To appear at MobileHCI 2020. Authors' copy. Official publication at: https://doi.org/10.1145/3406324.3410537

Let's Go There: Voice and Pointing Together in VR

Jaisie Sin
University of Toronto, TAGlab and University of Toronto, Faculty of Information
Toronto, ON, Canada
js.sin@mail.utoronto.ca

Cosmin Munteanu
University of Toronto, TAGlab and University of Toronto Mississauga, ICCIT
Toronto, ON, Canada
cosmin@taglab.ca

ABSTRACT

Hand-tracking has been advertised as a natural means to engage with a virtual environment that also enhances the feeling of presence in and lowers the barriers to entry to virtual reality. We seek to explore combining hand-tracking with voice input (which is then processed with automatic speech recognition) for a novel multimodal experience. Thus, we created *Let's Go There*, which explores this joint-input method for four functions in virtual reality environments: positioning, object identification, information mapping, and disambiguation. This combination may serve as a more intuitive means for users to communicate and navigate in virtual environments. We expect there to be multiple potential applications of this multimodal form of interaction across numerous domains including training, education, teamwork, and games. *Let's Go There*, the system described in this paper, was first accepted at CUI 2020, however we also believe there is value in showcasing it at MobileHCI.

CHI'19 Extended Abstracts, May 4-9, 2019, Glasgow, Scotland, UK.

ACM ISBN 978-1-4503-5971-9/19/05.
DOI: https://doi.org/10.1145/3290607.3312941

To appear at MobileHCI 2020. Authors’ copy. Official publication at: https://doi.org/10.1145/3406324.3410537

KEYWORDS

Virtual Reality; Pointing; Voice Technology; Voice User Interface; Virtual Agents; Technology Adoption

INTRODUCTION

Hand-tracking allows users to engage with a virtual environment with their own hands, rather than the more traditional method of using accompanying controllers in order to operate the device they are using and interact with the virtual world. Hand-tracking has been advertised as a more natural way to engage with virtual reality (VR) by enhancing one’s presence in the virtual space [9]. In turn, hand-tracking is considered to have a reduced barrier to entry [8]. However, the evidence for this claim and uses of hand-tracking, especially in combination with voice input, are still underexplored. Thus, we want to investigate:

1. For what functions might voice input enhance the naturalness of VR when using hand-tracking?
2. What are potential ways to implement voice and hand-tracking in VR so they work in tandem?

As such, we created *Let’s Go There*, which is an Oculus Quest app designed to explore the value of combining voice and hand-tracking as a joint-input method. With *Let’s Go There*, we explore the application of this joint-input method for four functions: positioning, object identification, information mapping, and disambiguation. Previous research [2] has explored some variations on the arrangement exemplified by *Let’s Go There*, such as navigation in a VR space using voice commands [5], the use of both voice and gestures as input [5, 6], and the multimodal interaction with virtual agents [1]. However, little to no work exists yet on the use of voice combined with pointing in VR.

Previous Demonstrations of This System

Let’s Go There has previously been accepted for CUI 2020 [11], and will be demonstrated at CUI 2021 conference [4]. The Oculus Quest is a truly mobile VR platform. This VR headset does not need to be connected to a power source or a computer in order to function. This mobile version of a VR headset allows the user great freedom of movement and mobility. Thus, we believe there is great value in showcasing *Let’s Go There* to the MobileHCI audience. In effect, MobileHCI will be the first international venue to showcase *Let’s Go There*.

THE FOUR FUNCTIONS IN LET’S GO THERE

Let’s Go There allows users to communicate in VR through two modalities in tandem: hand pointing and voice. Hand-tracking is used to identify objects and destinations of interest at which users are pointing. These objects and destinations that can be pointed at are preprogrammed by the developer. Voice input is mediated via automatic speech recognition. When the user’s hand points at a particular object or destination and specific keywords and phrases are used in the speech, an event is triggered accordingly. The following subsections describes four functions (positioning, object identification, information mapping, and disambiguation) of the multimodal hand pointing and voice input explored in *Let’s Go There*.

