this issue, anchor management process utilizes object recognition to load room-specific anchors. The process is shown in Figure 11 b).

3 DEVELOPMENT ENVIRONMENT

The platform is consisting of Unity3D, Microsoft HoloLens, HoloLens Toolkit, and cloud computing platform. HoloLens is spatial MR holographic device developed by Microsoft. With HoloLens, we scan the room to create a 3D space through which user, virtual objects, and virtual character are tracked. HoloLens supports Unity3D as a development platform. HoloLens toolkit provides an easy way of setting up initial project and settings. It also offers numerous

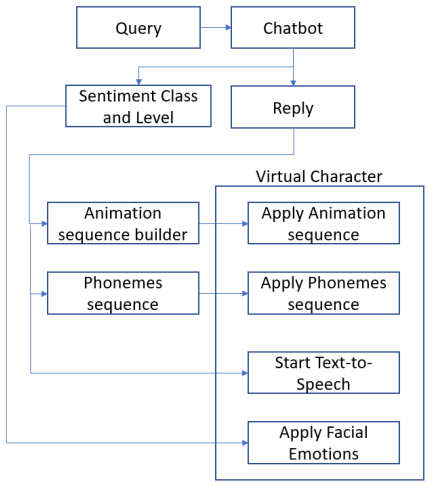

Figure 10: Applying speech, animation and facial emotions on the character

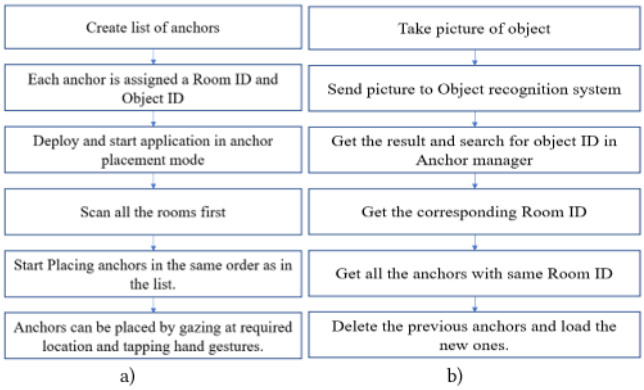

Figure 11: a) Anchor placement process. b) Anchor loading

helping functions with regards to mapping, gaze tracking, and speech recognition. HoloLens is a self-contained device with limited resources. It is not possible to efficiently run chatbot and object recognition from the device without compromising the quality of user experience. To solve this, our framework access cloud-based chatbot and object recognition system via module manager.

4 SCENARIO: A BOTANICAL GARDEN

The framework was utilized to create a demo system. The scenario was a tour at a Botanical garden. In this scenario, the user can interact with the virtual character to ask about different flowers. This scenario needed the following items:

- 3D humanoid model
- Flower dataset for the object recognition system
- Knowledgebase of flowers for chatbot

3D artists created 3D Models. It included body design, clothes, and facial emotions. Body animations were acquired from Mixamo[https://www.mixamo.com]. There were three models, i.e., Girl, Boy, and Old Man. Different voices from Text-To-Speech systems were assigned to them. Body animation also varies for each character. No customization was required in the framework to add multiple characters in one application. Categories of the flower were nine. We used artificial flower models. A dataset of images was prepared from those models. Our object recognition system was trained and deployed on custom vision API of Microsoft Azure. We trained the system using our dataset. Knowledgebase for chatbot was built with extensive research. The goal for chatbot was to include as much information as possible so that the user should not be forced to ask only specific questions.

Figure 12: Models

Figure 13: Nine flowers used in the scenario.

