Hands and Speech in Space: Multimodal Interaction with Augmented Reality interfaces

Mark Billinghurst
Human Interface Technology Laboratory New Zealand
University of Canterbury
Ilam Road, Christchurch, New Zealand
+64-3-364-2349
mark.billinghurst@hitlabnz.org

ABSTRACT

Augmented Reality (AR) is technology that allows virtual imagery to be seamlessly integrated into the real world. Although first developed in the 1960's it has only been recently that AR has become widely available, through platforms such as the web and mobile phones. However most AR interfaces have very simple interaction, such as using touch on phone screens or camera tracking from real images. New depth sensing and gesture tracking technologies such as Microsoft Kinect or Leap Motion have made is easier than ever before to track hands in space. Combined with speech recognition and AR tracking and viewing software it is possible to create interfaces that allow users to manipulate 3D graphics in space through a natural combination of speech and gesture. In this paper I will review previous research in multimodal AR interfaces and give an overview of the significant research questions that need to be addressed before speech and gesture interaction can become commonplace.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]: Artificial, augmented, and virtual realities, H.5.2 [**User Interfaces**]: Voice I/O, Interaction styles, I.2.7 [**Natural Language Processing**]: Speech recognition and synthesis, I.3.7 [**Three-Dimensional Graphics and Realism**]: Virtual reality.

General Terms:

Design, Experimentation, Human Factors.

Keywords

Augmented Reality, Multimodal Interfaces, Speech, Gesture

1. INTRODUCTION

Augmented Reality (AR) is technology that allows virtual imagery to be seamlessly overlaid on the real world, so both can be seen in the same place in real time [1]. Although the technology was first developed in the 1960's it has only recently become widely available, through platforms such as the web and mobile phones. In most AR applications user interaction is relatively simple, using on-screen touch or device input. However these methods break the illusion that users can directly interact with virtual content in the real world.

Recently, researchers have begun to explore gesture-based interaction with AR content. Depth sensing and gesture tracking technologies such as Microsoft Kinect [7] or Leap Motion [9] have made it easier than ever before to track hands in space and provide natural free-hand gesture into an AR application. For example, Bai et. al. [2] have shown how a Kinect can be used to allow users to directly manipulate AR content on a mobile phone.

In my research I am particularly interested moving beyond gesture interaction and adding speech input to explore multimodal interfaces for Augmented Reality. In the next section I provide a brief overview of earlier related work in AR multimodal interaction and then give a case study of a current system. Finally, I highlight areas of future research that will need to be addressed before speech and gesture interaction can become commonplace.

2. RELATED WORK

Previous research has shown that combined speech and gesture input can improve interaction in desktop or immersive 3D graphics applications. For example, Chu. et. al. [4] developed a multimodal interface for a Virtual Reality design application, while others have shown speech and gesture can be used to navigate through virtual worlds [8]. The main benefit of combining speech and gesture in this way is that they can be used to express different but complimentary types of input [5].

Based on this research we should expect the same benefits for AR applications, however multimodal input for AR has not been widely studied. One of the earliest examples was SenseShapes [10], an interface in which the user's gaze direction and gestures were combined with speech input to interact with virtual objects in an AR scene. The user wore a data glove, head mounted display and viewpoint tracking equipment. Irawati et al. [6] developed a computer vision based AR systems with multimodal input, allowing a user to pick and place virtual furniture in an AR scene using a combination of paddle gestures and speech commands. In the evaluation study they found that multimodal input enabled subjects to complete a task faster than with gesture alone.

3. EXAMPLE INTERFACE

Our earlier work [3] provides a good example of how to develop a multimodal AR interface. This system is a multimodal interface (MMI) that combines free hand gesture and speech input using a fusion architecture. Figure 1 shows the system architecture with three main components (1) speech and gesture recognition systems, (2) a fusion module that combines both inputs together, and (3) an AR scene manager that provides interaction with the virtual objects in the AR scene.

ICMI'13, December 9–13, 2013, Sydney, Australia.
ACM 978-1-4503-2129-7/13/12.
http://dx.doi.org/10.1145/2522848.2532202

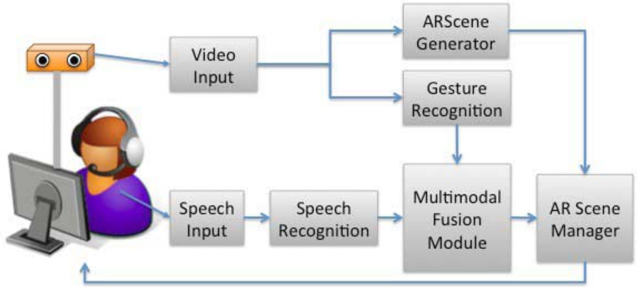

Figure 1. A Typical Multimodal AR System Architecture

In this application, a stereo camera is used to provide free hand tracking enabling the system to track the users hand and recognize simple gestures such as open hand, a fist and pointing. The fusion module used a sliding time window to find the speech commands associated with particular gestures and produce a single output command. This command is fed to the AR scene manager that then modifies the AR scene.

