Multimodal User-Defined inputs for Optical See Through Augmented Reality Environments

Adam Williams*
Colorado State University

Francisco Ortega†
Colorado State University

ABSTRACT

Augmented Reality headsets are becoming affordable and accessible; however, there are no standardized inputs or design guidelines available for these devices. It is critical that human-centric interaction techniques are designed for this emerging technology. We present existing research exploring that design space and a road map of future work to further flesh out a user-derived multi-modal interaction set for these devices.

Index Terms: H.5.2 [User Interfaces]: User Centered Evaluation—User Centered Design; H.5.m [Artificial, augmented, and virtual realities]: Evaluation/methodology

(a) First elicitation study (b) Proposed validation environment

Figure 1: Multi-modal interaction

1 INTRODUCTION

Understanding multi-modal interaction within augmented reality (AR) head-mounted displays (HMDs) is pivotal to improving user interaction. Consider the impact that the desktop computer, smartphone, and tablet have had on people's lives. Augmented reality is one of the key technologies expected to become pervasive, as these devices already have.

While we expect to see a transitional period where people use their phones with their AR glasses, overtime, more natural user interfaces may become the preferred interaction method. As an example, the Microsoft HoloLens already ships with gesture and speech as the default input. We are motivated by Mark Weiser's vision to make the computer invisible [13]. While watching novice users interact in AR and virtual reality (VR) environments, we have seen evidence that the learning curve for the inputs is steep. People often have difficulties with understanding and using the controls. Furthermore, there is not a clear standard [1] when it comes to mid-air gestures (compared to multi-touch devices). Our motivation is to enable people to do complex tasks more easily by decreasing the interactions learning complexity. In other words, allowing them to focus on the problem rather than the controls.

My Ph.D. research is dedicated to developing a set of multi-modal user interactions for simple and complex augmented reality environments, in particular, environments that can be used to build (e.g., Lego-like applications). I will focus on developing user-defined interactions for gesture, speech, and gesture with speech inputs.

2 RELATED WORK

Interface design must be intuitive [12]. Outside of human-computer interaction knowledge, there is a vast collection of human to human communication literature. Human to human interaction is what we as people are most familiar with. An interface that mirrors those interactions may reduce the learning curve needed for technology's use. With that framework in mind, it is important to have systems with multi-modal interactions (e.g., gesture and speech combined).

Using multi-modal inputs has many benefits, particularly when dealing with gestures and speech combined. Gesturing has been shown to help lower the cognitive load of a task [3], there are hints at sped up task completion time, and even lower error rates [6]. Each information stream (gesture, speech) contains non-redundant information [2]. Additionally, gesticulation is closely linked to the rhythmical structure of co-occurring speech [4], allowing for better error recovery in recognizers [5]. Information from each channel lends to mutual disambiguation of the inputs from either channel.

However, current AR-HMDs (i.e. Magic Leap One and Microsoft Hololens) are built with gesture sets that are limited and likely designed from the point of view of recognition accuracy, not ease of use. For example, Magic Leap's C gesture is fairly easy to detect (being a static symbolic gesture) but may not be the most intuitive. Occasionally gesture sets are derived from users; however, these may be expert users [15]. People typically prefer user-defined gesture sets to expert-designed sets [15]. There is also evidence that elicited gestures are up to 25% more memorable [10].

My work aims to develop efficient user interaction sets that will enable novice users to smoothly interact with AR-HMD systems. Metrics of success for this goal include task completion times, user frustration, and cognitive load.

3 CURRENT RESEARCH

I have finished the first of a series of elicitation studies aimed at deriving a maximally guessable set of interaction techniques for gesture, speech, and gesture and speech combined (submitted for publication). The first experiment is a single object user elicitation study. The study was a within-subjects experiment composed of one independent variable (input modality) with three levels (gestures only, speech only, gestures and speech). The study was conducted using Wizard-of-Oz (WoZ) design [14] and implemented on a Magic Leap One optical see-through AR-HMD. Participants were asked for the input modalities in a counterbalanced order; the referents were randomly presented. The canonical referents were used with the addition of create, delete, and select. The NASA TLX survey was administered after each input. An example of what the participant sees during the experiment can be seen in Figure 1a.

3.1 Findings

This elicitation study yielded insightful results as well as some design guidelines for future elicitation studies. We found that users almost always (95.8%) produced manipulation gestures for translation and rotation referents. Most of those gestures involved the user reaching out to directly manipulate the virtual object presented. The "select" referent showed high agreement rates (a measure of user

*e-mail: AdamWil@colostate.edu
†e-mail: fortega@colostate.edu

978-1-7281-6532-5/20/$31.00 ©2020 IEEE
DOI 10.1109/VRW50115.2020.00130

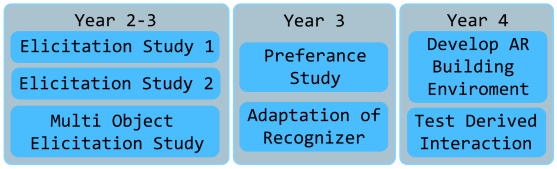

Figure 2: Proposed course of studies

consensus defined in [15]) due to the clicking gesture from touch screens and the mouse. Legacy bias [9] was also seen in the scale referents where users would preform a two-finger pinch as seen with a touch screens zoom in /zoom out gestures, as well as with a "bloom" gesture for create. This study also saw a minor shift from manipulation gestures to deictic gestures with descriptive words when allowing for both gestures and speech (e.g., "roll clockwise" with twisting hand and "roll" with a pointing gesture). Based on the NASA TLX results, users felt no statistically significant difference in perceived workload when producing gestures and speech versus when just producing either individually, although the mean for the gesture and speech block was highest.

