Comparing Single-modal and Multimodal Interaction in an Augmented Reality System

Zhimin Wang[1,2] Huangyue Yu[1] Haofei Wang[2] Zongji Wang[1] Feng Lu[1,2, *]

[1] State Key Laboratory of VR Technology and Systems, School of CSE, Beihang University
[2] Peng Cheng Laboratory, Shenzhen, China

Abstract

Multimodal interaction is expected to offer better user experience in Augmented Reality (AR), and thus becomes a recent research focus. However, due to the lack of hardware-level support, most existing works only combine two modalities at a time, e.g., gesture and speech. Gaze-based interaction techniques have been explored for the screen-based application, but rarely been used in AR systemsy configurable augmented reality system. In this paper, we propose a multimodal interactive system that integrates gaze, gesture and speech in a flexibly configurable augmented reality system. Our lightweight head-mounted device supports accurate gaze tracking, hand gesture recognition and speech recognition simultaneously. More importantly, the system can be easily configured into different modality combinations to study the effects of different interaction techniques. We evaluated the system in the table lamps scenario, and compared the performance of different interaction techniques. The experimental results show that the *Gaze+Gesture+Speech* is superior in terms of performance.

Keywords: multimodal interaction, augmented reality, gaze, gesture, speech, AR system

Index Terms: Human-centered computing—Human computer interaction (HCI)—HCI design and evaluation methods—User studies; Human-centered computing—Visualization—Visualization design and evaluation methods; Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Mixed/augmented reality

1 Introduction

Augmented Reality (AR) has been attracting great research interest in the field of human-computer interaction. Via overlaying virtual content into the real environment, AR techniques aim at providing immersive and intuitive interaction experience.

Towards this purpose, researchers have been investigating various human-computer interaction techniques for AR. The common modalities are hand gesture, speech-based command and eye gaze [1, 2]. Each modality has its own advantages and disadvantages. Hand gesture-based system is the most intuitive, speech-based system provides better system controllability, and eye gaze requires less physical effort. However, hand gesture-based system has to handle the occlusion problem [3], and is likely to cause arm fatigue after long-time usage. Speech-based system requires user to remember and pronounce the verbal command correctly [5]. This increases user's cognitive workload, especially for complex tasks. Gaze-based system often suffers from the Midas Touch problem [6] and insufficient eye tracking accuracy [4]. Thus, using these modalities individually might not bring the user optimal experience.

Multimodal interaction is expected to take advantages of each modal and provide better usability, such as *Gesture+Speech* [5], *Gaze+Gesture* [1] and *Gaze+Speech* [2]. For example, the *Gesture+Speech* technique uses gesture to select an object and uses speech to trigger an action. However, there are few works fusing three or more interaction modalities. It is unclear yet that whether more modalities lead to more superior performance. Therefore, it is worthy to investigate triple-modal interaction system.

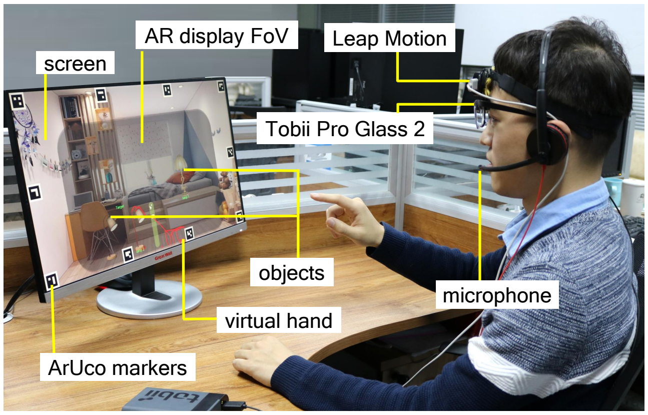

Figure 1: The gaze-gesture-speech AR system (GGS-AR) system setup. The user is interacting with the objects in augmented reality using the proposed system.

