Multimodal Speech-Gesture Interaction with 3D Objects in Augmented Reality Environments

---

A thesis submitted in partial fulfilment

of the requirements for the Degree

of

Doctor of Philosophy

in the

University of Canterbury

by

Minkyung Lee

University of Canterbury

2010

Publications from this dissertation

Material from this dissertation has been previously published in or submitted to the peer-reviewed papers or a journal listed below. The chapters of this thesis that relate to each publication are noted.

1. Lee, M. and Billinghurst, M.: 2008, A Wizard of Oz Study for an AR Multimodal Interface, Presented at International Conference on Multimodal Interfaces (ICMI2008), (Chania, Oct. 20-22, 2008) (Chapter 3).

2. Lee, M., Green, R., and Billinghurst, M.:2008, 3D Natural Hand Interaction for AR Applications, Presented at International Vision Conference New Zealand (IVCNZ 2008), (Lincoln, New Zealand, Nov. 26-28, 2008) (Chapter 3).

3. Lee, M. and Billinghurst, M.: 2009, Interaction Space-based Gesture Classification for MultiModal Input in an Augmented Reality Environment. Presented at International Workshop on Ubiquitous Virtual Reality (IWUVR 2009), (Adelaide, Austrailia, Jan. 15-18, 2009) (Chapter 4).

4. Lee, M. and Billinghurst, M.: 2009, User Observation to Design a Space-based Gesture Interface for U-VR Environments, Submitted to Journal of Research and Practice in Information Technology (Chapter 4).

5. Lee, M., Billinghurst, M., Baek, W., Green, R., and Woo, W. A Study on the Usability of a Seamless Multimodal Interface in an Augmented Reality Environment, Submitted to Virtual Reality, Springer Journal (Chapter 5, 6, 7).

Dedicated to

my family.

Abstract

Augmented Reality (AR) has the possibility of interacting with virtual objects and real objects at the same time since it combines the real world with computer-generated contents seamlessly. However, most AR interface research uses general Virtual Reality (VR) interaction techniques without modification. In this research we develop a multimodal interface (MMI) for AR with speech and 3D hand gesture input. We develop a multimodal signal fusion architecture based on the user behaviour while interacting with the MMI that provides more effective and natural multimodal signal fusion. Speech and 3D vision-based free hand gestures are used as multimodal input channels. There were two user observations (1) a Wizard of Oz study and (2) Gesture modelling. With the Wizard of Oz study, we observed user behaviours of interaction with our MMI. Gesture modelling was undertaken to explore whether different types of gestures can be described by pattern curves. Based on the experimental observations, we designed our own multimodal fusion architecture and developed an MMI. User evaluations have been conducted to evaluate the usability of our MMI. As a result, we found that MMI is more efficient and users are more satisfied with it when compared to the unimodal interfaces. We also describe design guidelines which were derived from our findings through the user studies.

Table of Contents

List of Figures

List of Tables

Abbreviations

ANOVA	Analysis of variance
AR	Augmented reality
MMI	Mulimodal Interface
VR	Virtual Reality
WOz	Wizard of Oz
TUI	Tangible User Interface
MR	Mixed Reality
HMD	Head Mounted Display
HHD	Hand Held Display
GIS	Geographic Information System
HMM	Hidden Markov Model
TAR	Tangible Augmented Reality

Acknowledgements

After all those years, I have got quite a list of people who helped in some way to this thesis. I would like to express my thanks here.

This thesis would not have been done without all the supports, the trenchant critiques, the probing questions, and the remarkable patience of my supervisor Mark Billinghurst. He was always accessible and willing to help me with my research. As a result, my research life became smooth and rewarding for me. I cannot thank him enough! I thank another supervisor, Richard Green, who encouraged me to stay positive throughout the PhD program. I would also like to thank my thesis examiners, Holger Regenbrecht and Didier Stricker, for taking the time to read, consider and evaluate my work.

Let me also say 'thank you' to all staff and students at the HIT Lab NZ: Senior researcher Raphael Grasset, Post-doc researchers Hartmut Seichter, and Andreas Duenser for their fruitful comments and advices on my research and life; Software engineer Julian Looser who helped me out at any time in many ways, especially, solving my unsolvable programming problems; Administration team, Ken Beckmam and Katy Bang, for their supports; my office mates, Christina Dicke and Mohammad Obaid and other staff and students at the lab.

I also would like to thank my friends, Nora & Phillip, Daniela, Cameron, Shaleen, Joerg, Eugene, and Keunjin, for their encouragements and supports.

My deepest gratitude goes to my family for their unflagging love, encouragement, and support throughout my life; this dissertation is simply impossible without them: My parents, Chunbae and Haesoog, deserve special

mention for their inseparable love, support, and prayers. They taught me the sense of perfection, honesty, hard work, ethics in life and the aptitude for knowledge. They always trusted and supported my independent decisions and have confidence in me. My sisters, Juhyun and Suyeon, thanks for being supportive. Finally, I would like to thank my husband, Kiyoung Kim. He has given me all his support, encouragement and love. I cannot imagine how I would have been able to complete this work without his love and support. Thank you, Kiyoung. I hope you finish up your study soon. Then, we will be able to get our lives back!

Chapter 1

Introduction

Augmented Reality (AR) is a technology that overlays computer-generated information onto the real world (Azuma, 1997). The goal of AR systems is to provide users with information-enhanced environments that seamlessly connect real and virtual worlds. To achieve this, accurate tracking and registration methods are essential for aligning real and virtual objects. In addition, natural interaction techniques for manipulating the AR contents should also be provided. Most AR interface research uses traditional Virtual Reality (VR) interaction techniques, such as a dataglove, without modifications. Adopting VR interaction techniques yield gaps between the virtual environment and real-world because they only consider interaction techniques useful in virtual environments. To provide seamless interaction in the AR environment, we should consider how to interact in the virtual world and real world at the same time.

Multimodal Interfaces (MMIs) are interfaces that process two or more combined user input modes in a coordinated manner with multimedia system

output (Oviatt, 2003). An intuitive interface is immediately understandable to all users who have neither special knowledge nor special education (Bærentsen, 2001). This implies that a user can walk up to the system with an intuitive interface; see what kind of functions the system affords and what needs to be done to operate it. The goal of a MMI is to provide an intuitive and efficient method of interaction by allowing a person to use multiple input modes. In human communication, gestures and speech are co-expressive; they arise from a shared semantic source but are able to express different but complimentary information (Quek *et al.*, 2002). The same use of co-expressive modalities can be used to create natural human computer interfaces. For example, speech input can be combined with pen gestures to create an intuitive command and control application (Cohen *et al.*, 1997).

In the past, MMIs have been used not only for 2D user interfaces but also for interacting with 3D virtual contents. Chu *et al.* showed how multimodal input can be used in VR applications to interact with virtual objects (Chu *et al.*, 1997) while Krum *et al.* used it to navigate a virtual world (Krum *et al.*, 2002). Laviola Jr. developed a prototype multimodal tool for scientific visualization in a immersive virtual environment (Laviola Jr., 2000). In his Multimodal Scientific Visualization Tool, a user could not only interact with virtual objects but also navigate through the VR scene by using gesture input from the pinch

gloves and triggering corresponding speech input. Wang proposed a multimodal interface with gaze, 3D controller, voicekey and keyboard to select and manipulate the virtual object in the desktop VR environment (Wang, 1995).

In our research we are studying how MMI techniques can be applied to Augmented Reality (AR) interfaces. An MMI may be an ideal interaction technique for AR applications; because the MMI supports interactions in real and virtual worlds at the same time. We develop an AR MMI system that allows us to combine gesture and speech input with a multimodal fusion architecture that merges the two different input modalities in a natural way. Prior to developing the AR MMI, we run two user studies to learn how people use the MMI in a given AR environment for a user-centred MMI and multimodal fusion architecture design. Our MMI system is tested in a simple AR application and evaluated using a user study that compared speech-only and 3D hand gesture-only conditions with an AR MMI. This comparison is done in order to study the usability of the MMI compared to unimodal interfaces. Note that the scope of this thesis is limited to 3D object manipulation in AR environments.

The main contributions of this thesis are:

(1) Development of a Multimodal AR interface with 3D natural hand gesture and speech input

(2) Development of a semantic multi-channel signal fusion architecture for an AR MMI based on the user observations

(3) User observations and formal user studies with the proposed AR MMI.

(4) Design guidelines for 3D AR MMI.

(5) A full process for building a user-centred MMI for AR

Chapter 2 gives an overview of the context and state of the art of research in MMI and AR. First, it reviews previous multimodal interfaces for various applications in 2D or 3D environments. Then, it gives an overview of AR interfaces involving tangible user interfaces, hand gestures, and multimodal AR interfaces. It also gives an overview of different multimodal fusion architectures. The chapter summarises the research gaps and identifies the research contributions that this thesis makes. We also give an overview of how the research is realised: how we implemented and evaluated. There are three components to the AR MMI we have developed: (1) vision-based hand gesture recognition, (2) speech recognition, and (3) a semantic multi-channel signal fusion architecture.

Chapter 3 presents findings from the first user observation using the Wizard of Oz method. We observe both how users will want to input multimodal commands, and how different AR display conditions affect these commands. We measure the frequencies of speech and gesture commands and the time gap between combined speech and gesture commands by watching users from recorded video. We also interview each subject after completing the three given tasks.

Chapter 4 describes another user study for gesture modelling. We explore gesture input by observing and comparing users' gesture pattern in different environments: the Real, the Augmented and the Mixed environment. The goal of the study is to investigate how different types of gestures were used to interact with various objects. We want to explore whether deictic and metaphoric gestures can be classified only by observing where gestures are made with real and virtual objects in 3D. We also want to explore how users felt while they triggered different gestures in the task environments.

Chapter 5 describes our complete AR MMI. Based on the two user observations in Chapter 3 and Chapter 4, we designed a multimodal fusion framework that uses adaptive filters to merge speech and natural hand gesture input to interact with AR objects. We also designed and developed a small AR application to evaluate the usability of the interface.

The sample application is a desktop AR interface that allows users to move virtual objects and change their colour and shape with speech and/or gesture input. We describe how we developed our speech and gesture interface, and how the multimodal fusion architecture connects the two input methods together.

Chapter 6 presents findings from the last experiment to evaluate the usability of our AR MMI that is described in Chapter 5. We described a pilot user study and a full usability test exploring the usability of the seamless AR MMI for object manipulation, compared with speech-only and 3D hand gesture-only conditions by considering all three aspects of usability: effectiveness (accuracy and completeness), efficiency (use of time and resources), and satisfaction (preferences). After running the pilot study, we found several problems from the users' feedback. Based on the findings from the pilot study, we updated our AR MMI and run a full usability test. In the full usability test, we measured the usability factors of (1) efficiency, (2) effectiveness, and (3) satisfaction for each interface.

Chapter 7 proposes design guidelines for MMIs in AR environments which will be helpful for researchers who want to develop AR MMIs in various applications. We also review the main findings of this work and outlines directions for future research.

Chapter 2

Related Works

2.1 Introduction

Over the last forty years, there has been a significant amount of research conducted in the AR Field. However, most of this has been about tracking or registration (Swan & Gabbard, 2005). Recently, researchers have started undertaking research on AR interface methods and attempting to provide a more natural end user experience. Among the many possible interaction methods, we are interested in exploring an AR MMI that allows a person to use combined speech and gesture input to interact with virtual contents.

According to Hansson et al.'s definition, an interface is considered a natural one when it builds upon knowledge that the user already possesses (Hansson, 1997). For example using real-world navigation skills for virtual-world navigation is a way of natural interaction. In the sense, an AR MMI provides a natural interface; because a user can use their everyday communication skills, speech-gesture combination, to interact with augmented virtual objects. The

AR MMI blends elements of AR, MMI, and usability testing and so is based on previous work in each of these areas as shown in Figure 2. 1.

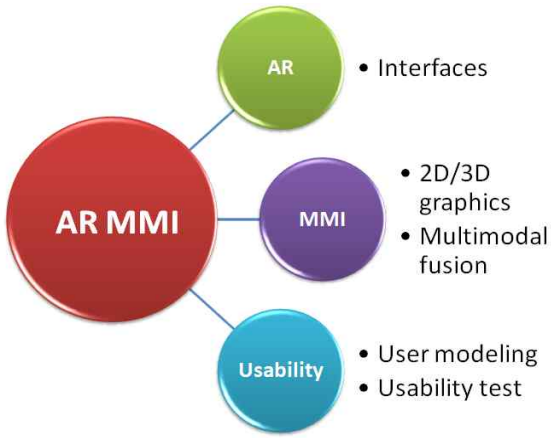


Figure 2. 1 Research outline

In this chapter we review previous research in the following areas:

- Multimodal interfaces for 2D/3D graphics
- AR Interfaces: Tangible User Interfaces, Gesture Interfaces, and MMIs
- Multimodal input fusion
- User modelling methods
- Usability testing

2.2 Previous MultiModal Interfaces

In this section, previous research on MMIs with 2D/3D graphics environments is studied and lessons learned are summarized.

Multimodal interfaces have a long history dating back to the 'Put-that-there' work (Bolt & Schamndt, 1980). Bolt used graphical actions with a space-sensing cube and synthesized speech as an interaction channel. One-handed pointing with a tracker was used to control virtual objects displayed on a large wall display. Users could create simple geometric shapes with the speech and gesture commands, give them names and details such as colour and size, move them around on a map, and delete them. A mixture of speech and gestures was used to select objects just like a mouse, even though multimodal interfaces can support much richer expressions.

Cohen *et al.* (1989) showed how a mixture of natural language and direct manipulation can overcome the limitations of each modality alone. The combination of speech and gesture provides a highly proficient communicative behaviour to interact with applications in a more transparent experience than GUI interfaces.

2.2.1 2D Interfaces

There have been a number of interfaces that have shown the value of an MMI on a desktop. QuickSet (Cohen *et al.*, 1997) is a multimodal interface for map-based tasks with pen and voice input. Users were able to issue combined speech and pen gesture commands. It also provided the same user input

capabilities for handheld, desktop, and wall-sized terminal hardware. It allowed users to label a map, or to put creative primitives or entities on the map. The multimodal interface was activated when a pen was on the screen for simultaneous speech and gesture fusion. Afterwards, each unimodal signal was processed in parallel. However, the pen-type interface was only designed for applications in 2D space.

DAVE-G (Rauschert *et al.*, 2002) is a collaborative Geographic Information System (GIS) application for a dialogue-assisted visual environment. Rauschert *et al*. developed a prototype for initial user studies which used speech commands for data queries, such as show, hide and locate, select and scroll, zoom and centre. Natural hand gestures were used to point and indicate an area and to outline contours using vision-based gesture recognition with Hidden Markov Models (HMM). The collaborative environment was generated by connecting stand-alone client applications via a network. However, Rauschert *et al*. did not evaluate their final application with user studies. DAVE-G only supports pointing and outlining contours with hand gestures. The fusion of speech and gesture recognition is based on a time-based analysis.

The GSI Demo (Tse *et al.*, 2006) was designed to allow users to rapidly create their own multimodal gesture/speech input wrappers. Tse *et al*. pointed out that most commercial applications had been designed for a single user using a keyboard or a mouse over an upright monitor. They adopted Cohen's unification-based multimodal integration algorithm (2002) to merge their speech and gesture input and to translate them to keyboard or mouse input. However, the multimodal mappings were limited in a certain way. For example, if clicking a menu option comes before than specifying a location, the multimodal command would fail. Thus their approach could be used only for specific existing single user applications. In addition, they did not conduct user studies to evaluate their research; thus, they cannot verify whether their approach is effective for users.

Some of the key lessons learnt from 2D multimodal interfaces include:

- That speech and gesture can be combined for extremely intuitive input
- The importance of having effective fusion techniques
- There have been few formal user studies conducted
- Map-based applications are considered as target application areas

2.2.2 3D Interfaces

Multimodal interfaces have also been used to interact with three dimensional computer graphics applications. Weimer and Ganapathy (1989) developed a virtual environment with speech and hand gesture input. They used a DataGlove for hand tracking; however only the thumb and index fingers were used for interaction because of the poor accuracy of the DataGlove. Thumb gestures are used to initiate a pick and the index fingertip is used like a stylus for locating. Speech assisted the system navigation with the hand tracking result.

ICONIC (Koons & Sparrell, 1994) is a descriptive interface to let users interact with computer-generated objects in a virtual environment with a free mixture of speech and depictive gestures. The proposed system did not allow users to manipulate virtual objects with their hands directly. Instead, users could describe the spatial and temporal aspects of a scene in the virtual environment. It was not necessary to learn a specific set of symbolic gestures for ICONIC because the system adopted the user's natural gesture instead of training users according to their symbolic gestures. ICONIC used a dataglove to capture the users' gesture input.

VisSpace (Lucente *et al.*, 1998) is a test bed for a deviceless multimodal user interface using computer vision techniques. Three dimensional graphical objects shown on a wall-sized display were controlled by speech and natural

gestures. VisSpace allowed users to manipulate and navigate through virtual objects and worlds and uses an integrator to integrate speech and gesture input sequentially for a valid command. The integrator assumed one second latency in the vision channel from a time-stamp from speech input. However, as mentioned in (Oviatt *et al.*, 2004), speech and gesture input did not happen sequentially all the time; thus, this kind of integration of speech and gestures could not support natural interaction.

Sowa and Wachsmuth developed an application which supports multimodal interaction to verify their Imagistic Description Tree (IDT) (Sowa & Wachsmuth, 2005). The IDT is a tree-like structure for information representation of imagistic and analogical nature. To show how their data representation structure worked, they developed an interface which allowed a user to make a certain shape of object with a voice command and body gesture; however, the user needed to wear three trackers on their back, hand, and elbow to allow system tracking of the users' body gesture. A data glove was also required to capture the users hand motion.