One of the most interesting findings is that the times between when gestures were initiated and an utterance was initiated in milliseconds were (M = 151.31, SD = 120.24, Median = 130). Based on a Wilcoxon Signed rank test ($p < 0.01$) we can assume that the true median for the data is above zero. This means that gestures are nearly always started before the speech. This finding shows that gesture and speech interactions in AR-HMDs have similar timings as found in human to human communication [7]. This again follows the more general model that the onset of a gesture proceeds the onset of an utterances [4].

4 Future Research

The next step in my research is to run an additional elicitation study where the participant is shown the action of the referent without words. This is to remove any speech biasing from the referent being displayed as text. Using methods refined from the first two interaction studies, a final multi-object elicitation study will be performed. This will feature an expanded set of referents to be more suitable to a multi-object building environment.

Based on the results of all of the elicitation studies, a set of multi-modal (gesture, speech) interactions will be created and tested in a user preference study. This preference study's measures will include ease of use, appropriateness for task, and accuracy using some of the methods outlined by Nielson et al. [12]. During this study, egocentric videos of user gestures will be collected and labeled for use as training data.

Upon completion of that study, a set of multi-modal interactions will be derived. The next phase of my work will be to adapt a recognizer system to validate the preferred gestures selected by participants and provide an interactive system. This goal will be aided by the Computer Vision Lab at CSU which has years of experience working with gesture and speech-based real-time interactive systems [11]. During this phase, the timing information found during the elicitation studies will be utilized to achieve higher accuracy.

Finally, I will implement the derived interaction set into an AR building environment (seen in figure 1b). With this tool, I can validate whether the set of interaction techniques developed with participatory design has any benefits over standard inputs in an in the wild scenario. This study will have high external validity (whereas elicitation alone has high internal validity). This tool will measure task completion time, accuracy, error rates. As well as surveys for perceived workload, frustration, and ease of use. By combining those a clear picture of how this interaction set preforms can be painted. I hypothesize that user-defined inputs will result in lower frustration and more accurate interactions in this environment. A summary of the tasks to be completed during my Ph.D. is found in Figure 2.

5 Contributions

This work will help expand upon the knowledge in working with head-mounted (ego-centric) cameras, along with issues of self-occlusion caused by this perspective. It will contribute to participatory experimental design and input design. It will also bridge the work of linguists [4, 8] into computer science, which may be one of the most important contributions from my doctoral studies.

6 Conclusion

With AR-HMDs becoming more commonplace, design guidelines for guessable and ergonomic inputs for these devices are critical to their acceptance. These inputs should be designed from the end-user up. My work will contribute to developing user-defined interaction sets with multi-modal inputs for AR-HMDs. This work will be tested in the field against available inputs to validate whether or not they perform better than and if they are preferred over stock inputs by end-users. These human-centric inputs will stand as a contributing factor to AR-HMDs widespread adaptation and ease of use.

7 Questions

1. Do I need to narrow the scope of this proposal? 2. Has an important metric been overlooked? 3. Should the validation be done within a controlled setting?

References

[1] A. Dünser, R. Grasset, H. Seichter, and M. Billinghurst. Applying hci principles to ar systems design. 2007.

[2] S. Goldin-Meadow, M. W. Alibali, and R. B. Church. Transitions in concept acquisition: using the hand to read the mind. *Psychological review*, 100(2):279, 1993.

[3] S. Goldin-Meadow, H. Nusbaum, S. D. Kelly, and S. Wagner. Explaining math: gesturing lightens the load. *Psychol. Sci.*, 12(6):516–522, Nov. 2001.

[4] A. Kendon. Gesticulation and speech: Two aspects of the process of utterance in M. 25, Jan. 1980.

[5] D. B. Koons, C. J. Sparrell, and others. Integrating simultaneous input from speech, gaze, and hand gestures. *MIT Press: Menlo Park, CA*, 1993.

[6] M. Lee, M. Billinghurst, W. Baek, R. Green, and W. Woo. A usability study of multimodal input in an augmented reality environment. *Virtual Real.*, 17(4):293–305, Nov. 2013.

[7] D. P. Loehr. Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology*, 3(1):71–89, 2012.

[8] D. McNeill. *Gesture and Thought*. 2005.

[9] M. R. Morris, J. O. Wobbrock, and A. D. Wilson. Understanding users' preferences for surface gestures. In *Proceedings of graphics interface 2010*, pp. 261–268. Canadian Information Processing Society, 2010.

[10] M. A. Nacenta, Y. Kamber, Y. Qiang, and P. O. Kristensson. Memorability of pre-designed and user-defined gesture sets, 2013.

[11] P. Narayana, J. R. Beveridge, and B. A. Draper. Gesture recognition: Focus on the hands. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5235–5244, 2018.

[12] M. Nielsen, M. Störring, T. B. Moeslund, and E. Granum. A procedure for developing intuitive and ergonomic gesture interfaces for HCI. In *Gesture-Based Communication in Human-Computer Interaction*, pp. 409–420. Springer Berlin Heidelberg, 2004.

[13] M. Weiser. The computer for the 21st century. *Mobile Computing and Communications Review*, 3(3):3–11, 1999.

[14] J. O. Wobbrock, H. H. Aung, B. Rothrock, and B. A. Myers. Maximizing the guessability of symbolic input, 2005.

[15] J. O. Wobbrock, M. R. Morris, and A. D. Wilson. User-defined gestures for surface computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pp. 1083–1092. ACM, New York, NY, USA, 2009.