Several commercial AR devices that support multimodal interaction have been available in the past decades. The commonly used AR devices, such as Microsoft HoloLens and Oculus Rift , use speech and gesture as system inputs. In addition, it has been reported that Magic Leap one supports eye tracking, hand gesture recognition and speech recognition. However, the hardware configuration of the above-mentioned AR devices cannot be easily changed to adapt to the different needs in various studies. Thus, it becomes necessary to build a lightweight system that could be flexibly configured in different interaction modalities. Such system brings the opportunity of studying the effects of different interaction modalities in the same experimental settings.

In this paper, we propose a gaze-gesture-speech AR system (GGS-AR), as shown in Fig. 1. The GGS-AR system supports accurate gaze tracking, robust hand gesture recognition and speech recognition simultaneously. The head-mounted parts of the system sense user's input from different communication channels. The modality fusion and display parts of the system are left to the remote end. In the current implementation, we use a 2D computer screen to display the environment. However, our system can be easily adapted to display 3D environment once we use a stereo display. The size and shape of AR viewport can be arbitrarily modified. Our system can be flexibly configured into single-modal, double-modal and triple-modal interaction modalities. We evaluated the proposed system in two scenarios: table lamp and cube scenarios. The experimental results suggest that the *Gaze+Gesture+Speech* setting is superior to other modality combinations in terms of task performance. The

[*] Corresponding Author: Feng Lu (lufeng@buaa.edu.cn).

GGS-AR system provides initial insights for designing multimodal interactive AR system.

2 Experimental Setup

In our GGS-AR sytem, the hardware setup is shown in Fig. 1. The user wears the Tobii Pro Glasses 2, the Leap Motion Controller and the Plantronics Microphone. To achieve accurate eye tracking, we used the Tobii Pro Glasses 2. The manufacturer reports an accuracy of 0.6° and precision of 0.05°. The weight of Tobii Pro Glasses 2 is only 45 grams, which is comfortable for the user to wear. The Leap Motion controller is mounted on the user's head using a strap. It is placed above the eye tracker and faces forward for better hand gesture tracking. We place the microphone near the user's mouth to reduce the environmental noise. The user is required to sit at about 50 cm in front of a screen. We attach ten markers at the boundary of the screen (see Fig. 1).

The total weight of our head-mounted devices is 380 grams approximately, and its FoV is 50×35° in the current settings. The FoV can be adjusted by using varisized displays. Besides, each sensor of GGS-AR could also be replaced. For instance, The Tobii Pro Glasses 2 can be substituted with Pupil Labs' eye trackera or even a remote eye tracker.

3 Experiment Procedure

There is no earlier research fusing gaze, gesture and speech interaction simultaneously in one AR system. There are three types of combinations including three single-modal techniques, three double-modal techniques and one triple-modal technique, as shown in Fig. 2. In this work, we investigated five representative modalities: two single-modal techniques (*Gesture Only* and *Gaze Only*), and three multimodal techniques (*Gesture+Speech*, *Gaze+Speech* and *Gaze+Gesture+Speech*), as highlighted in gray color in Fig. 2.

single-modal	double-modal	triple-modal
Gesture only	Gesture+Gaze	Gaze+ Gesture+ Speech
Gaze only	Gesture+Speech	
Speech only	Gaze+Speech	

Figure 2: The table shows three types of combinations. We choose five representative modalities including two single-modal techniques and three multimodal techniques.

We conducted a 5 (modalities) × 1 (task scenario) user study to compare the usability of different modalities for AR interaction. The experiment has a repeated measure within-participants design, with interaction modality as the independent variables. The dependent variables include objective speed, accuracy.