Some of the key lessons learnt from 3D multimodal interfaces include:

- that main application domains were virtual environments

- that speech and gesture input was main components for multimodal interaction
- data glove to track users hand gesture input was cumbersome

2.3 Previous Research on AR Interfaces

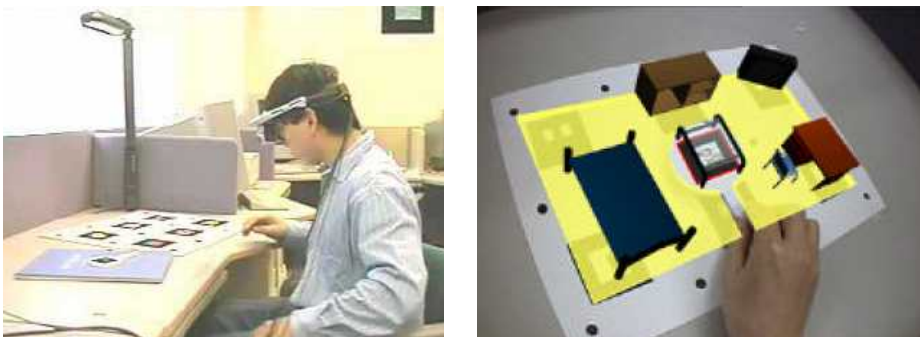
Although AR technology offers new possibilities for interacting with computer generated contents, much of the previous research in AR was about viewpoint tracking or virtual object registration but not interaction techniques (Swan II and Gabbard 2005). In this section we summarize related work AR interfaces, include previous work on Tangible User Interfaces, hand gesture input, and AR MMIs.

2.3.1 Tangible User Interfaces

The concept of a Tangible User Interface (TUI) was first defined by Ishii and Ullmer (1997). A TUI connects the real world of atoms with the virtual world of bits and bytes. The two different basic properties of two different worlds, atoms and bits, are closely coupled by mapping the virtual information to physical objects.

Kato and Billinghurst (1999) released the ARToolKit software library which made camera viewpoint tracking easy by using black square markers with unique shapes in the markers for identification. As a result, building AR applications became much easier than before and this contributed to the rapid growth of AR research. Many of the AR interfaces developed with ARToolKit used the TUI metaphor. Markers were attached to conventional interaction devices or physical objects, and the positions the objects were tracked corresponding to the position of the attached markers.

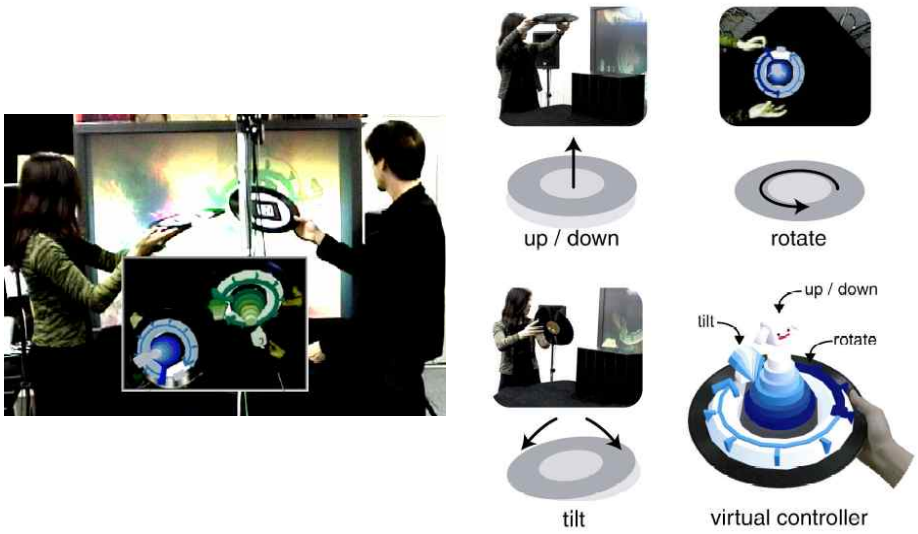
One of the early examples of a Tangible Augmented Reality (TAR) was the VOMAR application produced by Kato *et al*. (2000). In VOMAR, people had a marker-attached paddle for interacting with virtual furniture in a real book (see Figure 2.2). The paddle was used for picking up, moving and placing virtual furniture from one position to the desired place.



(a) (b)

Figure 2.2 VOMAR application: (a) system configuration and (b) AR view - a user interacts with the paddle

A dynamic gesture with the paddle was used to remove the virtual furniture objects either from the paddle or the target position. This interface enabled users to easily interact with augmented virtual objects; however, it required the user to carry a special paddle.



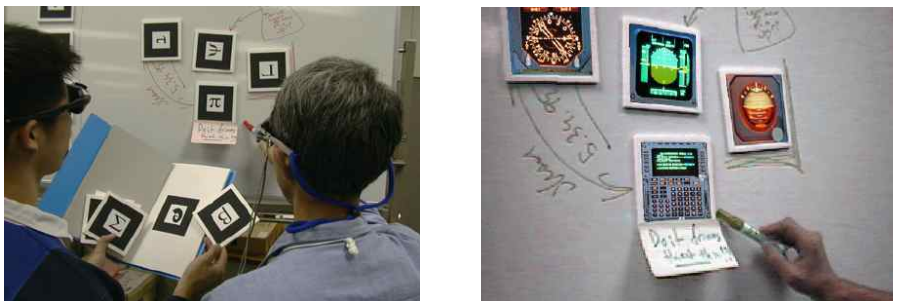
(a) (b)

Figure 2.3 Augmented Groove: (a) The users playing music with Augmented Groove and (b) Gesture interface for Augmented Groove

Poupyrev *et al.* (2001) developed the Augmented Groove an interface for electronic music performance (see Figure 2.3). A fiducial marker attached to real records was used to tracking record motion and to overlay 3D virtual controllers on top of them. Different fiducial markers were used to map corresponding music sequences to the markers. The markers gave users instant feedback on the status of the musical performance. The users could move the records up and down, rotate them, or tilt them, to have different modulations,

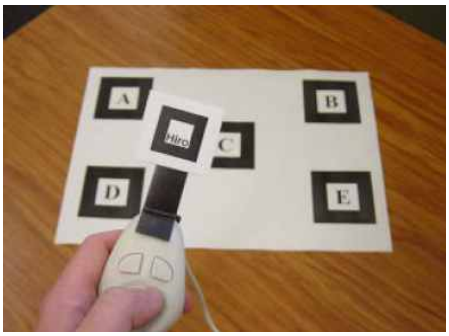
such as pitch, distortion, amplitude, and so on. Using the records enabled users to map musical contents to the controllers in an intuitive way. However, users could only compose their own music phrases when they had the marker-attached records.

The Tiles system (Poupyrev *et al.*, 2001a) is a collaborative Mixed Reality (MR) TUI that is based on a metal white board. Several users wearing a single camera-attached Head Mounted Display (HMD) stood around the physical working space, the white board, and interactively arrange the marker-attached tiles to create their own MR scene (see Figure 2.4). The system allowed users to add, remove, copy, duplicate and annotate virtual objects on top of each tile. The augmented tile could be placed anywhere in the 3D physical workspace. Additionally, users were able to put physical annotations on the virtual objects by writing on the white board. Although they only showed a prototype of aircraft instrument panel, the Tiles system might be easily used to create many other applications.

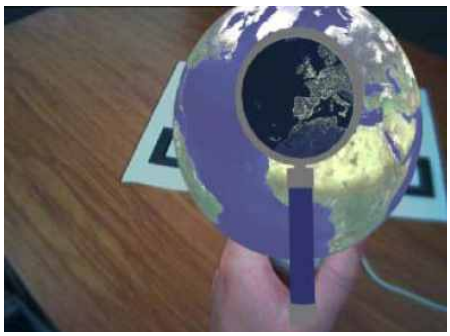


(a) (b)

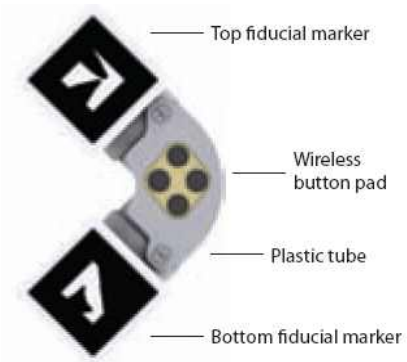
Figure 2.4 The Tiles System: (a) real environment with marker attached tiles and (b) a user's AR View



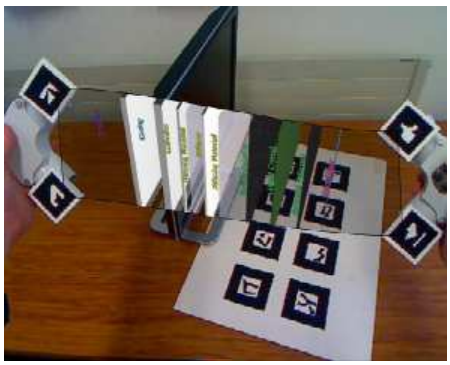
(a)



(b)



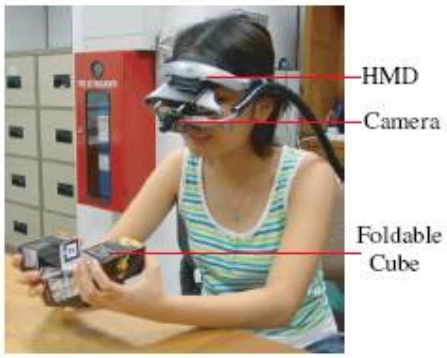
(c)



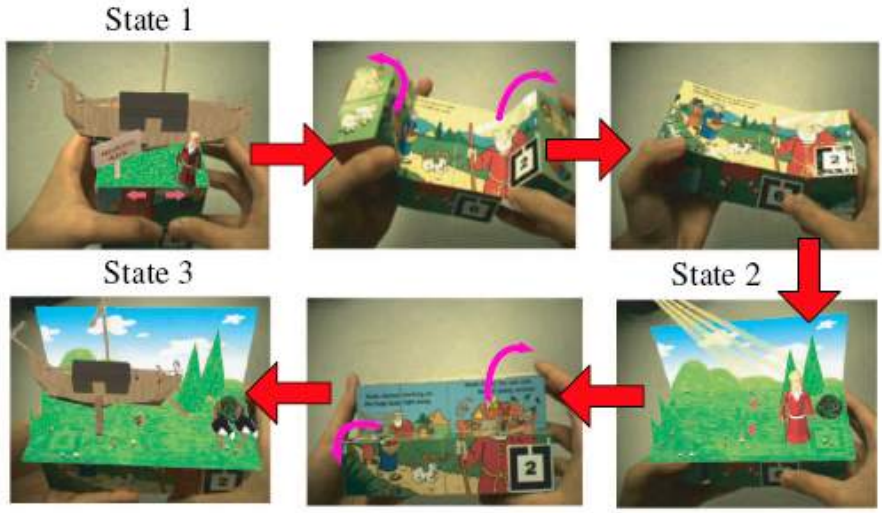
(d)

Figure 2.5 AR Magic Lenses: (a)(c) Two hardware configurations of AR Magic Lenses , (b)(d) AR View with AR Magic Lenses.

The AR Magic Lenses (Looser, 2007) is another example of an AR TUI (see Figure 2.5). The AR Magic Lenses allowed users to magnify an augmented object, to browse global datasets, the internal structures of the augmented 3D objects, and to access to additional layers of information through different software applications; however, the AR Magic Lens did not provide direct manipulation of the augmented virtual objects.



(a)



(b)

Figure 2.6 Magic Story Cube - (a) Physical setup and (b) state transition of the storytelling. (Zhou et al., 2004)

The Magic Story Cube (Zhou *et al.*, 2005) is a foldable cube for 3D mixed media storytelling interface (Figure 2.6). The system let a user unfold the cube in a unique order. As a result, the system provided a different stage of story corresponding to the different cube states. This enabled users to have continuous storytelling by unfolding the cube using their two hands. Using a

physical cube would be attractive to users and provided a new way of exploring a story interactively; however, the Magic Story Cube only supported interaction with sequence of the story, not the contents of the story. In follow-up work (Zhou *et al.*, 2004) they enhanced interaction by supplying new functions, such as moving, resizing, and deleting augmented objects. The system still had the same disadvantages of TUIs which were the user had to carry their interaction devices or physical objects with them to interact with the augmented virtual objects.

Some of the lessons learned from this earlier research include:

- Most AR TUIs are based on marker-based tracking technology to get the position of the tangible object relative to the virtual objects that the user can interact with.
- AR TUIs provide easy and fast input in an AR interface.
- AR TUI physical objects provide tactile feedback
- User often has to carry special tangible objects for input

2.3.2 Hand gesture

As we learned in Section 2.1.1, one of the disadvantages of AR TUIs was that a user had to carry the interaction tool. To overcome the limitation of the AR

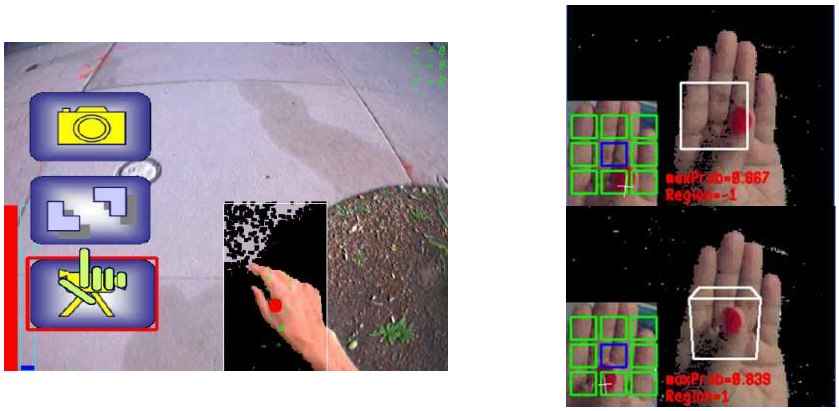
TUIs, researchers were developing hand-based interfaces and a wider range of input devices.

For example, the software architecture for the wearable AR system, Tinmith (Piekarski & Thomas, 2001), was designed to support developing AR applications with trackers, input devices, and graphics. Interfaces for the wearable AR applications had evolved according to the development of its hardware and software system. They developed Tinmith-Hand (Piekarski & Thomas, 2002), a unified user interface technology for mobile outdoor AR and indoor VR using 3D interaction techniques. A pinch glove with an ARToolKit marker on the thumb for tracking was used to control a menu and a 3D modelling system. The interaction with the Tinmith-Hand was done through head and hand gestures. Head tracking was for an eye cursor to specify objects and planes along the line of sight relative to the body. Hand tracking was done in two ways: a one-handed cursor was used for both selection and translation: and two-handed cursors were used for multiple selections and relative rotations and scaling. Although the Tinmith-Hand was designed to leave the users' hands free from input devices, a user had to wear a marker-attached pinch gloves all the time.

(a) (b)

Figure 2. 7 Tinmith system: (1) Tinmith architecture, (2) Tinmith-Hand

HandVu (Kölsch, 2004) is a computer vision-based software library which can be used to build a hand gesture interface by detecting a standard hand posture, and recognizing key postures in real-time without camera or user calibration (see Figure 2. 8(a)). Hand detection for the HandVu used Violar and Jones's Method (Kölsch & Turk, 2004) and the hand tracking used Flocks of Features and Multi-Cue integration (Kölsch & Turk, 2004a). The HandVu provided fast 2D natural hand interaction without additional devices, such as data gloves or colour markers. However, the posture of the hand was limited to certain shapes to improve the accuracy of the hand detection algorithm. Moreover, the hand interaction was done in 2D. As a result, direct manipulation with the augmented virtual objects was limited.

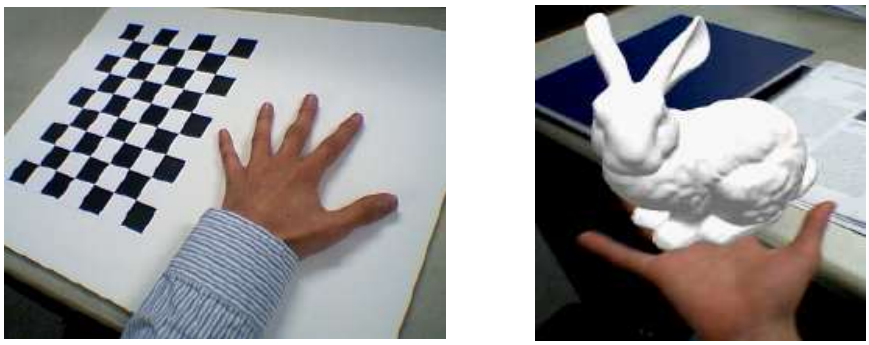


(a) (b)

Figure 2. 8 Natural hand interface examples: (a) Hand Vu (Kölsch, 2004) and (b) ThumbStick (Man *et al.*, 2005)

ThumbStick (Man *et al.*, 2005) is a hand gesture interface for a wearable AR environment with the complex background (see Figure 2. 8(b)). A user wearing a Head Mounted Display (HMD) with camera attached had an AR view through the HMD. By moving his/her thumb as a pointer, the user could interact with virtual objects. The other four fingers were used as a control region. The user had to paint his/her thumbnail in red to capture the centre of the thumbnail as a pointer. There were nine control regions shown in the bottom left corner of the screen. As the thumb entered the four finger regions, the user could control the augmented object in five degrees of freedom. However, the hand interface worked exactly as the mouse does and so did not provide very natural 3D object manipulation.

Lee and Höllerer (2007) have developed Handy AR, which enabled a user to use his/her open hand as a marker, placing a virtual object on top of the hand. In the off-line calibration step, a user constructed a hand pose model by measuring fingertip positions relative to a checker board which was placed next to the open hand. The origin of the coordinate system was translated to the centre of the hand. The users could change the view of the augmented objects by moving around their hand. As a result, a user did not need to carry around fiducial markers to get an AR view. However, the hand was used as an interface to augment a virtual object, not to manipulate the augmented object. Additionally, the hand posture had to be fixed in an open hand shape.