This same architecture could be used in a wide variety of applications, but in this case it was used in a simple scene arranging application. Using an AR mirror set up (see figure 2), a user could use combined speech and input to select 3D objects and change their shape and colour to match that of a target object.

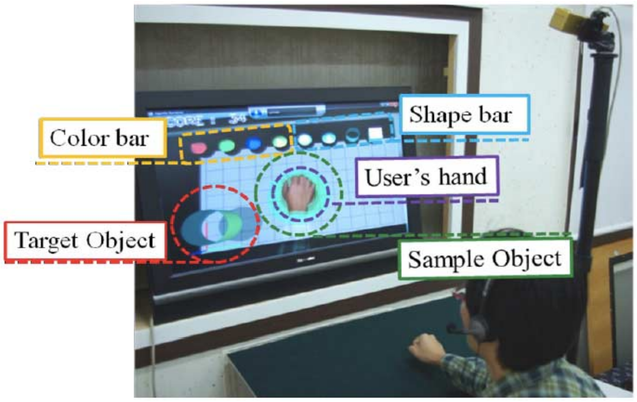

Figure 2. AR Mirror Multimodal Interface

A user study was conducted where subjects had to complete a set of ten tasks involving changing object shape, colour and position. This was done in three different interface conditions (1) speech only, (2) gesture only, and (3) MMI. Users were able to complete the tasks significantly faster with the MMI than gesture only input, and users felt that the MMI was more efficient, faster, and accurate than the gesture-only condition. However there was no difference between the MMI and speech only conditions. A total of 70% of the users preferred the MMI to the single modal interfaces (25% preferred speech only and 5% gesture only).

4. AREAS FOR FUTURE RESEARCH

In this paper we have provided a brief summary of earlier work in multimodal interfaces for AR applications, and then an example of the type of system that can be implemented. The results from using this system seem positive and provide a more natural interaction than what is possible with other types of AR interfaces. However there are still a number of research areas that need to be explored before multimodal interaction is common.

The first is to provide improved hand tracking and gesture recognition. Most existing multimodal interfaces use simple hand gestures such as pointing or grasping. However research could be conducted on using depth sensors to create more accurate 3D kinematic models of the hands allowing for a richer range of gestures to be recognized. This would support the use of pantomimic or iconic gestures rather than simple deictic gestures.

A second area of research is to capture not just the user's gestures, but also the natural environment around them, allowing the AR application to understand the geometry and structure of the real world and for the virtual content to interact with real objects. In order to do this, research will need to be conducted in semantic representation of the captured scene.

Finally, more research needs to be conducted on multimodal fusion strategies and how to reliably combine speech and gesture input so that the user's intent can be conveyed to the AR application. This is particularly important as a wider range of speech and gesture types are used for input.

As can be seen multimodal interaction for AR applications will provide a rich field of research for a number of years to come.

5. REFERENCES

[1] Azuma, R. 1997. A Survey of Augmented Reality. *Presence,* 6 (4), 355-385.

[2] Bai, H., Lee, G. A., and Billinghurst, M. 2012. Freeze view touch and finger gesture based interaction methods for handheld augmented reality interfaces. In *Proceedings of the 27th Conference on Image and Vision Computing New Zealand* (pp. 126-131). ACM.

[3] Billinghurst, M. and Lee, M (2012). Multimodal Interfaces for Augmented Reality. In Dill, John, Earnshaw, Rae, Kasik, David et al (editors), *Expanding the Frontiers of Visual Analytics and Visualization.* London : Springer.

[4] Chu CP, Dani TH, and Gadh R. 1997. Multimodal Interface for a virtual reality based computer aided design system. *Proceedings of 1997 IEEE International Conference on Robotics and Automation* 2: 1329-1334

[5] Cohen, P. R., Dalrymple, M., Moran, D. B., Pereira, F. C., and Sullivan, J. W. 1989. Synergistic use of direct manipulation and natural language. In *ACM SIGCHI Bulletin* (Vol. 20, No. SI, pp. 227-233). ACM.

[6] Irawati, S., Green, S., Billinghurst, M., Duenser, A., and Ko, H. 2006. An evaluation of an augmented reality multimodal interface using speech and paddle gestures. In *Advances in Artificial Reality and Tele-Existence* (pp. 272-283). Springer Berlin Heidelberg.

[7] Microsoft Kinect Website: http://www.xbox.com/KINECT

[8] Krum, D. M., Omoteso, O., Ribarsky, W., Starner, T., and Hodges, L. F. 2002. Speech and gesture multimodal control of a whole Earth 3D visualization environment. In *Proceedings of the symposium on Data Visualisation 2002* (pp. 195-200). Eurographics Association.

[9] Leap Motion Website: https://www.leapmotion.com/

[10] Olwal, A., Benko, H., and Feiner, S. 2003. SenseShapes: Using statistical geometry for object selection in a multimodal augmented reality system. In *Proceedings of the 2nd IEEE/ACM International Symposium on Mixed and Augmented Reality* (p. 300). IEEE Computer Society.