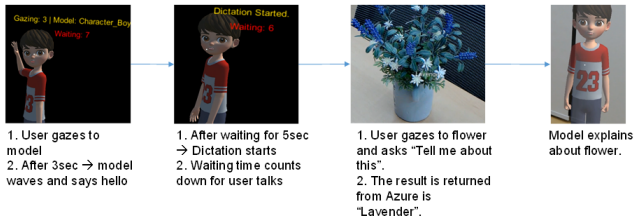

Figure 14: Interaction with model

For this scenario, two rooms were set up with five flowers in 1 room and 4 in 1 room. Scanning was performed, and anchors were placed. Girl model was assigned room 1, and Boy model was assigned room 2. Then we successfully demonstrated the framework. The outcome was a seamless interaction with response time ranging between 2-4 seconds. This time includes speech-to-text conversion, sending text query to a chatbot, receiving a response, and building animation list. This response time, however, does not include the time of silence to detect the end of user speech. The response time will be higher if object recognition is also triggered

Table 1: Total average response time for each query

Query	Query A	Query B	Query C
Total queries	30	30	30
Object recognition	5.4s	n/a	5.3s
Chatbot	n/a	2.1s	2.0s
Processing time	0.5s	1s	1s
Total Time	5.9s	3.1s	8.3s
Std Dev	0.99	0.98	0.99

typically in the range of 5-8 seconds total. As this delay is more significant, so during this time, our character shows the animation of thinking about the object and say something like "let me see, let me think about it." This way, the user does not feel a time delay. As our modules were deployed on cloud-platform and access through the internet, so a significant portion of the time was utilized in web requests. In Table 1, we summarize the average results for three queries. Query A was loading of anchor map, and it required object recognition. Query B was general conversation. Query C was about flower, and it needed object recognition. Above table proves the usability of anchor map to reduce response time.

5 CONCLUSIONS

We studied and conceptualized a framework by integrating spatial mapping, virtual character, multi-modal interaction, chatbot, and object recognition for wearable mixed reality. We successfully demonstrated that such a framework could reduce the time to develop an augmented reality environment with an intelligent virtual agent. It can prove to be more engaging and entertaining. The approach presented both theoretical and experimental aspects to emphasize the importance of using multimodal interactions for humans and agents. The potential of spatial mapping, virtual character, multi-modal interaction, chatbot, and object recognition was effectively integrated to enhance the human-agent interaction and to stimulate the realism of the mixed reality environment.

ACKNOWLEDGMENTS

This research is supported by the Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research and Development Program and KIST Flagship Project.

REFERENCES

[1] M. Alencar and J. Netto. 2011. *Improving cooperation in Virtual Learning Environments using multi-agent systems and AIML.* Frontiers in Education Conference (FIE). Rapid City. https://doi.org/10.1109/FIE.2011.614302

[2] M. Anabuki, H. Kakuta, H. Yamamoto, and H. Tamura. 2000. Welbo: an embodied conversational agent living in mixed reality space. *In CHI EA '00 CHI '00 Extended Abstracts on Human Factors in Computing Systems (CHI '00). The Hague, The Netherlands* (2000), 10–11. https://doi.org/10.1145/633292.633299

[3] J. Arroyo-Palacios and R. Marks. 2017. POSTER Believable Virtual Characters for Mixed Reality. 2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct). *Nantes* (2017), 121–123. https://doi.org/10.1109/ISMAR-Adjunct.2017.45

[4] V. Avramova, F. Yang, C. Li, C. Peters, and G. Skantze. 2017. A Virtual Poster Presenter Using Mixed Reality. *Intelligent Virtual Agents* (2017), 25–28. https://doi.org/10.1007/978-3-319-67401-8_3

[5] A. Bogdanovych, T. Trescak, and S. Simoff. 2015. Formalising Believability and Building Believable Virtual Agents. *Lecture Notes in Computer Science* (2015), 142–156. https://doi.org/10.1007/978-3-319-14803-8_11

[6] L. Cai, B. Liu, J. Yu, and J. Zhang. 2015. Human behaviors modeling in multi-agent virtual environment. *Multimedia Tools and Applications* 76, 4 (2015), 5851–5871. https://doi.org/10.1007/s11042-015-2547-z

[7] R. Castano, G. Dotto, R. Suma, A. Martina, and A. Bottino. 2015. VIRTUAL-ME: A Library for Smart Autonomous Agents in Multiple Virtual Environments. *Communications in Computer and Information Science* (2015), 34–45. https://doi.org/10.1007/978-3-662-46241-6_4