(a) (b)

Figure 2. 9 Handy AR: (a) A hand model construction by putting the hand next to the checkerboard pattern and (b) Augmenting a bunny model on top of the user's natural hand

Some of the lessons we learned from previous research with Hand Interfaces in AR include:

- The user did not need to carry around interaction tools like AR TUIs.
- Most of hand interfaces required users to wear markers or the user had to fix their hand posture.
- Interactions with the user's hand were done in 2D. Thus, it was more like mouse input.
- The interaction with the hand was limited to a few functions.

2.3.3 Multimodal AR Interfaces

The functions the hand interface provides is limited and users have to wear a marker or to have a fixed hand posture. Using speech to provide an additional input modality to the hand interface will overcome the limitations of gesture input alone. Although we studied previous MMI with 2D/3D graphics environment in Section 2.2, now we review earlier MMIs that have been used in AR applications.

There has been some earlier work in applying MMI in AR applications.

(a) (b)

Figure 2. 10 Multimodal systems: (a) SenseShapes (Kaiser *et al*., 2003) and (b) Multimodal interface in AR scenario(Heidemann *et al.,* 2004)

Kaiser *et al*. created SenseShape (2003); a multimodal AR interface in which volumetric regions of interest that are attached to the users' eyesight or hand to provide visual information about interaction with augmented or virtual objects. SenseShapes increased the predictability of object selection. Object selection was available with the help of multimodal AR interface based on statistical data sets. Speech recognition provided information where the user wanted to move the object, by using words such as "this" or "that". However, a user must wear a data glove to detect users' gestures and 6 DOF trackers to monitor hand position for interaction with objects. Kaiser *et al.* also did not conduct user studies to measure the effectiveness of their system.

Heidemann *et al.* (2004) developed a prototype of a situated intelligent system with multimodal interfaces for information retrieval in AR. It supported online

acquisition of visual knowledge and retrieval of memorized objects. To achieve this, an inertial sensor on the top of users head was employed to watch the users' movement. Hand gestures and speech were adopted to move between menu options. Two cameras on the users head and computer vision software was used to recognize when the user moves their hand underneath the target menu and was adopted to record skin colour samples with a voice command according to the change of lighting of the environment. However, the two cameras were not aligned to the same position of users' eyes, and the video of the real environment was offset. So it was not an easy way for users to interact with it. In addition, the system only supports 2D menu navigation, and the speech input was used to select the menu item that the user wanted to choose, in the same way a mouse did; thus, the system did not use multimodal input fully.

Irawati *et al.* (2006a) developed a computer vision based AR system with a multimodal interface. They extended the VOMAR application by adding support for speech recognition. The final system allowed a user to pick and place virtual furniture in an AR scene using a combination of paddle gestures and speech commands. A semantic fusion method was used to improve the recognition rate more than a simple time-stamping method. They also conducted a user study (Irawati *et al.*, 2006), which verified that combined

multimodal speech and paddle gesture input was more accurate than using one modality alone. However, the system could not provide a natural gesture interface for users, and required the use of a paddle with computer vision tracking patterns on it.

Lessons we have learnt from the previous research on AR MMI include:

- that MMI provided an effective and easy way of interaction in AR environments.
- that there has been little research on MMI in AR.
- that user studies on AR MMI were not fully explored.

2.4 Previous MultiModal Fusion Architectures

The main difference between a unimodal interface and a multimodal interface is that the multimodal interface requires a multimodal fusion architecture to merge two or more modality input in an efficient and effective way. Multimodal fusion systems can be classified in two groups: (1) feature level fusion and (2) semantic level fusion (Oviatt *et al.*, 2000).

Feature level fusion is done before the input signals are sent to their respective recognizers. Feature level fusion is considered as a good strategy for integrating the closely coupled and synchronized input signals, for example,

lip movement and speech input (Wark *et al.*, 1999) whose signals correspond to each other. Typical drawbacks of the feature level fusion are that it is complex to model, intensive to compute, and difficult to train. Mostly, feature level fusion requires a large amount of training data.

Semantic level fusion is done after the signals are interpreted from their respective recognizers. Semantic level fusion is appropriate for integrating two or more signals which provide complementary information, such as, speech and pen input (Cohen *et al.,* 1997). Individual recognizers are used to interpret the input signals independently. Those recognizers can be trained with existing unimodal training data.

For our multimodal interface with gesture and speech input, semantic level fusion needs to be adopted to integrate two input signals. Thus, in this section, we will concentrate on previous works in semantic level fusion.

2.4.1 Semantic level fusion

Johnston *et al.* (1997) proposed a unification-based multimodal integration. The integration strategy was designed based on Oviatt *et al.*'s (1997) user observations of subjects using pen-based gesture and speech commands. Johnston *et al.* represented the recognition results of each modality in typed feature structures to translate them into as commands for any interfaces. In their fusion approach, the integration of speech and gesture was decided by

using two factors: (1) tagging of input as either complete or incomplete and (2) time-stamping. Integration was done when speech or gesture was marked as incomplete and speech followed gesture within a time window of three to four seconds. To use their integration architecture, typed feature structures for all of the possible commands had to be predefined.

Johnston (1998) proposed another fusion architecture which has multi-dimensional parser: first, N-best recognition results were listed: second, a single spoken utterance has one or more associated gestures by using temporal constraint and spatial constraints; at the end, a time-stamping method and unification of typed feature set are adopted to finalize the fusion process.

Medl *et al*. (1998) proposed a slot-filter method to integrate speech, hand gesture, and gaze input in a monitor-based 2D graphics application. 'Frames' contain information about commands and 'Slots' include name of the principal objects name in a frame. Their multimodal fusion was done 'first-come first-serve' based rule. However, there was no synchronization among modalities. As a result, errors from multimodal integration were high.

Rauschert *et al.* (2002) proposed the DAVE-G system (Rauschert *et al.*, 2002) which had free hand gestures and speech as input channels. The multimodal

fusion was based on the time analysis of incoming signals. Extracted features from the speech and gesture signals were used to measure co-occurrence.

Sharma *et al.* (2003) proposed a multimodal integration algorithm based on a time stamp and a searching window. They proposed two different types of semantics: (1) static and (2) dynamic. Static semantics are the place where the knowledge can be stored. In general, the semantics of language, user knowledge, and world knowledge are dealt with static semantics. In task or domain specific cases, user models, tasks, and structures are stored in static semantic forms. Dynamic semantics are the place where current states of the interaction are stored. Discourse history, attentional states will be represented in dynamic semantics in general.

Kaiser *et al.* (2003) developed a multimodal fusion architecture for speech, gesture, and head tracking input for SenseShape (Olwal *et al.,* 2003). Their fusion architecture was also based on time-stamps. The N-best candidates of the objects that were referenced at that time had been listed, and the gesture recognition results filtered by the speech recognition results. They also adopted mutual disambiguation to improve the error avoidance and resolution. The mutual disambiguation in a multimodal architecture is a particular advantage of multimodal interface over unimodal interface, and provides superior error handling. Kaiser *et al.* extended Johnston's architecture to

handle 3D hand gestures instead of 2D pen-based input of the QuickSet (Cohen *et al.*, 1997). They took advantage of additional 3D sources of information such as object identification, head tracking and visibility.

Another fusion architecture for fusing 3D gesture and speech input was proposed by Irawati *et al.* who used an ontology for semantic integration in 3D MMI (Irawati *et al.*, 2006b). They proposed a multimodal interaction in virtual environments to integrate several input modalities to an interaction command in the virtual world. Ambiguities from the users' commands were solved by using the spatial ontology that included the information about virtual objects and described the spatial relationships between the virtual objects. However, in their work, it was not clearly described how they integrated different modalities using the spatial ontology. Additionally, it was not mentioned how they used the time stamp to merge several input and what kind of semantic representation of the recognized input was adopted.

Some of the key lessons learnt from semantic level fusion include:

- that input channels needed to have complementary information to each other.
- that time-stamp played an important role to match two different modalities for integration.

- that semantic representation of the recognized input was essential for multimodal fusion
- that mutual disambiguation was necessary to improve error handling and resolution.
- that user-observation to learn users' interaction pattern was useful to design a unification method.

2.5 Limitations of Prior AR MMI Research

In the first part of this chapter, we studied related research in several AR interfaces: (1) TUI, (2) hand interfaces, and (3) MMIs.

TUIs were the most popular interfaces in the early stage of AR interface research. With the help of the ARToolKit, any physical object could be a controller to interact with augmented virtual objects by attaching fiducial markers on the physical tools to track the pose and orientation of the tools. However, a user needed to carry around the physical objects to use them as an interaction tool in the AR environments. To overcome the limitation of the AR TUIs, natural hand-based interface have been considered.

Hand interfaces did not require users to carry a physical object for an interaction. Instead, in most of research, a user had to wear markers or had to

fix their hand posture. Moreover, the hand interfaces did not support direct manipulation because the interactions with the users' hand were done in 2D. Thus, we need to have a hand interface which does not require users to wear markers or datagloves, and also which supports user interactions in 3D. Morever, hand interfaces were not able to support descriptive commands, such as changing colours or shapes of the target object. MMIs have a fast and accurate way of interaction by letting users have two or more input channels. Users can combine different modalities to deliver their commands to the system in an efficient way. As we observed earlier, the combination of hand interface and speech would be useful for interactions in AR environments.

However, there is no AR interface research which provides natural hand interaction in 3D with corresponding speech input. Additionally, user studies on AR MMI are not fully explored, and there has been little study of fusion architectures which are designed according to the users' interaction behaviour in AR MMI environments.

2.6 Proposed Method

To overcome the limitations mentioned above, we will (1) develop an AR MMI and (2) evaluate usability of the AR MMI.

2.6.1 Multimodal AR interface development

To develop a multimodal interface in AR environment, we first need to implement each of the following components:

- vision-based hand gesture recognition
- speech recognition
- a semantic multi-channel signal fusion architecture

2.6.1.1 Vision-based gesture analysis

For vision-based gesture input, stereo vision tracking can be used to find out hand position and pose in 3D for free or natural hand interaction. At first, a rough hand position will be obtained by using a centre of mass algorithm, then, depictive gestures using second moments will be implemented. An occlusion-free AR view is very important for providing a natural sense of interaction to users. This will be done by considering the 3D position and pose relative to the augmented object.

2.6.1.2 Speech Recognition

Speech Recognition will be used to give commands directly to the system. We will use the Microsoft Speech API (SAPI) for speech recognition.

2.6.1.3 Semantic multi-channel signal fusion architecture

MMIs, unlike unimodal interfaces, require having a multi-signal fusion architecture to merge two or more input commands in a natural and efficient way. We should have a history of each mode of signal. With the analysis of each signal, statistical characteristics will be obtained. Then, multi-channel signal fusion is available with the provided statistical characteristics. Additionally, environmental context and task context should be considered to provide better recognition result.

2.6.2 User study to evaluate multimodal interface in AR

User evaluations for verifying the usability of the implemented AR MMI is necessary. Video analysis is important for analysing user behaviour, such as how they use speech and gestures together, how they interact with the objects, and missed functions of the system to be a natural interface for users.

We will run three user studies; (1) a user observation to see how users interact with virtual objects in an AR environment, (2) another user observation to closely observe how different types of gestures were used to interact with various objects, and (3) a full usability test to evaluate the usability of a complete MMI application. Quantitative measurements are necessary to objectively evaluate the proposed MMI, such as, how many errors are occurred during the experiments. Afterwards, we will give questionnaires to

users to check users' satisfaction with the system, the impact of interface, involvement with the task, and awareness and distractions.

Chapter 3

User Observations I – Wizard of Oz Study

3.1 Introduction

To build a user-centred AR MMI, we need to observe both how users will want to input multimodal commands and how different AR display conditions will affect these commands. This can be accomplished through a Wizard of Oz (WOz) study where the users' commands are interpreted by a human 'Wizard' who controls the interface and gives the illusion that the application is capable of perfect speech and gesture recognition.

This chapter describes the results of user observation with the WOz method. Observed data includes the frequencies of speech or gesture commands, the time for speech and gesture commands, and the time gap between combined speech and gesture commands. In addition, there are also findings by watching users from recorded video. Finally, we interview each subject after completing the experiment tasks.

WOz methods have often been used for prototyping speech and gesture recognition systems in the past; however, there has been no research on using WOz user study to explore natural human behaviour in a multimodal AR interface.

3.2 Related work

Salber and Coutaz (1996) provided a good overview of how WOz techniques can be applied to multimodal interfaces. Their NEIMO system (Coutaz & Salber, 1996) used these methods in a multimodal usability laboratory for evaluating 2D user desktop interfaces. They observed users using MMI with a mouse-based application along with simulated speech recognition and interpretation of facial expressions. Through the observation of users' behaviour, they identified users' needs when they use the MMI relative to the given tasks. There are many other examples of how WOz techniques can be used for system prototyping in various research areas. For example, Oviatt *et al*. (1992, 1994) have shown the value of using high-fidelity WOz simulations in comparing speech-only, pen-only, and combined speech-pen input modalities in a variety of applications such as checking bank accounts or using maps.

Most relevant to our work is the use of WOz studies with multimodal input in 3D graphics applications. For example, Hauptmann (1989) provided an early example of using a WOz technique to simulate multimodal interaction with a 3D graphics environment; in this case rotating blocks on a screen. He found that users typically used short spoken commands and that gesture input was the preferred method for manipulating the blocks. Corrdini and Cohen (2002) described using a WOz technique for navigating through a 3D virtual environment. Molin (2004) made a WOz prototype for cooperative interaction design of graphical interfaces. After this WOz study, Molin concluded that the WOz experience triggered an analysis of the interaction which produced new design ideas that could be tested, and the recordings of screen and video could provide clarification and examples of good or bad design.

As can be seen, there have been few examples of multimodal AR interfaces, and none have used computer vision techniques for 3D natural hand interaction with speech input. There has also been very little evaluation of AR multimodal interfaces in general, and no previous studies that have used a Wizard of Oz technique.

The research in this chapter is novel because it uses computer vision to support natural hand input in a multimodal AR interface for 3D object manipulation. Most importantly, it is the first WOz user study in a multimodal AR interface.

We are interested in both how users will want to input multimodal commands as well as how different AR display conditions will affect these commands. This research is essential to design a user-centred AR MMI and will be useful for others trying to develop multimodal AR interfaces.

3.3 Proposed solution and Experimental setup

We have developed an AR system that combines 3D vision based hand tracking with simulated speech input and screen-based and hand held display (HHD) AR output. We have also developed a simple command trigger tool for supporting the WOz experiment. In this section we describe our system in more detail. Figure 3. 1 shows how the system components are connected.

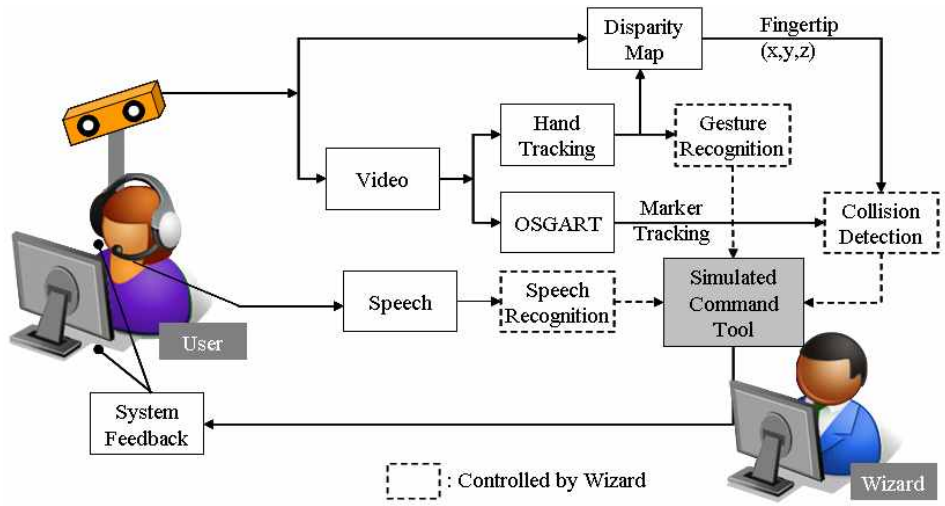


Figure 3. 1 Software components of the proposed AR WOz system: User input is analysed and triggered by the Wizard using Simulated Command Tool.

From previous research (Hauptmann 1989, Corrdini & Cohen 2002, Molin 2004) an ideal Augmented Reality WOz study should have the following attributes:

- A tool for capturing user input for later analysis.
- The ability to observe the frequencies of each gesture or speech command (which command and how often) and the time window size needed to detect related speech and gesture.
- Support for remote control from the WOz expert user.
- An interview exploring how users feel about multimodal input and different display types.
- Several experimental conditions for comparing speech and gesture input in.
    - 2D, 3D, 2D/3D mixed environments
        - Changing characteristics of objects
            - Colour, shape
        - Manipulating objects
            - Pick up, Drop, Delete

The study we have designed satisfies each of these attributes. In addition we developed a method for computer vision hand tracking, a WOz command input tool and an AR viewer as described in the next sections.

3.4 3D Natural hand interface

It is not easy to simulate normal 3D natural hand interaction in real time in a WOz application. Thus, we have implemented a 3D vision-based hand tracking system (Figure 3. 2). Our hand tracking is based on three methods: (1) segmenting skin colour, (2) finding feature points for the centre of the palm and fingertips, and (3) finding the hand direction. We used a BumbleBee2 stereo camera (2009) and our software is based on the OpenCV library (2009).