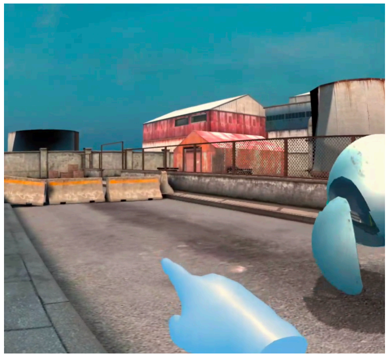

Figure 1: The user points at the barriers and utters “You go to the barriers,” and the virtual agent behaves accordingly by walking there.

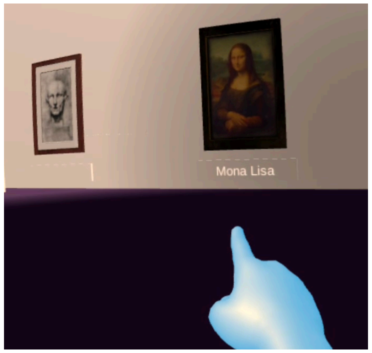

Figure 2: The user points at the Mona Lisa painting and utters “That is the Mona Lisa” to give it the “Mona Lisa” label.

Function 1: Positioning

Voice and pointing can be used as a means of directing relocation and locomotion in a VR environment. In *Let’s Go There*, when the user points at a potential pre-programmed destination and utters specific phrases such as “I’ll go to the X” (with X being a destination, e.g. “barriers”), locomotion will commence and their camera will automatically begin to move them towards the destination. If the user instead speaks a phrase that is meant to command something else, the other object will relocate instead. For example, speaking “You go to the barriers” can make an accompanying virtual agent position itself beside the barriers instead (Figure 1). In this case, pointing directs attention to the target destination and the voice input triggers the relocation. The non-verbal component of pointing and the verbal component of voice each play a role in the successful execution of the positioning.

Function 2: Object Identification

Pointing can be used to identify an object being referenced without specifically uttering the name of the object. *Let’s Go There* demonstrates this by supporting ambiguous words such as “here/there” or “this/that one.” For example, in the aforementioned example from 2.1 of positioning oneself or a virtual agent beside some “barriers,” the user can utter a phrase like “I’ll go there” or “You go there” to achieve the same result. In this case, the user does not need to mention the target destination of “barriers” by name. In this case, the pointing gestures identifies the target object and the voice triggers the action. Once again, the non-verbal component of pointing and the verbal component of voice each play a role in the successful execution of object identification.

Function 3: Information Mapping

The joint-input of voice and pointing can be used to assign information to objects inside the VR environment. In the case of *Let’s Go There*, this allows the user to label objects in the environment. the painting being referenced and the voice input provides information to assign this painting as the “Mona Lisa” (see Figure 2). Such an action of information mapping would not be possible, or at least not as natural, if either of the input modalities of pointing or voice was missing.

Function 4: Disambiguation

When two objects in view can fit a given physical verbal description, pointing can be used to identify the specific object that is being referenced. *Let’s Go There* demonstrates this in its “museum” level in which there are several vases in view of the user. When two vases are in two separate locations, pointing at one of them and uttering “that vase” can successfully isolate the target object. In this case, pointing clarifies the ambiguity that is introduced by the utterance. Conversely, if the two vases are both in front of the user, the user can utter “the farthest vase” to indicate they are referring to the vase that is located the farthest away (see Figure 3). Here, the use of voice clarifies the ambiguity that is made by the pointing. In both of these cases, pointing and voice complement each other to successfully disambiguate and provide clarity in communication.

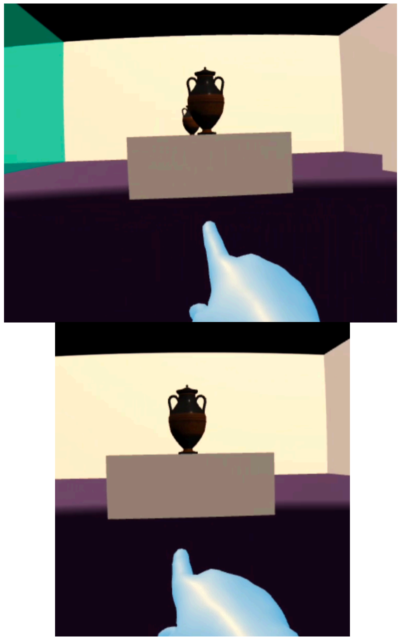

Figure 3: The user points in the direction of the two vases and utters "Make the farthest vase disappear," which makes the vase at the back vanish. The closer vase remains unchanged.