[8] V. Catterson, E. Davidson, and S. McArthur. 2012. Practical applications of multi-agent systems in electric power systems. *European Transactions on Electrical Power* 22, 2 (2012), 235–252. https://doi.org/235--252.10.1002/etep.619

[9] F. Freigang and S. Kopp. 2016. This Is What's Important – Using Speech and Gesture to Create Focus in Multimodal Utterance. *Intelligent Virtual Agents* (2016), 96–109. https://doi.org/10.1007/978-3-319-47665-0_9

[10] J. Kolkmeier, J. Vroon, and D. Heylen. [n.d.]. Interacting with Virtual Agents in Shared Space: Single and Joint Effects of Gaze and Proxemics. *Intelligent Virtual Agents* ([n. d.]), 1–14. https://doi.org/10.1007/978-3-319-47665-0_1

[11] S. Kopp and et al. [n.d.]. Max - A multimodal assistant in virtual reality construction. *KI - K u nstliche Intelligenz* 4 ([n. d.]), 11–17.

[12] W. Liu, O. Fernando, A. Cheok, J. Wijesena, and R. Tan. 2007. Science Museum Mixed Reality Digital Media Exhibitions for Children. *Second Workshop on Digital Media and its Application in Museum and Heritages (DMAMH* 2007 (2007), 389–394. https://doi.org/10.1109/DMAMH.2007.61

[13] O. M. Machidon, M. Duguleana, and M. Carrozzino. 2018. Virtual humans in cultural heritage ICT applications: A review. *Journal of Cultural Heritage* 33 (2018), 249–260. https://doi.org/10.1016/j.culher.2018.01.007

[14] S. McArthur, E. Davidson, V. Catterson, and et al. 2007. Multi-agent systems for power engineering applications âĂŞ part I: concepts, approaches, and technical challenges. *IEEE Transactions on Power Systems* 22, 4 (2007), 1743–1752. https://doi.org/1743--1752.10.1109/TPWRS.2007.908471

[15] S. McArthur, E. Davidson, V. Catterson, and et al. 2007. Multi-agent systems for power engineering applications âĂŞ part II: technologies, standards, and tools for building multi-agent systems. *IEEE Transactions on Power Systems* 22, 4 (2007), 1753–1759. https://doi.org/1753--1759.10.1109/TPWRS.2007.908472

[16] C. Rebman, M. Aiken, and C. Cegielski. 2003. Speech recognition in the human–computer interface. *Information and Management* 40, 6 (2003), 509–519. https://doi.org/10.1016/s0378-7206(02)00067-8

[17] D. Richards. 2012. Agent-based museum and tour guides. (2012).

[18] S. Uchino, N. Abe, K. Tanaka, T. Yagi, H. Taki, and S. He. 2007. VR Interaction in Real-Time between Avatar with Voice and Gesture Recognition System. *Advanced Information Networking and Applications Workshops (AINAW '0* 7 (2007), 959–964. https://doi.org/10.1109/AINAW.2007.370

[19] S. Vosinakis, N. Avradinis, and P. Koutsabasis. 2017. Dissemination of Intangible Cultural Heritage Using a Multi-agent Virtual World. *Lecture Notes in Computer Science* (2017), 197–207. https://doi.org/10.1007/978-3-319-75789-6_14

[20] R. Xiao, J. Schwarz, N. Throm, A. Wilson, and H. Benko. 2018. MRTouch: Adding Touch Input to Head-Mounted Mixed Reality. *IEEE Transactions on Visualization and Computer Graphics* 24, 4 (2018), 2018. https://doi.org/10.1109/TVCG.2018.2794222

[21] J. Xie and C. Liu. 2017. Multi-agent systems and their applications. *Journal of International Council on Electrical Engineering* 7, 1 (2017), 188–197. https://doi.org/10.1080/22348972.2017.1348890

[22] B. Yumak, Z. van den Brink and A. Egges. 2017. Autonomous Social Gaze Model for an Interactive Virtual Character in Real-Life Settings. *Computer Animation and Virtual Worlds* 28 (2017), 3–4. https://doi.org/10.1002/cav.1757