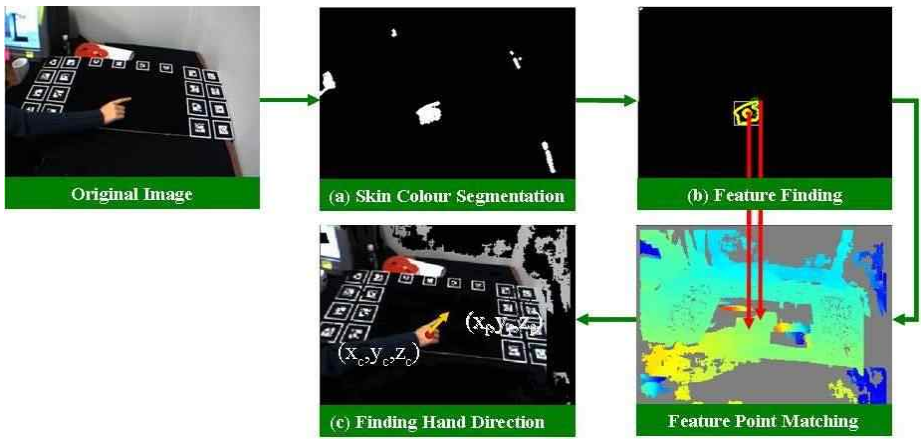


Figure 3. 2 Our 3D Natural Hand Interface: (a) Segmenting skin colour, (b) Finding feature points for palm centre and fingertips, and (c) Finding hand direction.

The user's hand is found by detecting skin colour in the input video images. We converted the camera image from RGB values into the HSV colour space which is more robust against lighting changes (Zhu *et al.*, 2000). We then used a sample skin image and its histogram of the hue plane to find out the proper threshold value to extract just the user's hand region.

After the skin colour segmentation, we find the biggest contour (Freeman, 1974) of the segmented area to extract the user's hand more accurately. Afterwards, a distance transformation (Borgefors, 1986) is performed to find the centre of the palm which is the farthest point inside the contour. Next we find the candidate's fingertips and the farthest fingertip from the palm is used to calculate the direction of the user's hand. The positions of two feature points, the centre of palm and the fingertip, are mapped to a disparity map to estimate the 3D information of each point for AR interaction.

We were able to track the user's fingertip with accuracy from 3mm up to 20mm depending on the distance between the user's hand and the stereo camera. The frame rate was 11-13 frames per second. The accuracy and the frame rate were enough to support our tasks in real time.

3.5 The Simulated Command Tool

We also created tools for WOz input. A command menu interface was written to provide simulated speech or gesture input for when users gave commands to the application. A human expert sat out of sight behind the user and entered commands in response to the user actions in the AR system. Figure 3. 3 shows the dialog menu used by the Wizard to quickly input commands. It has three functions for gesture commands ("pick-up", "drop", and "delete"), and two groups for speech: "change colour" and "change shape".

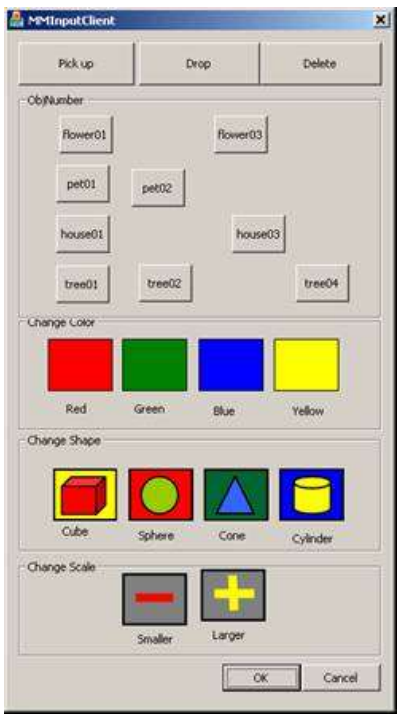


Figure 3. 3 The Simulated Command Tool: Three functions for replacement of gesture commands ("pick-up", "drop", and "delete"); Two groups for speech: "change colour" and "change shape".

3.6 The Augmented Reality View

To provide an AR view we used the osgART rendering and interaction library (Looser *et al.*, 2006) which includes the ARToolKit (Kato & Billinghurst, 1999) computer vision tracking library to track the user's real camera position relative to square fiducial markers. Once the camera position is known, osgART can create a 3D graphics scene which is overlaid on the live video view to create an AR view. We added lighting and shadow effects to improve the realism of the AR scene.

3.7 User study setup

In our research we wanted to use a WOz interface to explore the type of speech and gestures people would naturally use in a multimodal AR system. We were also interested in testing if different AR display conditions would have any effect on the multimodal input pattern. In this section we describe our experimental set up and tasks, while in the next section we present the results.

3.7.1 Experiment setup

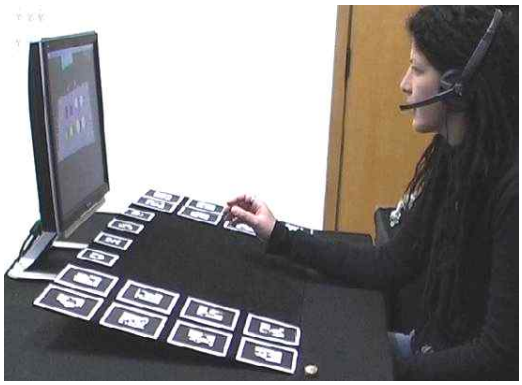
The primary goal of the experiment was to investigate the speech and gesture input and the time window for fusing speech and gesture input. The secondary

goal was to explore how the display or the task types affected the user's multimodal commands. Through interviews, the subjects were asked which interface they preferred and how easy they found it to complete the task, etc. We declared hypotheses of the study as following:

- H1: Different types of tasks lead to different usage of speech and gesture commands in multimodal interfaces.
- H2: Different types of tasks lead to different patterns of multimodal time windows.
- H3: A multimodal interface is preferred by the users compared to speech-only or gesture-only conditions
- H4: A multimodal interface is easy to interact with compared to speech-only or gesture-only conditions.
- H5: The display type affects the interaction pattern of multimodal interface.

There were 12 participants in the experiment (2 females and 10 males) with ages from 23 to 49 years old and an average age of 30.5 years old. The users completed three tasks in each of two display conditions; a screen display (Figure 3. 4(a)) and a Hand Held Display (HHD) (Figure 3. 4 (b)). We had to

attach a stereo camera on the front of a Head Mounted Display (HMD) (a widely adopted AR display). The stereo camera was too heavy to be worn attached to HMD, so we used Hand Held Display as another display.



(a)

(b)

Figure 3. 4 System Display Configurations: (a) Screen-based AR system and (b) Hand-Held Display-based AR System.

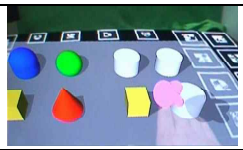
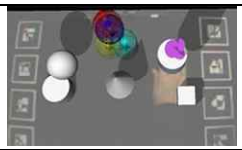
The HHD was custom hardware created from a display module of an e-Magin head mounted display (800x600 pixel resolution and 30 degree field of view) and BumbleBee2 camera attached to a handle. The screen display condition

involved the user looking at a 21 inch LCD screen with 1024 x 768 pixel resolution while the BumbleBee camera was fixed to show a view of the workspace in front of it. This view from the BumbleBee camera placed on the users' right side was combined with 3D virtual image overlay to create an AR view shown on the screen. The screen was placed about 80cm away from the user. The simulated command menu (see Figure 3. 3) provided users with the impression that the system had perfect speech and gesture recognition. We provided a different order of tasks and display conditions to each user to avoid learning effects using a Latin Square method (6 x 6).

3.7.2 Experimental tasks

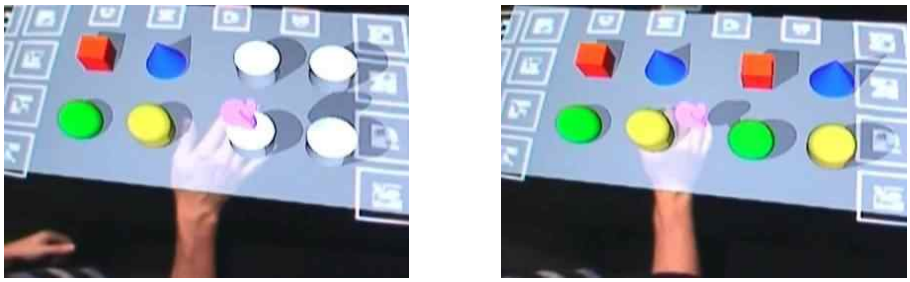
The experiment consists of subjects performing three simple tasks involving virtual object manipulation. Most interaction in an AR environment involves one or more of; moving virtual objects, rotating or translating virtual objects, or changing object colour or shape. Thus, we designed our tasks to include these interactions. In particular, each task included different dimensions of interaction spaces (2D, 3D, 2D/3D). The available interaction sub-tasks are shown in Table 3. 1.

Table 3. 1 Task Types and Available Interaction Modes in Different Dimensions

	Task I	Task II	Task III
			
Changing colour	O	O	×
Changing shape	O	×	×
Selecting object	2D	3D	2D/3D
Moving object	2D	3D	2D/3D

3.7.2.1 Task I

For the first task the system showed a set of simple AR primitive objects appearing on the table in front of the user, displayed over video of the real world (see Figure 3. 5). The users were supposed to change the colour and shape of four white cylinders, placed on the right side of a user, to the same shape and colour of target objects, which were placed on the left side of the user. Subjects needed to let the system know the colour or shape of which object they wanted to change. However, they could not change the position of any object displayed.



(a) (b)

Figure 3. 5 Task I: (a) initial view for the task and (b) completed view after user interactions.

In this case, the virtual objects were positioned on a table so gesture input was a largely 2D task where users would touch or point an object and say a shape and colour. Thus, the gestures which would be used in this task were almost 100% deictic gestures.

3.7.2.2 Task II

The second task involved moving sample objects distributed in 3D space into a final target arrangement of objects. The subjects needed to move their hands in all three directions to select and move objects. Figure 3. 6 shows the system recognizing a user's hand in 3D. When the user's hand is located within the object, then the system recognizes it as a collision and the object is rendered in wireframe. Once an object is selected the user must arrange the piece in the same layout as the final target configuration.

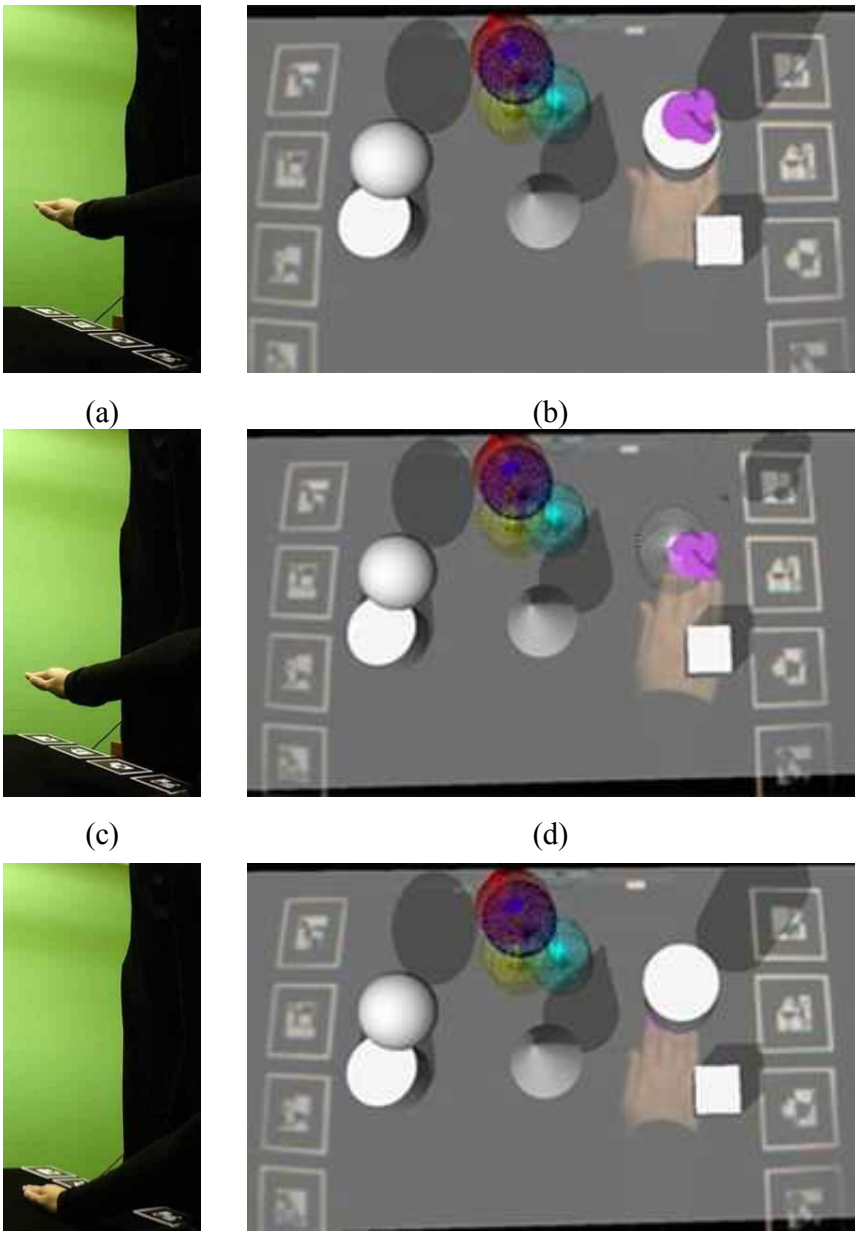


Figure 3. 6 Task II - 3D interaction with AR objects: (a)(b) when the user's hand is located on top of the object, (c)(d) within the object, and (e)(f) under the object.

3.7.2.3 Task III – Scene assembly task

The final task was to create an AR scene with detailed models instead of simple primitives. Using the models, subjects were told to create their own AR scene, using any gestures and or speech commands. The subjects used their gestures to move the models in 2D or in 3D. For example, dragging it on the table surface is a 2D interaction, and picking up the model and moving in space is a 3D interaction. The users were also asked to use their speech input to select the objects or to drop the objects to the target area.

3.8 Result and Analysis

Video data of user interaction was collected from each of the task conditions for all subjects. The collected video was analysed by a single observer. An independent video analysis with multiple observers would have been more reliable with an established protocol for the analysis. However, we only had a single observer because of time limitations. From this we counted the frequencies of speech or gesture commands to see which were used and how often they were used. We also analyzed the time for speech commands, gesture commands, and the time gap between combined speech and gesture commands. In addition, there were also findings by watching users from recorded video. Finally, we interviewed each subject after completing the

experiment tasks. After the experiment, the recorded video is analysed by a single observer to save time to train multiple observers to annotate the recorded video.

3.8.1 Frequencies of Speech

From the video data we analyzed the users' speech based on the number of following types of words used; colour, shape, deictic, and miscellaneous (misc) commands. The group of deictic words includes pointing in a direction, using "*here*" or "*there*", and pointing to an object, using "*this*" or "*that*". For example, a phrase "*Pick this*" consists of a misc word (pick) and a deictic word (this).

Table 3. 2 shows the number of words spoken in the experiment broken down by categories and tasks. Across all tasks subjects used a total of 1232 words (612 words with the screen display and 620 words with the HHD). According to our analysis, 74% of all speech commands were phrases of a few discrete words, and only 26% of commands were complete sentences. On average the phrases used were 1.25 (std=0.66) words long and the sentences used were 2.94 (std=1.08) words long. There was no significant change in speech patterns over time.

Table 3. 2 The numbers of words used for speech input: colour, shape, deictic, and miscellaneous speech commands with different display types and different task types.

Task	Display	Deictic	Colour	Shape	Misc.	Total
Task1	Screen	36	83	86	33	238
	HHD	29	47	87	50	213
Task2	Screen	26	61	11	80	178
	HHD	48	62	14	107	231
Task3	Screen	58	13	31	94	196
	HHD	41	19	29	87	176
Total		238	285	258	451	1232

3.8.2 Gesture Frequency

Table 3. 3shows the numbers of gestures used. The subjects used a total of 926 gestures (495 with screen display and 431 with HHD). We found that main classes of gestures were deictic (65%) and metaphoric (35%) gestures.

Table 3. 3 Numbers of gestures

Task	Display	Deictic	Metaphoric	Beat	Iconic	Total
Task1	Screen	72	0	0	1	73
	HHD	61	0	0	0	61
Task2	Screen	122	90	3	0	215
	HHD	124	57	0	0	181
Task3	Screen	112	94	1	0	197
	HHD	106	83	0	0	189
Total		597	324	4	1	926

From the experiment video we analyzed users' gestures according to the gesture classification scheme of McNeill (1992) (Deictic, Metaphoric, Iconic, and Beat-like gestures). The classifications of the gesture are:

- Deictic gesture: mainly pointing.
- Metaphoric gesture: representing an abstract idea.
- Iconic gesture: depicting an object.
- Beat gesture: formless gestures, utterance rhythm.

3.8.3 Speech and Gesture Timing

In addition to counting speech and gesture events we also wanted to investigate the relationship between speech and gesture input in creating multimodal commands. We wanted to identify the optimal time frame for combining related gesture and speech input based on the users' natural response. This is important because the size of the time frame may affect not only the accuracy of the multimodal fusion but also the system delay.

To do this we measured the Multimodal window, a time frame that contained the combine gesture and speech input as shown in Figure 3. 7. This is made up of:

- *Gesture Window*: how long the user holds a particular gesture for.
- *Speech Window*: how long it takes the user to issue the speech command.
- *Front Window*: the time delay of the speech input before (-) or after (+) the corresponding gesture input.
- *Back Window*: how long the user held their gesture after their speech input was finished.

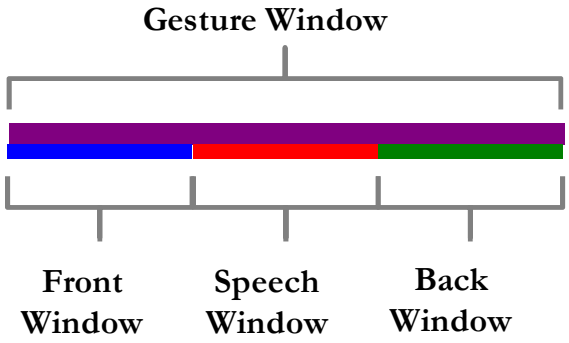


Figure 3. 7 The definition of Multimodal window: (a) Gesture Window, (b) Speech Window, (c) Front Window, and (d) Back Window.