IMPLEMENTATION

Let's Go There is built as an app (as an Android application package, or APK) for the Oculus Quest. The Oculus Quest is a fully standalone VR headset that runs an Android operating system. *Let's Go There* has three primary components: the VR environment, the hand-tracking, and the voice recognition. The VR environment was built using the Unity Game Engine. The environment assets and the virtual agent used by *Let's Go There* [8, 9] are from third-party sources found in the Unity Asset Store. The hand-tracking feature is implemented through the Oculus software development kit for hand-tracking and has only been recently (December 2019) released for developer use. Automatic speech recognition is performed by sending the Quest's microphone input to Microsoft Azure's Cognitive Services. *Let's Go There* logs the transcript (i.e. the voice commands as text, with timestamps) and the directions and landmarks users point to for later data analysis.

CONCLUSIONS AND FUTURE WORK

The use of voice and pointing together can be a means to interact with VR environments in a natural manner. Moreover, each input modality grants critical information for the successful execution of events; the event would not work as well or at all if either of the two input modalities were missing. *Let's Go There* demonstrated this principle for four functions in VR environments: positioning, object identification, information mapping, and disambiguation.

Formal usability testing of *Let's Go There* is planned to evaluate the intuitiveness of the current configuration for navigating oneself and directing virtual agents though the VR environment. We also intend to explore more meaningful or "real-life" applications of the four functions. Some applications come to mind readily: issuing relocation commands to a battalion (positioning); navigation in a virtual museum (object identification); scribing text to a sticky note in a virtual workplace (teamwork); and, identifying between chess pieces of identical class on a virtual reality game board (disambiguation). We also plan on expanding the repertoire of interactions, such as through the application of casting magical spells in virtual reality (incantation using voice combined with a relevant hand gesture or movement).

REFERENCES

[1] 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. *Proceedings of the 32nd International Conference on Computer Animation and Social Agents*, Association for Computing Machinery, 47–52.

[2] Mark Billinghurst. 2013. Hands and speech in space: multimodal interaction with augmented reality interfaces. *Proceedings of the 15th ACM on International conference on multimodal interaction*, Association for Computing Machinery, 379–380.

[3] Richard A. Bolt. 1980. "Put-that-there": Voice and gesture at the graphics interface. *Proceedings of the 7th annual conference on Computer graphics and interactive techniques*, Association for Computing Machinery, 262–270.

[4] CUI 2020. 2020. CUI 2020 – COVID-19. Retrieved June 19, 2020 from https://cui2020.com/.

[5] Andrea Ferracani, Marco Faustino, Gabriele Xavier Giannini, Lea Landucci, and Alberto Del Bimbo. 2017. Natural Experiences in Museums through Virtual Reality and Voice Commands. *Proceedings of the 25th ACM international conference on Multimedia*, Association for Computing Machinery, 1233–1234.

[6] David M. Krum, Olugbenga Omoteso, William Ribarsky, Thad Starner, and Larry F. Hodges. 2002. Speech and gesture multimodal control of a whole Earth 3D visualization environment. *Proceedings of the symposium on Data Visualisation 2002*, Eurographics Association, 195–200.

[7] Dmitry Kutcenko. 2019. RPG/FPS Game Assets for PC/Mobile (Industrial Set v2.0). Retrieved from https://assetstore.unity.com/ packages/3d/environments/industrial/rpg-fps-game-assets-for-pc-mobile-industrial-set-v2-0-86679.

[8] Oculus Blog. 2019. Introducing Hand-tracking on Oculus Quest—Bringing Your Real Hands into VR. Retrieved from https://www.oculus.com/blog/introducing-hand-tracking-on-oculus-quest-bringing-your-real-hands-into-vr/.

[9] Oculus Blog. 2019. Thumbs Up: Hand-tracking on Oculus Quest This Week. Retrieved from https://www.oculus.com/blog/thumbs-up-hand-tracking-now-available-on-oculus-quest/.

[10] RAZGRIZZZ DEMON. 2019. Robot Sphere. Retrieved from https://assetstore.unity.com/ packages/3d/characters/robots/robot-sphere-136226.

[11] Jaisie Sin and Cosmin Munteanu. 2020. Let's Go There: Combining Voice and Pointing in VR. *2nd International Conference on Conversational User Interfaces (CUI 2020)*, ACM.