Scenario: Table Lamps Table lamps scenario requires participants to adjust the brightness of different lamps to match the brightness of the target lamps. Table lamps are placed on a wooden box and on the floor. When a certain lamp is selected through different modalities, the target lamp appears on the left of the lamp with red edge surrounding it, indicating that the lamp is selected. Then, participants can brighten or darken the lamp by sliding the control bar, pressing buttons, or looking at buttons. Finally, they can deselect the lamp when they think the brightness of adjusted lamp and target lamp is the same. In this scenario, there are two lamps need to be adjusted.

4 Experimental Results

We recruited 12 subjects on campus to conduct the experiment (8 male, 4 female), the average age is 23.8 (SD = 1.6). All participants are able to see the hint on computer screen clearly without glasses, even the myopia participants. All the participants can read and speak English fluently.

4.1 Completion Time

We used a repeated-measures ANOVA ($\alpha = .05$), in conjunction with post hoc pairwise T-tests with Bonferroni correction to identify whether the significant difference exists among modalities on the task time. The statistical analysis showed that it failed to reject the equality of the levels of modalities on completion time for the table lamps scenario (p = 0.641).

4.2 Accuracy

We performed a repeated-measures ANOVA ($\alpha = .05$), in conjunction with post hoc pairwise T-tests with Bonferroni correction to identify whether the task accuracy varied significantly according to the modalities. The statistical analysis indicated that the effect of modalities on accuracy was statistically significant (F(4, 44) = 2.949, p = .052, $\eta^2 = .21$). Specifically, we found that *Gaze+Gesture+Speech* outperformed *Gaze Only*, *Gesture+Speech* and *Gaze+Speech* in terms of accuracy (Bonferroni-corrected p-values of 0.027, 0.027 and 0.026 respectively).

5 Discussion And Conclusion

We found that there was no significant difference in table lamps scenario in terms of trial completion time. This might due to the table lamps are large, which has less requirement for the time of modality manipulation. We analyzed the reason that the object size had a leading impact on the performance of modality. The size of table lamps is 7 × 4 cm, while the size of cubes is only 3 × 3 cm. Therefore, it may be easier to select and manipulate the table lamps. Future extensions of the GGS-AR may compare the results of different initial scales: 25%, 50%, 75%, 100% of target size.

In this paper, we proposed a novel GGS-AR system to investigate the benefits of multimodal interaction in AR system. We also evaluated and compared various combinations of three interactions in this system. Our lightweight AR head-mounted system integrates different sensors that support accurate gaze tracking, hand gesture and speech recognition simultaneously. The experimental results indicate that the *Gaze+Gesture+Speech* modality is superior to the other interaction modalities in terms of the interaction accuracy. This study offers preliminary insights to design multimodal interactive AR systems in a more flexible way.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 61972012.

References

[1] I. Chatterjee, R. Xiao, and C. Harrison. Gaze+gesture: Expressive, precise and targeted free-space interactions. In *International Conference on Multimodal Interaction*, pp. 131–138. ACM, 2015.

[2] M. Elepfandt and M. Grund. Move it there, or not?: the design of voice commands for gaze with speech. In *Workshop on eye gaze in intelligent human machine interaction*, p. 12. ACM, 2012.

[3] A. O. S. Feiner. The flexible pointer: An interaction technique for selection in augmented and virtual reality. In *ACM Symposium on User Interface Software and Technology*, pp. 81–82. ACM, 2003.

[4] M. Kytö, B. Ens, T. Piumsomboon, G. A. Lee, and M. Billinghurst. Pinpointing: Precise head- and eye-based target selection for augmented reality. In *ACM Conference on Human Factors in Computing Systems*, p. 81. ACM, 2018.

[5] M. Lee, M. Billinghurst, W. Baek, R. D. Green, and W. Woo. A usability study of multimodal input in an augmented reality environment. *Virtual Reality*, pp. 293–305, 2013.

[6] P. Mohan, W. B. Goh, C. Fu, and S. Yeung. Dualgaze: Addressing the midas touch problem in gaze mediated VR interaction. In *International Symposium on Mixed and Augmented Reality*, pp. 79–84. IEEE/ACM, 2018.