By viewing the videos of the user interaction we could measure the time difference between when the subject issued related speech and gesture commands. We analyzed the size of windows to improve the accuracy of input in a multimodal interface with a multimodal signal fusion architecture. The mean multimodal windows for each task with different display types are shown in Figure 3. 8.

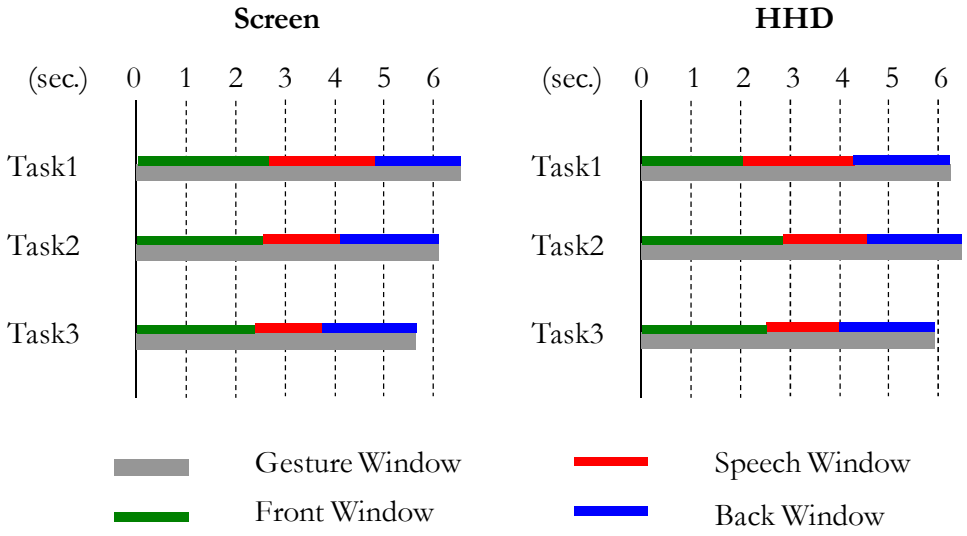


Figure 3. 8 The mean multimodal window (in seconds) for each task with different display types.

We realized that if we took mean values of each window, a lot of data would be missed and so the accuracy of multimodal input would be reduced. Thus, we decided to take the time window which covers 98% of data set. The mean size of the gesture time window which covers up to 98% of gesture time windows was 7.9 seconds (std=1.20), the mean size of the speech time window was 2.6 seconds (std=1.41), the mean size of the front window was 4.5 seconds (std=1.46), and the mean size of the back window was 3.6 seconds (std=1.13). Each window size with different task and display conditions is shown in Table 3. 4.

Table 3. 4 The optimal multimodal window (in seconds) for each task with different display types.

	Display	Gesture Window	Speech Window	Front Window	Back Window
Task1	Screen	7.636 (1.670)	3.091 (1.700)	4.182 (1.328)	2.727 (0.786)
	HHD	7.200 (1.550)	3.300 (1.418)	2.800 (1.033)	3.400 (1.174)
Task2	Screen	8.333 (1.970)	2.583 (1.564)	4.750 (1.288)	3.833 (1.337)
	HHD	8.909 (2.468)	2.727 (1.555)	5.273 (1.618)	3.546 (0.934)
Task3	Screen	7.400 (1.265)	1.900 (0.876)	4.700 (0.949)	3.900 (1.197)
	HHD	7.800 (1.229)	2.100 (0.738)	5.100 (1.197)	3.900 (0.994)

We also found that gesture commands were almost always issued before the corresponding speech input in a multimodal command. Overall, 94% of the time gesture input came before the related speech input. Breaking this down for the three tasks, 94%, 92%, and 96% of gestures come before speech in tasks 1, 2, and 3, respectively. So in order to combine related speech and gesture commands, the final multimodal AR system should have a search window at least 7.9s long, and should look for related speech input issued on average 4.5s after the gesture command is made.

3.8.4 Dependences on task or display type

We used a two-factor (task type and display type) repeated measures ANOVA with post-hoc pair wise comparisons (with Bonferroni correction) to see how task or display types affected the numbers of words for each speech command type, the numbers of gestures for each gesture command type, and the window sizes of multimodal input windows.

3.8.4.1 Dependences of speech input

The numbers of words for colour ($F(2,10)=7.212$, $p= .012$), shape ($F(2,10)=19.843$, $p< .001$), and miscellaneous commands ($F(2,10)=9.520$, $p= .005$) differed significantly across task type. Post hoc multiple comparisons showed that task 1 was different from both task 2 and task3 with a higher number of words for shape. This was expected because only task 1 included changing the shape of the objects based on the target objects. The number of other words in task 1 was significantly different from task 2 ($p= .010$). Most of the words spoken in task 1 were about colour and shape. Moreover, users did not move any virtual objects in task 1, but did in task 2 and 3. In case of deictic words and number of words, no significant difference was found. None of the speech command type was dependent on the display type.

3.8.4.2 Dependences of gesture input

A two factor (task type, display type) repeated measures ANOVA with post-hoc pair wise comparisons (with Bonferroni correction) was applied to the gesture analysis as well to find out differences between the numbers of gestures depending on task or display type. There was a significant difference in the numbers of deictic gestures by task type ($F(2,10)=10.023$, $p= .004$). Task 1 was significantly different from task 2 ($p= .003$) because the gestures in task 1 were all pointing gestures. Therefore, compared with task 2 which included more other gesture types, task 1 had more deictic gestures than task 2. In case of metaphoric gestures, there was a significant difference across task type ($F(2,10)=13.676$, $p= .001$). Task 1 was significantly different from task 2 ($p= .001$) and task 3. Users did not use metaphoric gestures at all in task 1. However, we could not find a significant difference between task 2 and task 3. The number of gestures was significantly different by task type ($F(2,10)=119.207$, $p< .001$). Task 1 was different from task 2 ($p< .001$) and task 3($p< .001$). Task 1 was a simpler task than the other two tasks. Thus, the mean number of gestures in task 1 was significantly smaller than task 2 and task 3. There was no difference in gestures used depending on the display type ($F(2,10)=2.585$, $p= .136$).

3.8.5 Dependences of Speech and Gesture Timing

We also investigated how the window sizes of multimodal input changed according to task types or display types. There was no significant difference in the gesture window size among the tasks or between display types. In case of speech input, there was a significant difference between the phrase lengths in each task ($F(2,6)=8.145$, $p= .020$). Task 1 was different from task 2 ($p= .041$) and task 3 ($p= .025$). Task 1 had a longer speech timing window (mean=3.50, std=0.34) than task 2 (mean=2.69, std=0.35) and task 3 (mean=2.00, std=0.23). Task 1 was more descriptive, such as changing colour or changing shape, than task 2 or task 3. Thus, users gave longer commands to describe what they wanted to change. There was no difference between task 2 and task 3 and no significant difference in display type. We did not find a significant difference among tasks or between display types for the front time window size. However, there were significant differences in the back time window among task types ($F(2, 6) = 9.297$, $p< .015$). Task 1 showed a smaller size of the back time window than task 3.

3.8.6 Subjective Questionnaire

To get more information from users, we analyzed their feedback using a subjective questionnaire. We adopted Looser's Magic Lens Questionnaire to develop specific questions related to the task (Looser, 2007). In addition extra questions were adapted from the NASA TLX questionnaire to measure the

cognitive workload (Hart and Staveland 1998). The exact questions can be found in Appendix A. We asked users to score the naturalness of speech, gesture, and mixture of speech and gesture input on a Likert scale (1: disagree, 5: agree). The questions for the naturalness of interface were:

- *Q1: It was natural to use speech input in this task.*
- *Q2: It was natural to use gesture input in this task.*
- *Q3: I felt that it was natural to manipulate the virtual object with combined speech and gesture input.*

When users were asked whether they thought speech was natural, we got a mean score of 3.94 out of 5 (std=1.07). When asked if gesture was natural, users gave a mean score of 3.61 out of 5 (std=1.18). When we asked users whether the combination of speech and gesture was natural, they gave a mean score of 2.61 out of 5 (std=0.52). Using a two way ANOVA within subjects we found no significant differences between different task or display types in response to the questions about the naturalness of speech, gesture, and the combination of speech and gesture input.

We also asked users how helpful the speech and gesture input was with the following questions:

- *Q8: I think the use of speech helped me communicate descriptively with the system.*

- *Q9: I think the use of gestures helped me communicate spatially with the system.*

When users were asked whether they thought speech was helpful for the descriptive communication with the system, they gave a mean score of 3.96 out of 5 (std=1.11). When asked if gesture was helpful for the spatial description with the system, we got a mean score of 3.71 out of 5 (std=0.97). Using a two way ANOVA within subjects, we found no significant difference between task types or between display types in the above questions.

We also asked users how much physical demand, mental demand, and frustration were caused by the tasks and displays with the following questions:

- *Q10: I found using this technique was physically demanding.*

- *Q11: I found using this technique was mentally demanding.*

- *Q12: I found this technique frustrating.*

When we asked users whether the MMI was physically demanding, they gave a mean score of 2.61 out of 5 (std = 0.52). When asked if the MMI was

mentally demanding, we got a mean score of 2.44 out of 5 (std = 0.278). When the users were asked whether the MMI was frustrating, they gave a mean score 2.38 out of 5 (std = 0.57).

Using a two way ANOVA within subjects we found there was no significant difference between the physical demand rating for different display types, the screen display and the HHD, even though the HHD required the user to be holding something for the entire time. However, there were significant differences among the task types ($F(2,10)=14.809$, $p< .010$) from statistical analysis on the physical demand results. Task 1 was rated more demanding than the other two tasks.

In case of mental frustration issues, there were significant differences among task types ($F(2,10)=9.655$, $p< .005$), but there was no significant differences between display types.

After users finished their overall conditions, we also asked them to pick one display type based on their preference, enjoyableness, and ease of use. In total 66.7% of people preferred the screen display over the HHD and said it was more enjoyable, while 83.3% people said that it was easier to do the task with the screen display. According to the users' comments, the ease of watching and interaction was the main advantage of the screen display. No limitations of

movement, and being less physical demanding were other advantages. However, from the users' comments, we learned that the AR experience provided was not as immersive or compelling when the users were using the screen display.

On the other hand, users felt that the HHD provided a natural AR view because the view point of the camera was exactly same as where the users were looking. The novelty of the HHD was also attractive to users. However, the HHD did have a lot of disadvantages compared with the screen display. Holding the HHD was physically demanding and the tracking was not as good as the screen display because the camera moved around according to users' view. The users' interaction area was much smaller than with the screen display because the stereo camera on top of HHD required a minimum distance to calculate the 3D information of the user's hand for interaction.

These results show that display type does not affect physical demand, mental demand, or user frustration. However, users preferred the screen display over the HHD, and they felt it is enjoyable and easy to interact with the objects. Thus, screen display should be better than a HHD for a system design. Around 75% of users did not feel it was natural to talk to the computer.

3.8.7 Observations

We have several observations from watching the users do the experiment, such as considering the users' response to the Wizard's errors. First of all, when the Wizard did not react to their gesture commands properly, most of users repeated the same commands again to let the system respond to them properly. In case of speech commands, they tried to find out other commands for the system. In addition, when the Wizard made a mistake simulating the users' command, the users thought they did something wrong, not the system. The users sometimes wanted to know what they did wrong by asking the Wizard. In this sense, it may be better to provide a channel to let the users know which of their commands the system did not understand well. We also found that if the system did not have fixed commands the users may be initially frustrated. For example, a user said "What can I say?", then tried to figure out which commands were available, such as saying "Move the target. Does it work?". However, when they learned how the system worked, they improved their interaction speed. Moreover, they tried to explore which new functions were supported by the system. For example, one user said "Change the shape to a box." Although this changed the target object to the box, he still tried to change other objects to similar shapes with other commands, such as "Change it to a dice. Change this to a cube. Oh, they work as well!"

Although user's used few types of gestures, the gestures inferred different meanings based on the context. For example, a static gesture opening the user's hand was used for pointing, grabbing, moving, and dropping objects. However, the gesture meaning varied according to the combination of speech or with the certain movement of user's hand. We also observed that users keep their gestures the same while they were moving the objects as shown in Figure 3. 9. The users used different static hand gestures to point to the virtual object to interact with.

Figure 3. 9 User's hand gesture for moving an object.

We also observed user's head movement while they were using handheld display device. As shown in Figure 3. 10, users moved their head first followed by their hand movements. The users also changed their head pose to change the AR view depending on their view point or to have a magnified AR view (see Figure 3. 10).

Figure 3. 10 User's head movement for view change with HHD.

Users also gave exclamations such as ‘Oh’, ‘Wow’, ‘Awesome’, ’It is very cool.’, ‘It is kind of fun’, etc. in between spoken commands.

3.9 Discussion

Although gestures got the highest mean score for natural input technique, when we looked at the usage of speech and gesture, combined speech and gesture input was the most used command modality. Counting the number of commands issued, commands that combined speech and gesture input were 63% of the total (49% combined word commands and gestures, and 14% combined sentence commands and gestures), whereas gesture input only commands were 34%, and speech only input was 3.7% (0.4% of words and 3.3% of sentences). This implies that multimodal AR interfaces for object manipulation will rely heavily on accurate gesture recognition, as almost 97% of commands involved gesture input. From the post experiment interview, we

found that all the users did not want to talk with the computer in the same way as they did with other people.

We expected that the display type would affect the way users interacted with the virtual contents since the size of the interaction area varied according to display type. However, from the analysis of results, none of experimental measures showed a significant difference due to display type. We only had twelve users to evaluate how different display types affect pattern of multimodal interface. The small number of subjects can introduce higher variablility in the result, thus, we cannot definitely reject the hypothesis related to the effect of the display type. Although users preferred the screen display over the HHD, and felt it was more enjoyable and easier to interact with the objects. These results are interesting because they imply that people will use similar multimodal speech and gesture patterns in an AR interface regardless of the display type.

3.9.1 Design Recommendations

From the results of the WOz study we can derive some design recommendations that could be used to guide the development of future AR multimodal interfaces. These include:

- Use a gesture-triggered MMI system to reduce delay

- Make sure that the gesture recognition input is as accurate as possible, and is particularly good at recognizing deictic and metaphoric gestures.
- Use speech commands in a phrase, not in a sentence
- Use context-based multi-signal fusion system to improve the accuracy of the system response
- Screen based AR may provide a better user experience

Firstly, the gesture input signal should be used to trigger the multimodal command recognition system. Most current MMI systems are triggered by speech input with a certain size of timing window to look for related commands coming from the gesture input stream. However, as we mentioned earlier, in our task 94% of the time the user gave a gesture command before the related speech input, showing that the onset of the gesture command should be used as the trigger to find the related speech input.

To provide natural hand gesture input, we need to consider a gesture recognition algorithm which recognizes static hand shape and the movement of the hand. In addition, we need to have gesture recognition as accurate as possible because most of multimodal input commands relied heavily on gesture input.

Based on our analysis of the speech commands, we found that most of the speech input was short phrases rather than complete sentences. Although sentence-based speech input can work based on a predefined grammar, it can cause more recognition errors than word-format speech input because commands in sentences include fewer lexicon words than commands in words.

A context-based multi-signal fusion architecture is necessary to improve the accuracy of the system response. During the video analysis, we found that the classification of speech input or gesture input depended on the input context. Thus, we need to have a context-based signal analysis with the help of proper signal fusion architecture.

Finally, it seems that a large screen based AR environment provides a better experience for the users for this type of task. Our analysis has shown that for these tasks the speech or gesture commands used depended on task type not display type. Although we did not see the effect of display within the experiments, the screen display was overwhelmingly preferred by users.

3.10 Conclusions

In this chapter we have described a Wizard of Oz study for an AR multimodal interface and model manipulation tasks that allowed users to use natural speech and gesture input. We found the frequencies of multimodal input and

the optimal size of the multimodal input time window. Deictic gestures (65%) and metaphoric gestures (35%) were the main types of gestures used. We also found that subjects used same gestures with meanings that varied depending on how they moved and which speech command they used. Thus, we need to consider a context-based multi-signal fusion architecture to analyze them more accurately.

Task related words, such as words for colour or shape, were the main speech commands. From the speech input analysis, we found that most of speech commands were given in phrases with a few discrete words (74%), and not full sentences (26%). Overall, in 94% of the multimodal commands, gesture commands came earlier than the corresponding speech commands.

After the formal study with the exploratory data, we found that the MMI used depended on task types, but not on display types. In addition, users preferred the screen display over the handheld display. Thus, for the multimodal system integration in AR, a screen display may be preferable. The size of time window for combining speech and gesture input depends on tasks as well. Moreover, although users felt gesture input alone was a more natural interface than speech or the combination of speech and gesture, 68% of the input involved combined speech and gesture commands.

Based on these findings, the next step is to develop a functioning multimodal AR interface with real speech and gesture recognition. To do this we need to implement an accurate hand gesture recognition module with a multi-signal fusion architecture to give more accurate and natural feedback to users. In addition, the interface has to be compared in formal user studies with the system which does not allow users interact multimodally. As we observed in Section 3.8.7, the same gesture has different meanings according to the corresponding speech. Thus, it is necessary to observe users how they trigger their gestures in AR environments. In the next chapter we explore more in detail on users' gesture input.

Chapter 4

User Observations II – Gesture Pattern Curves

4.1 Introduction

In the previous chapter, we found that a gesture is the main cue to decide a multimodal fusion with a time window. We also found that the gesture has different meaning according to the corresponding speech. In the sense, providing an accurate and stable gesture interface is essential aspect for effective gesture interaction. However, user-centred gesture interface design also has to be considered prior to the implementation of the actual gesture-based system. Much of the research on gesture recognition (such as hand shape recognition) has been done with American Sign Language (ASL) or similar gesture language (Yang & Ahuja, 1998). Mapping sign language to a gesture for interacting with a virtual object could be one option; however, we can provide a more natural gesture if we adopt the hand gesture that users use in their everyday lives for interaction with the real world. To do that, we need to observe how users use gestures when they interact with a target object in the real world.

In this chapter, we explore gesture input by observing and comparing users' gesture pattern in different environments: the Real, the Augmented, and the Mixed environment. The goal of the study is to investigate how different types of gestures are used to interact with various objects. We can see if there is a significant difference between deictic and metaphoric gesture spaces in 3D, by looking at where gestures are made with real and virtual objects. If we can recognize a gesture according to its interaction space, we may be able to develop a novel way of gesture recognition based on the users' gesture pattern. We also want to explore how users felt while they triggered different gestures in the task environments.

4.2 Related work

There has been a substantial amount of previous research on observing gestures in human-human communication. For example, McNeill and Pedelty (1995) observed normal speakers and right hemisphere damaged speakers performing gestures while describing the same scene. They were interested to see how right brain affects the damage on gestures using space. Another study, McNeill (1992) classified gesture into metaphoric, iconic, deictic, and beat gestures: the metaphoric gesture class represents an abstract idea; the iconic gesture depicts an object; deictic gesture mainly includes pointing; the beat gesture includes formless gestures and utterance rhythm. He defined a gesture

space in front of a seated adult, and found the main types of gestures used were iconic and beat gestures.

Lee and Billinghurst (2008) followed McNeill's gesture classification (McNeill, 1992). They found the main types of gesture input used in an AR environment were metaphoric (moving) and deictic gestures (pointing). This shows that human-human interaction may be different from human-computer interaction. Thus, we need to study how people use gesture in the AR environment to explore if this is different from in the Real environment.

The easiest way to achieve this is observing users' behaviour while they are using a prototype gesture interface in a virtual and a real environment. Hauptmann observed users using their gesture and speech to manipulate graphic images (Hauptmann, 1989). For the gesture input, he counted number of fingers and hands presented in the experiment. He also analyzed the type of gestures used into four groups: rotation, sliding, and growing/shrinking. However, their study does not focus on the gesture movement patterns of each.

Epps *et al*. (2006) observed user hand shapes for tabletop interaction. In their study they let users perform the gesture as they wanted. As a result, they captured the typical hand gesture shapes and their main usage. This would be helpful for deciding which gesture has to be included in the tabletop interface

design. However, they were also not interested in gesture movement patterns. Mason *et al*. studied how haptic and visual feedback affects the movement type and the peak velocity of reach-to-grasp movements (Mason *et al.*, 2000). The goal of their research was observing the performance according to the type of feedback, not observing users' gesture patterns. There has been research on observing people's gesture in television talk shows (Kipp, 2007), but this was for synthesizing 2D/3D avatar gestures, and not for a natural interface design.

This chapter represents the first research on different gesture patterns (pointing, touching and moving) in different environments: Real, Augmented, and Mixed. We also captured how users felt while they perform the given tasks in different environments.

4.3 Proposed solution

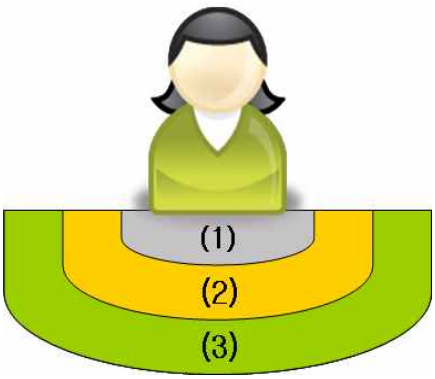


Figure 4. 1Gesture spaces: (1) Preparation area; (2) Deictic gesture space; (3) Metaphoric gesture space

A gesture space was defined as shown inFigure 4. 1. The closest area to the body is the Preparation area where no gesture is made. A subject has to start and end their gesture from this area. The next area is the Deictic gesture space where pointing gestures are triggered and the furthest area is the Metaphoric gesture space where direct object manipulation occurs. The defined gesture space was not told to users because we were interested in whether the natural gesture interaction was able to be classified according to the position where the gesture was triggered.

We used the OSGART (2009) rendering and interaction library to provide an augmented reality view and adopted a hybrid tracking system with the ARToolKit (2009) and ARTTrack3 (2009) computer vision software and hardware for accurate tracking results. The experiment setup is shown in Figure 4. 2.

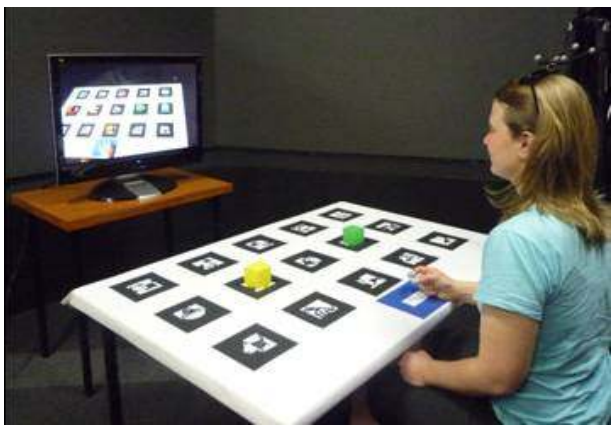


Figure 4. 2 Experiment Setup

Bare-hand interaction would be an ideal interface for natural interaction. However, a stable markerless 3D hand tracking method was not available. Thus, we used two thimble-like reflective trackers to track the position of a thumb and index finger of a user. We also asked the user to wear reflective marker-attached glasses to track his/her head movements. We recorded video of the user for further analysis.

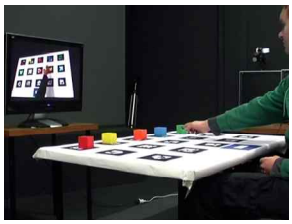
In our research we observed how users interacted with three types of cubes: (1) Real, (2) Augmented and (3) Mixed cubes, using three gestures: (1) pointing, (2) touching, and (3) moving. The three gestures were chosen based on the findings from the WOz study in Chapter 3; the main types of gestures were metaphoric (moving and touching) and deictic (pointing) gestures. In the Mixed cubes condition, half of the cubes were real and half of the cubes are virtual objects.

We provided five different coloured but same sized (40mm) cubes to a user in each scene. Each cube was placed on top of a square marker. The order of the provided task environments was randomized to reduce learning effects.

The user triggered gestures after the experimenter's instructions. Before the user triggers a gesture, his/her hands had to be in the Preparation area. For example, when the experimenter said, "Point at the red cube", the user would

then move their hand from the Preparation area and point at the red cube. After that, the user puts his/her hands back into the Preparation area to finish performing the gesture input. The experiment instructions were designed to let users perform three different gestures at least five times each. We provided a different order of gesture and cube type conditions to each user to avoid learning effects. The subjects were asked to fill out questionnaires after each test condition and when the experiment was finished.

Table 4. 1 Task Table

Real	AR	Mixed
		

Our hypotheses for the study are following:

- H1: The gesture type can be classified by observing only the distance between the position of the target object and the user's hand.

- H2: A different environment will lead to different patterns of gesture curves.

- H3: The time a user spends to complete a task varies according to the position of the target object and the type of the triggered gesture.

4.4 Results

A total of twelve users (9 males and 3 females) participated in the experiment. The average age was 30 years old. They were all right handed except one user. We recruited users from the HIT Lab NZ who were working in the AR research field, but were not familiar with gesture interfaces/MMIs. We were considering MMIs in AR environments; thus we needed to have subjects who were already familiar with AR environments.

4.4.1 Objective User Study

To analyze the users' gesture patterns, we visualized the tracked hand movements in 3D. Our initial idea was that the 3D visualization of tracked information would help to classify gestures based on the distance from the user's body to their hand. A visualization result from one user is shown inFigure 4. 3.

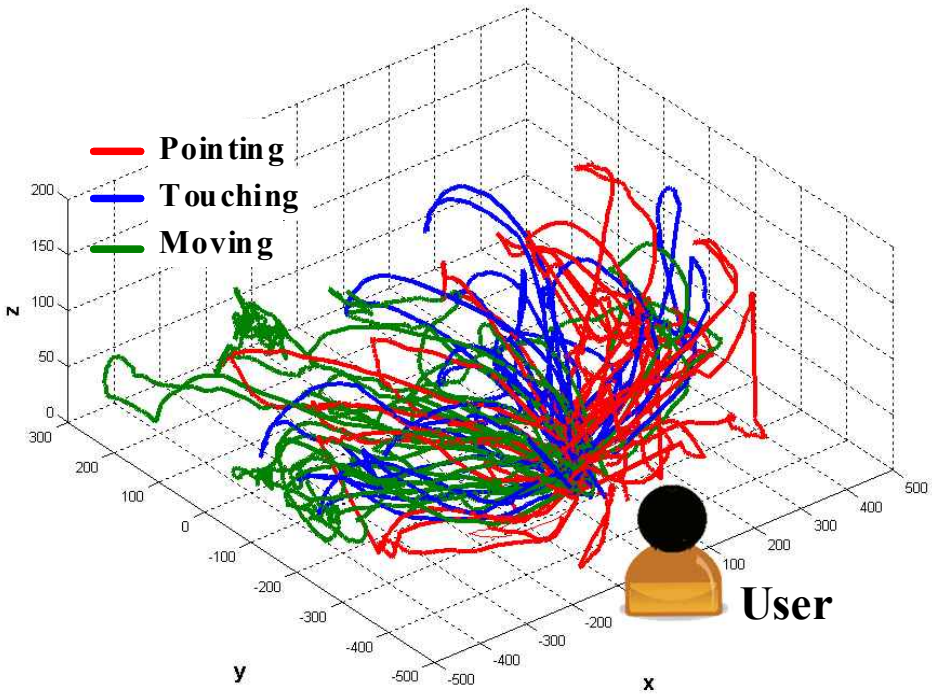


Figure 4. 3 Gesture Path Visualisation in 3D

4.4.2 Normalized Pattern Curves

Using the 3D plots we could not clearly distinguish different gestures because the distance was relative to the position of the target object. Thus we used a second technique where we normalized the range of the users hand based on the initial distance from the subject's hand to a target object. Normalization has been done as shown in Figure 4. 4.

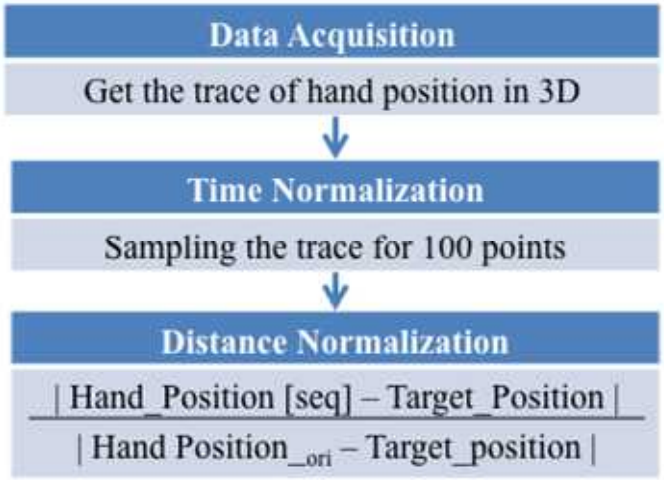


Figure 4. 4 Normalization procedure.

Figure 4. 5 shows the plot of the normalized range to the object for gestures in the real and AR conditions interacting with near and far objects. By comparing curves for each gesture, we found the minimum absolute distance of the pointing gesture was around 0.4 and the distances of the touching and moving gesture were about 0.2. This implies that touching and moving gestures (metaphoric gestures) were triggered further away from the subjects' body than the pointing gesture (deictic gesture). This observation supports our assumed gesture space in Figure 4. 1.

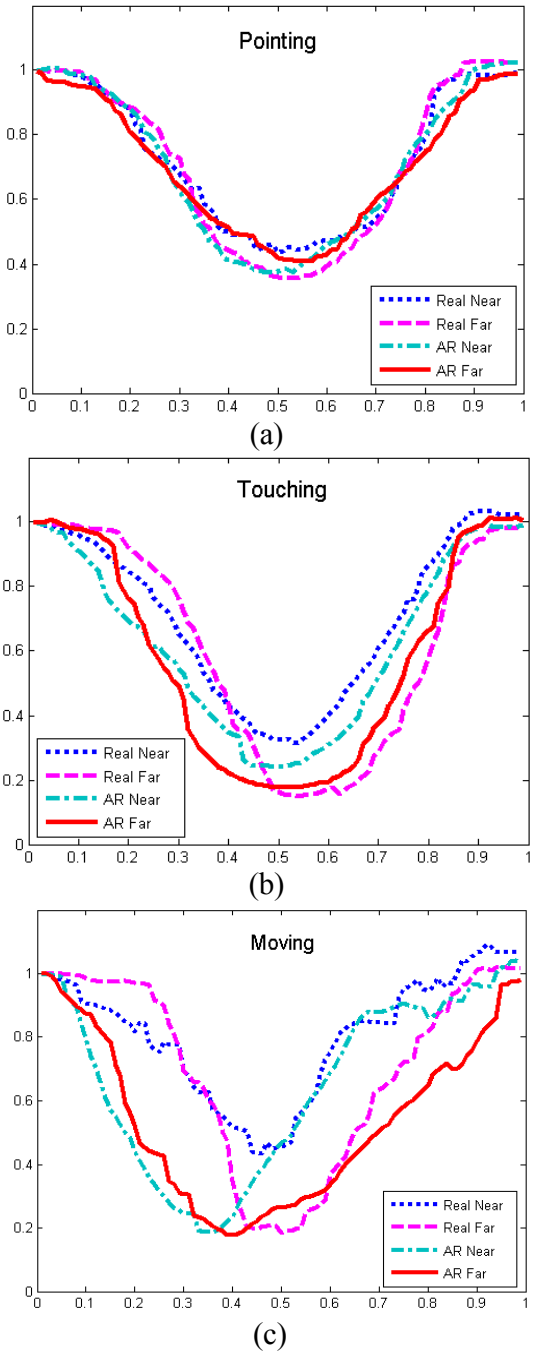


Figure 4. 5 Normalized gesture curves of different gesture patterns in the Real and AR environment.

We classified objects into two groups: The objects on the first row of markers from the subjects were the ‘near’ objects (an average of 30 cm from the user) and the objects on the third row the ‘far’ objects (an average of 63 cm from the user). Then we visualized the normalized curve in each environment and which each group of objects.

There was no significant difference in the pointing behaviour for different environments and for different object groups (Figure 4. 5(a)). This implies that the characteristic of the pointing gesture can be generalized as a single curve. As a result we may be able to detect pointing gestures by observing the movement of a user’s hand or fingertip in an absolute distance.

From Figure 4. 5(b) we observed that touching in the AR environment forms a wider curve than for the Real environment. This implies that performing the touching gesture in the AR environment took longer than in the Real environment. We assume this is because of the lack of haptic or tactile feedback. By comparing curves for touching gestures with near objects in two environments, we found that users moved their hand more in the AR environment than in the Real environment. However, the subjects moved their hand more in the Real environment than in the AR environment for far objects.

This can be thought of as an example of Poulton's range effect (1973) (overshooting near targets and undershooting far targets).

We also observed that users spent more absolute time performing the moving gesture in the AR environment than for the Real environment (Figure 4. 5(c)). We found that the users' hand distance in the Real environment was increased according to the position of objects. However, the distance in the AR environment was not affected by the position of the objects. This may be because the users' hand could move into the cubes when they interact with the objects in the AR environment, but not in the Real environment.

4.4.2.1 Estimating users' gesture pattern in the Mixed environment.

Our initial idea was that users' gesture in the Mixed environment could be described from gesture curves in the Real and Augmented environments. To describe this with a mathematical model, we have decided to apply a regression algorithm. However, the data for each gesture was not normally distributed, so we could not apply a regression algorithm on the gesture pattern curves. Instead, we decided how the combined gesture (RAR), mean of the Real and the Augmented, pattern curves is different from the gesture curve in the Mixed environment by comparing their shapes.

The pointing curve in the Mixed environment with near objects has a different pointing pattern compared to other pointing patterns (Figure 4. 6(a)). Based on this curve, when the users pointed to a near object in the Mixed environment, they spent a longer time to get to the object than the pointing gesture recovery period.

We could not see any significant difference in the touching behaviour for different environments and for different object groups (Figure 4. 6 (b)). This implies that the characteristic of the touching gesture in the mixed environment can be derived by the average of touching curves in the AR and the Real environment.

In Figure 4. 6(c), the moving curves with far objects had a similar pattern. By comparing two moving gesture patterns with near objects, we found that users moved their hand closer to the target object in the Mixed environment than the combined moving curve. Interestingly, the moving gesture pattern with near objects in the Mixed environment looks like the moving pattern with far objects in the AR environment (Figure 4. 6(c)).

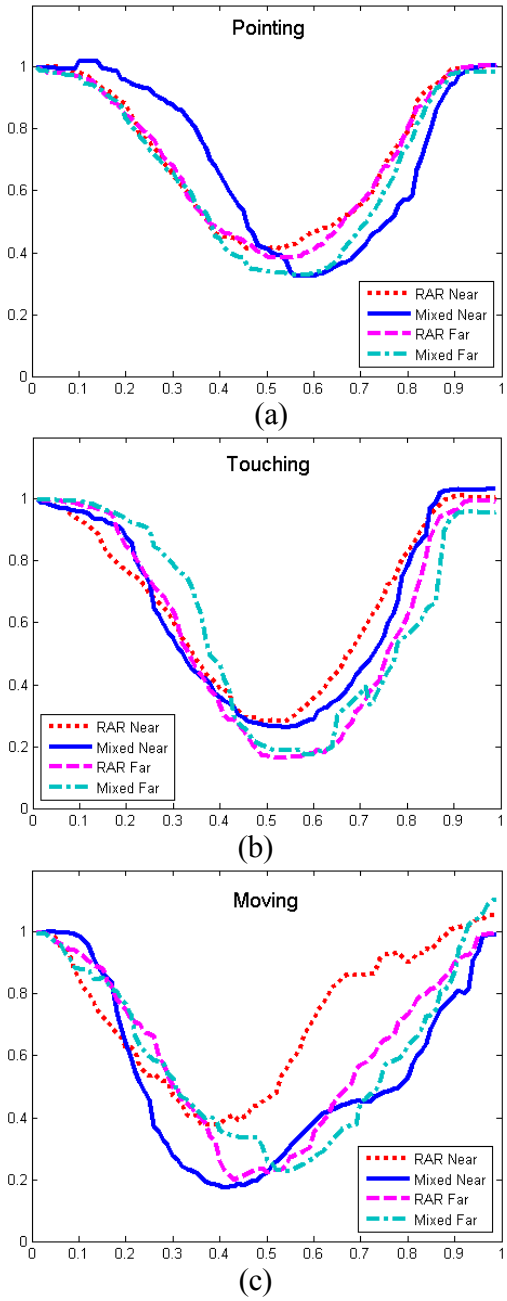
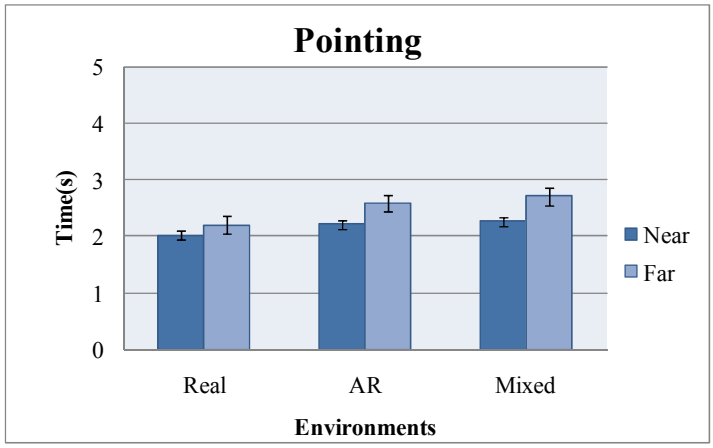


Figure 4. 6 Gesture curves from Real and AR Combination and from Mixed: (a) pointing gesture curves, (b) touching gesture curves, and (c) moving gesture curves.

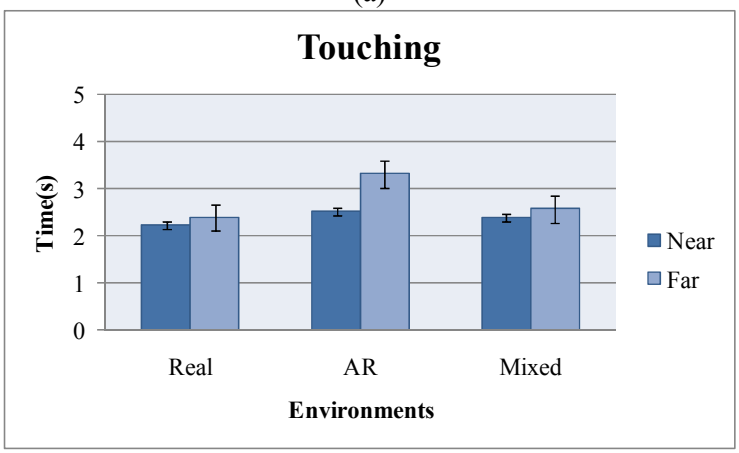
4.4.2.2 Time Analysis

In our gesture pattern curves, we exclude the effects of distance from the objects and the time on the curves. However, in a real interface implementation, the time is also an important factor in recognizing a gesture. Thus, in this section, we analyze the average time for each gesture in each environment with two different types of objects (Figure 4. 7).

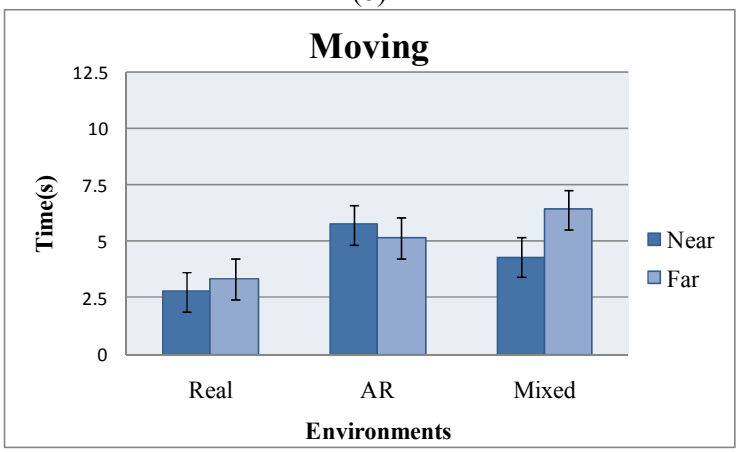
Overall, the average time for near objects was shorter than for the far objects; however there was an exceptional case for a moving gesture in the AR environment. Moving gestures require a direct manipulation with the target object. We assume that the visual feedback after selecting an object might not be enough feedback to users so that users may spend more time to pick-up the object.



(a)



(b)



(c)

Figure 4. 7 Average Time Analysis: (a) Pointing, (b) Touching, and (c) Moving.

We also found that the average time for moving gestures in the mixed environment was very different for each user (large between subject standard deviation). In the Mixed environment the users have real and augmented objects at the same time. This may cause user confusion about the type of object (real or augmented). For example, if a user was expecting that a target cube was real they would think that they could easily pick up the object. However, if they reached the object and found that it was a virtual model, they would need to perform the pick-up gesture more carefully.

4.4.3 Subjective User Study

We also collected subjective feedback to see how the users felt while using the various gestures in different environments. The subjects answered questions on a Likert Scale from 1(very low) to 7(very high). Unlike the first user study, we switched the range of the Likert scales from 5 (Chapter 3) to 7 to create a more sensitive instrument. The subjects' familiarity with AR was 3.75 out of 7 (std = 1.29) and with gesture interfaces was 2.67 (std = 1.07). For analysis, we applied a One-way ANOVA within subjects with the Bonferroni post-hoc test.

To see how natural the gesture interface is for the users, we asked:

- *Q1: It was natural to use gesture input*

We found a significant difference on naturalness of gesture interface among the different environments ($F(2,10) = 8.545$, $p<.01$). The subjects felt that using gesture was more natural in the Real environment (mean = 6.25, std = 0.97) than in the Mixed environment (mean = 5.00, std = 1.54) or in the AR environment (mean = 4.92, std = 1.54).

We asked users how easy it was to use different gestures in different environments with following questions:

- *Q2: It was easy to point to the objects.*
- *Q3: It was easy to touch the objects.*
- *Q4: It was easy to move the objects.*

We could not find any significant differences for ease of pointing. However, we could find significant differences for ease of touching ($F(2,10) = 14.02$, $p = .01$) and for ease of moving ($F(2,10) = 29.60$, $p < .01$) in different environments. The mean scores for each gesture in different environments are shown in Table 4. 2. As can be seen in all cases the real was the easiest, and the augmented was the least easy, although there was no significant difference in the pointing case.

Table 4. 2 Ease of pointing, touching, and moving in different environments.

		Real	Mixed	AR
Pointing	mean	6.75	6.33	6.17
	std	0.62	0.98	1.11
Touching	mean	6.75	5.25	4.50
	std	0.62	1.22	1.45
Moving	mean	6.75	4.08	4.08
	std	0.62	1.21	1.44

We were also interested in how the subjects felt wearing markers and using the Preparation area. The questions were:

- *Q5: I think wearing the thimbles affected my concentration when performing gestures.*

- *Q6: I think putting my hand in the preparation area is uncomfortable or unnatural.*

Interestingly, wearing the thimbles in the Real environment affected their concentration more than other environments ($F(2,10) = 4.24$, $p < .05$). However, we did not find a significant difference in the unnaturalness by having the Preparation area. The mean scores are shown in Table 4. 3.

Table 4. 3 Distractions from the experimental setup.

		Real	Mixed	AR
Thimble	mean	4.00	3.42	3.25
	std	2.00	1.83	1.76
Preparation Area	mean	3.00	3.08	3.33
	std	1.86	1.78	1.83

We also asked users how quickly and accurately they performed the gestures:

- *Q7: How quickly did you perform the tasks?*
- *Q8: How accurately did you perform the tasks?*

We found significant differences for both speed of gesture ($F(2,10) = 4.36$, $p = .04$) and accuracy of gesture $F(2,10) = 9.85$, $p < .01$). The mean values are shown in Table 4. 4.

Table 4. 4 Speed and accuracy of performing gesture

		Real	Mixed	AR
Speed	mean	5.58	5.00	4.33
	std	1.31	1.28	1.23
Accuracy	mean	6.17	5.00	4.33
	std	0.83	1.21	1.23

The subjects answered how physically demanding, mentally demanding, and frustrating it was to perform the gesture. The questions were:

- *Q9: How physically demanding was it?*
- *Q10: How mentally demanding was it?*
- *Q11: How frustrating was it?*

We found that the different environments only affected mental demands ($F(2,10) = 5.034$, $p = .03$). Performing gestures in the AR environment was more mentally demanding (mean = 3.58, std = 1.56) than in the Mixed environment (mean = 3.25, std = 1.36) or in the Real environment (mean = 2.17, std = 1.11).

After the experiment we asked users to answer a post-experiment questionnaire comparing all conditions. We asked users to rank which environment they preferred in based on the following questions:

- *Q1: Which environment was the easiest to use the gesture input?*
- *Q2: Which environment was the most enjoyable to use the gesture input?*
- *Q3: Which gesture was the most enjoyable to use the gesture input?*
- *Q4: Which environment do you prefer overall?*

For all users the Real environment was the easiest to use the gesture input, compared to the Mixed or AR conditions. Seven people answered that the Mixed environment was more difficult than the AR environment. The reason why they felt the Mixed condition was more difficult than the AR one was that the Mixed one has two types of objects (real and virtual) in the same environment. It was hard to figure out which one was real or virtual from the visual cue. In addition, the lack of tactile feedback from the interaction with the augmented cube made them frustrated. As a result, the accuracy of their gesture with the AR cubes was not good as interacting with the real cubes.

Seven users ranked the Mixed environment as the most enjoyable to use for gesture input and three users answered that the AR was the one. Only two users picked the Real environment as the most enjoyable.

Eight users picked the moving gesture as the most enjoyable gesture among three gestures. The pointing gesture was the most enjoyable gesture for three users. Only one user answered that the touching gesture was most enjoyable. Users felt that the moving gesture was the most enjoyable because it is very interactive compared to other gestures.

Only two subjects preferred the AR environment overall. The Real and Mixed environments were the most preferred for five users respectively.

After users finished the experimental tasks, we had an interview with individual users. We wanted to know why more than half users (7 users) ranked the Mixed environment as the most enjoyable to use with gesture input. According to the users' comments, it was something that they had never tried earlier.

4.4.4 Further Finding

In the previous sections we described gesture patterns using movement curves and average completion time. A subjective user study was also pursued. In this

section we have descriptions from observing users from the recorded video. As shown inFigure 4. 8, most of time, users watched the monitor even when they interact with a real object. When we consider the order of environmental condition was randomized this finding is very interesting. From this observation we assume that the separation of interaction area (marker-attached table top) from the visual area (monitor display) in our experimental setup does not distract users concentrating on the interaction tasks. This implies that our experimental setup facilitated seamless interaction.

In the after-experiment interviews, users said that it took a while to figure out they did not need to watch the monitor to interact with the real object.

Figure 4. 8Users watching monitor while they are interacting with the real cubes.

4.4.4.1 Design Recommendations

From our experimental analysis, there are some lessons learned which would be helpful for designing the gesture interfaces:

- Use multimodal interface with speech input

- Use different gesture windows for each gesture type

From the gesture pattern curve, especially the pointing curve, we could see that it has a common shape. Thus if we know which gesture users are performing, we can estimate how they will move their hand based on the gesture pattern curve. In this sense, having a multimodal interface with speech input would provide better recognition results than gesture input alone. For example, if users begin making a gesture after saying 'this', the gesture would be mostly a pointing gesture.

From the time analysis, we found that the average time for each gesture in each environment was different. Thus, having a different sized time window to recognize a different gesture would be helpful to improve the speed of recognition.

4.5 Conclusion

In this chapter we have proposed a gesture classification method based on hand distance from the user's body. We normalized the distance of the users hand based on the initial distance from a subject's hand to a target object with normalized time to exclude effects of object position in gesture pattern. Using

this we observe that touching and moving gestures (metaphoric gestures) were triggered further away from the subjects' body than the pointing gesture (deictic gesture). We also compared the gesture pattern curves from the Mixed environment with the combined (Real and AR) gesture pattern curves.

From the subjective user study, we found how people felt using gestures in the AR and Real environments. Users felt that using gestures in the Real environment was more natural, easier, quicker, more accurate, and less mentally demanding than in the Mixed or in the AR environment. In addition, all the users answered that the Real environment was the easiest one to complete the given tasks compared to the other two environments.

We found a consistent pattern from the normalized pointing gesture curves. We found that metaphoric gestures were triggered further away from the subject's body than the pointing gesture. However, we did not find a common pattern from the touching or moving gesture curves. Additionally, although there is a certain pattern on the pointing gesture curves, it would not be easy to apply the pattern curves to predict the pointing gesture in real time. For example, we cannot estimate how far users hand would reach to point a certain object. Thus, the system response corresponding to the pointing pattern would be delayed after the normalisation process. In the next chapter, we will

describe our multimodal fusion architecture which is based on previously mentioned two user observations.

Chapter 5

Final MMI system

5.1 Introduction

In this chapter, we describe our AR MMI which combines 3D natural hand gesture with speech input. We also present our own multimodal fusion architecture. Additionally, we will describe a sample application which demonstrates how the two interfaces are connected to the fusion module.

5.2 Related Work

As shown in Section 2.3.3, there have been only few examples of AR MMIs, and none of them has used computer vision techniques for natural 3D hand interaction. There has also been very little evaluation of AR multimodal interfaces, especially on the usability of AR MMI.

MMI typically involves understanding two or more input modes at the same time (e.g. speech, gesture, gaze, etc). One of the features which distinguishes MMI from unimodal interfaces is the fusion of modalities into a single input

command (Dumas *et al.*, 2009). Thus well-designed fusion architectures are needed because they can enable natural and effective multimodal interfaces.

As shown in Section 2.4, there are two approaches for multimodal fusion: early fusion and late fusion (Pfleger, 2006). Early fusion is used for merging highly correlated input at the feature level. The combination of speech input and video of lip movement is an example combination for the early fusion. Late fusion is adopted for integrating modalities from different modes. For example, a rule-based method for integrating speech and gesture input. In our case, we consider different type of input modes, so we use a late fusion approach. The research described in this chapter is novel because it uses computer vision to support natural hand input in a 3D AR environment for 3D object manipulation. From the previous research, we found that there is no research which tested the usability of an AR MMI with 3D natural hand gesture and speech input. Unlike previous work, our research is targeting AR applications and uses adaptive filters based on observations of real user behaviour for simple modality fusion (from Chapter 3).

5.3 Proposed Augmented Reality Multimodal Interface

In this section we describe our AR MMI that combines 3D stereo vision-based natural hand gesture interface and speech interface. In addition, we also explain our multimodal fusion architecture which merges MMI.

Our AR MMI system is made up of a number of components that are connected together. They include input modules for capturing video, recognizing gestures and speech input, a fusion module for combining speech and gesture input, and AR scene generation and AR scene manager modules for generating the AR output and providing feedback to the user. Figure 5. 1 shows how the AR MMI components are connected.

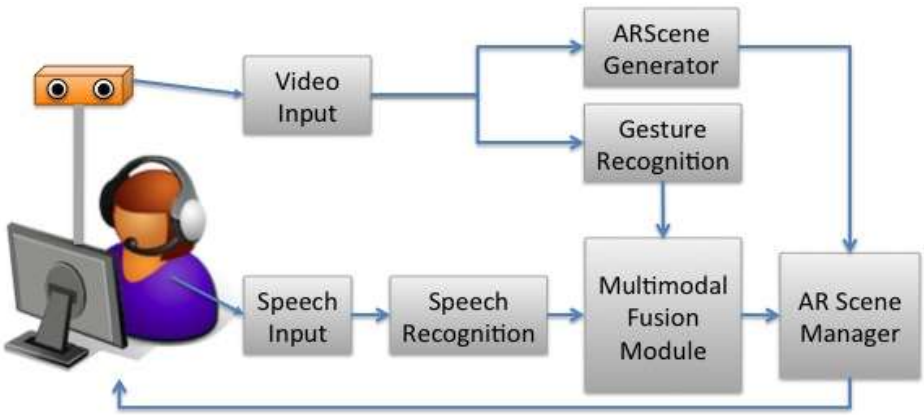


Figure 5. 1 The architecture of the AR MMI.

In the Wizard of Oz study (Chapter 3), we observed how users use their natural gesture and speech input in an AR environment and how the users

integrate and synchronize two different modalities. As a result we found that the same gestures had different meanings based on the context; that is, the meaning of a gesture is varied according to its corresponding speech command. We also found that users mostly triggered gestures before the corresponding speech input, meaning that a gesture-triggered time window needs to be used to capture related commands. From the study, we found that people used three different types of gestures: (1) open hand, (2) close hand, and (3) pointing. In the next section we describe the computer vision techniques we have used to capture free hand gestures.

5.4 3D Hand Gesture Interface

We have implemented a gesture recognition method to capture 3D hand gestures from a stereo video input. Our approach is based on five steps: (1) Camera calibration (off-line), (2) Skin colour segmentation, (3) Fingertip detection, (4) Fingertip estimation in 3D, and (5) Gesture recognition (see Figure 5. 2).

5.4.1 Camera calibration

First of all, we need to have 3D information about the user's hand position for bare-hand interaction in AR environments. For this, we need to calculate an accurate 3D position of the fingertips. The first step was to map 2D image

points to corresponding 3D positions by triangulating two points. To do this we needed to have accurate camera calibration. We adopted Zhang's calibration algorithm to find out the intrinsic and extrinsic parameters of the two cameras (Zhang, 2000).

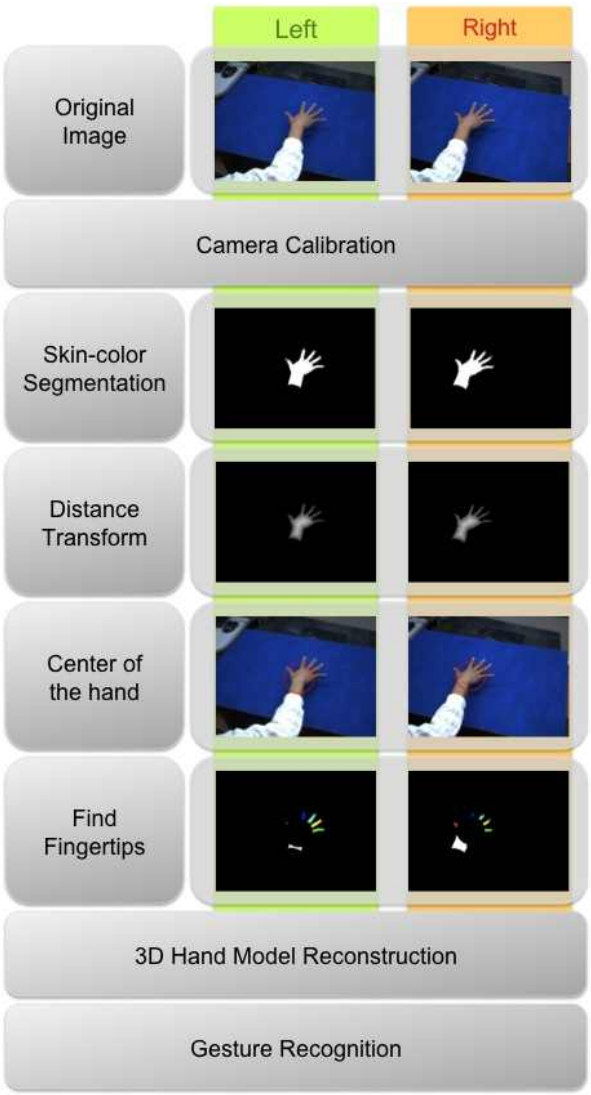


Figure 5. 2 Hand gesture recognition procedure

The parameters from the calibration are used not only to reconstruct the fingertips in 3D but also to augment virtual object in the real environment.

5.4.2 Skin-colour segmentation

To find the users' hand in the camera image, we used a skin-colour segmentation method. We adopted a statistical model-based skin-colour segmentation algorithm in our gesture interface module for supporting real-time interaction; specifically, Chai and Bouzerdoum's algorithm that uses a Bayesian approach for skin colour classification in YCbCr colour space (Chai & Bouzerdoum, 2000). The statistical model-based skin colour segmentation is based on a large sample of ethnically diverse people to determine an accurate statistical skin colour manifold of humans. The distribution of skin colour in normal RGB colour spaces is irregular and widely distributed, and it is very sensitive to noise. Thus, in their research, the input image in RGB colour space is converted to the YCbCr colour space. They found the distribution of the skin colour in the YCbCr colour space is concentrated in a small area. To guarantee stable skin colour segmentation, we controlled the environment with a single coloured background.

5.4.3 Fingertip detection

From the study in Chapter 3, we learned that people naturally used a small number of hand gestures. The number of fingertips which were visible to the camera was limited to 0 (for closed hand), 1(for pointing), or 5(for open hand). Thus, recognizing the number of visible fingertips is one of the easiest ways to recognize these gestures. Thus, we estimate fingertip positions by (1) drawing the convex hull based on the segmented hand region, (2) applying a distance transform (Borgefors, 1986) to find out the centre point of the hand (the furthest point would be the centre of the hand), (3) removing the palm area to leave only the segmented fingers, (4) finding the contour of each finger blob, (5) calculating the distance from points on each contours to the hand centre, and (6) marking the furthest point on each finger blob as a fingertip. The algorithm we proposed is simple and it works effectively with the reduced computational complexity.

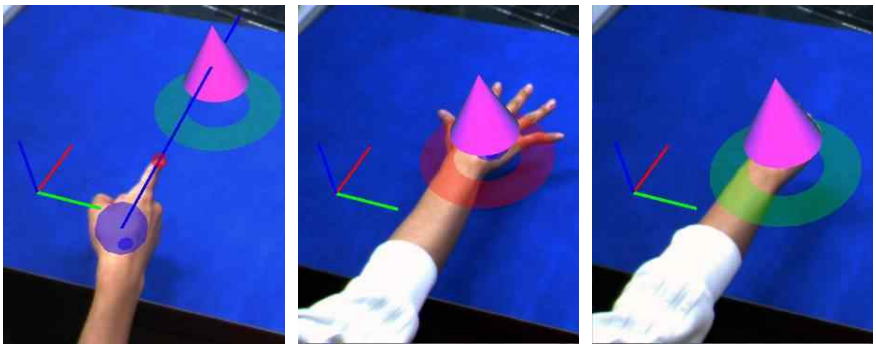
5.4.4 Fingertip estimation in 3D

Once we know the fingertip locations and calibration matrices, we can estimate the 3D position of the fingertips in real-time. This is done by performing a triangulation which solves the linear equation generated from two corresponding fingertip-points observed at each camera (Hartley &

Zisserman, 2004). That is, the triangulation is used to get the 3D point X which satisfies $x_1 \sim P_1X$ and $x_2 \sim P_2X$ where P_1 and P_2 are projection matrices of two cameras and x_1 and x_2 are the observed points on each images, respectively. In the ideal case, the two vectors from the optical centres to the 3D point meet at one position so that we can get a unique solution. However, there are only a few possibilities when two vectors meet in 3D space in practice. Mostly, the vectors are at a skew position. To estimate the position of a fingertip from two vectors at the skew position, we find the point that has the minimum distance between two vectors satisfying the epipolar constraint. This gives reasonable estimation results without reducing the frame rate.

5.4.5 Gesture Recognition

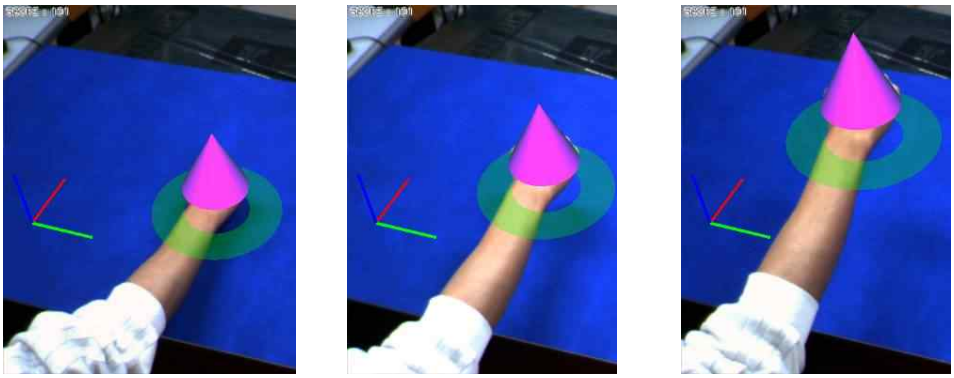
Based on earlier Wizard of Oz study work (Chapter 3), there are three gestures we need to have in AR MMI: (a) open hand, (b) closed hand, and (c) pointing. It is easy to recognize these gestures by considering the number of fingertips visible; an open hand has 5 fingertips; closed hand has 0 fingertips; and a pointing gesture has only one fingertip. The moving gesture is recognized as a continuous movement of the close hand. We were able to track the user's fingertip with accuracy from 4.5mm to 26.2mm depending on distance between the user's hand and the cameras. The accuracy was enough to support our tasks. Figure 5. 3shows the three hand gestures we implemented.



(a) (b) (c)

Figure 5. 3 Hand gestures interacting with the augmented object (a) pointing gesture, (b) open hand gesture, and (c) close hand gesture

The hand tracking works in three dimensions (see Figure 5. 4). The user places their hand inside the virtual pink cone model and then closes their hand to select it. While their hand is closed the user can pick up and move the cone in 3D (Figure 5. 4 (a)). As the hand moves higher, the pink cone gets bigger (Figure 5. 4 (b) and (c)).



(a) (b) (c)

Figure 5. 4 Hand tracking on 3D: as users moving their hand close the camera, the augmented cone is bigger

5.5 Speech Interface

For the speech input, we used the Microsoft Speech API 5.3 with Microsoft Speech Recognizer 8.0 (2009). Speech recognition results are described in a unified form like the gesture recognition results. The arrival time of the speech input is passed to the multimodal fusion module. We define the type of speech command in advance to use it later for integrating it with gesture input. The supported speech commands are shown in Table 5. 1.

Table 5. 1 Supported speech commands

Colour	Shape	Direction
Green Blue Red Yellow	Sphere Cylinder Cube Cone	Backward Forward Right Left Up Down

5.6 Multimodal Fusion Architecture

We designed and implemented a user-centred multimodal fusion architecture which generates a single system input out of input from two different modalities. Figure 5. 5 shows how the proposed multimodal fusion

architecture works. It consists of three sub modules: (1) unification, (2) integration, and (3) scene management modules.

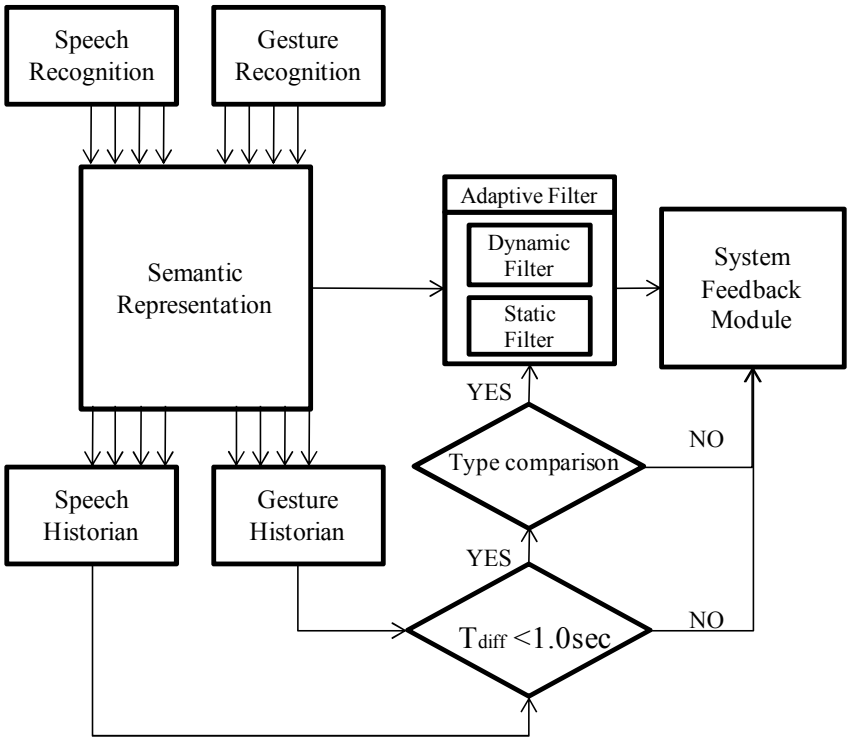


Figure 5. 5 The proposed fusion architecture

We assume that the gesture or speech recognition is done by an independent speech or gesture recognition module and only the recognition results are passed to the multimodal fusion architecture. Recognition for each modality needs to be done separately in parallel. Although we cannot guarantee the accuracy of the fusion system, we built our multimodal fusion system based on observation the user. This method does not require a large number of training

or test data sets. Additionally, it is easy to build a multimodal fusion architecture; although user observation using certain types of interface is essential.

5.6.1 Unification Module

Once the recognition results are available, they are passed to the unification module. The Unification module consists of two parts; (1) the semantic representation module and (2) the historian module.

In the semantic representation module, the speech and gesture recognition results are represented to a unified form.

First, as we saw in 2.4.1, the multimodal fusion architecture cannot be free from timing issues. Thus, the semantic representation template has to have a time stamp slot. The signal arrival time of an input will be stored in this slot.

Second, we need to know what the gesture or speech means. The recognition result is stored in the 'Function' slot. According to the function of a command, we can classify whether the function belongs to deictic group or metaphoric group. Thus, we also need to put the 'type' of the command into the semantic template as well. After the first user study in Chapter 3, we found that the main type of gestures when a user interacted with virtual objects in an AR

environment were deictic and metaphoric. For example, a gesture which is used for pointing at a green cube is a deictic gesture and one for moving a red sphere is a metaphoric gesture. Speech can also be classified into three groups; (1) deictic, (2) metaphoric, (3) miscellaneous. Deictic commands are *this, that, here*, and so forth, while metaphoric commands are *move, drop, stop, etc*. Miscellaneous commands include the speech commands which describe characteristics of the target object, such as, *red, green, sphere, cube,* etc.

When we consider the 'put-that-there' example (Bolt, 1980), we need to consider two reference points to know 'where' it is and 'where' to put it. Thus, we need to have a semantic template for a gesture which requires two reference points. The template for representing the recognition result is shown in theTable 5. 2. There are three different forms for unification: single-point required gesture, two points required gesture, and speech.

Table 5. 2 Semantic attribute-value pairs (a) for pick-up and drop gesture recognitions, (b) for point and move gesture recognitions, and (c) speech recognition

ID # Time Stamp C1 – Function Type Position – x Position – y Position – z	ID # Time Stamp C1 – Function Type Position – x1 Position – y1 Position – z1 Position – x2 Position – y2 Position – z2	ID # Time Stamp C1 – Function Type Colour Shape
(a)	(b)	(c)

The historian module is where the input is stored in the order of arrival. From the semantic representation module, we could unify the recognition result in the semantic form. We will store unimodal input for ten seconds because we may need to refer the previous command in a short time later.

5.6.2 Integration Module

As we studied earlier in Chapter 3, AR MMI is a gesture-driven interface. All unimodal input is described in semantic representation and the input is stored in the historian in the order of arrival. Using the most recent gesture input, the system will search through all the speech input which arrived up to a second after the gesture input arrived. If there is a speech input that arrives within a second after the gesture input has been triggered, the input is considered as a

multimodal input, unless the gesture input goes directly into the system after a second. When we have a valid multimodal input, the fusion module will check whether the types of the speech and gesture commands are compatible and can be resolved into a single command.

Using speech recognition in a quiet environment with a trained recognizer typically produces more stable results than computer vision based gesture recognition. From Chapter 3, we learned that the meaning of some gesture can vary according to accompanying speech input. Thus, in our fusion architecture, we have a procedure where the system can change the meaning of the gesture according to the corresponding speech input.

The gesture and speech input is merged according to the type of the input modality. The type of the function is decided automatically based on the pre-description of the enabled commands (Chapter 3). We have two types of filters in the Adaptive filter module: one is for moving commands (Dynamic Filter) and the other is for static commands (Static Filter). In the case of the Dynamic Filter, it handles two points, the starting point and the destination point. In the case of the Static Filter, we only need to have a single point. To easily handle the objects in an AR scene, we need to know which object the user wants to interact with. Thus, based on the pointing or moving spot, we can estimate which object the user wants. The template for each filter is shown in Table 5. 3.

Table 5. 3 Types of output from the adaptive filter module template: (a) Dynamic Filter and (b) Static Filter

(a)	(b)
ID #	ID #
Time Stamp	Time Stamp
Function	Function
Target object ID_{start}	Target object ID
$P_{start}(x,y,z)$	Characteristics
$P_{end}(x,y,z)$	

5.6.3 Scene Manager

The fusion result is passed to the system to interactively update the AR scene. Thus, we have a trigger to update the AR scene and the data base of the AR view. According to the fusion result the AR scene is changed and audio-visual feedback given to users.

5.6.4 Illustration how the architecture works

When a speech or gesture input arrives, the recognition modules for each input will recognize what the speech or gesture input means. For example, if a user triggered a pointing gesture and spoke “red”, then each recognition module will recognize each input as “pointing” and “red” Then the result is passed to

the semantic representation module with its arrival time. The gesture and speech recognition results in the semantic form are shown inTable 5. 4.

Table 5. 4 Example of semantic recognition result representation: (a) gesture recognition result in the semantic form and (b) speech recognition result in the semantic representation

ID # G124 Time Stamp: 20:08:11:30 C1 –Function: Point Type: Deictic Position X1– 50.0 Position Y1– 132.5 Position Z1– 80.45 Position X2– NULL Position Y2– NULL Position Z2– NULL	ID # S176 Time Stamp: 20:08:12:01 C1 – Function: Red Type: Misc Colour: Red Shape: NULL
(a)	(b)

The output from the semantic representation module is passed to the speech and gesture historians respectively. The system will take the latest speech input from the speech historian and compare the time difference with the latest gesture input from the gesture historian. The fusion architecture will compare the time difference between two inputs. For example, the time gap between gesture input and speech input was 31 ms. This difference is smaller than 1 second that we set as a threshold to decide whether it is multimodal or unimodal. The gesture and speech input is checked whether they can be merged based on the type of the input modality. The pointing gesture has only one reference point, and speech input is *Misc* which represents characteristic

of a target object. Thus, we will proceed to have multimodal input from two unimodal interfaces. The pointing gesture has only one reference point. Thus, two independent unimodal inputs are merged with the static filter. The merged multimodal result is shown in Table 5. 5.

Table 5. 5 Example of the result from the static filter

ID # M84
Time Stamp: 20:08:11:30
Function: Misc
Target object ID: 04
Characteristics: Red

The result has an ID as a multimodal input with. The time stamp decided by referring to the time tamp of the first arrived unimodal input. The type of the function is changed to Misc according to the speech function. The target object ID is decided by calculating the closest distance between the reference points from the pointing gesture and each object's position. Finally, the characteristic we want to change with the multimodal input is setting the color of the object to red. All necessary information is filled out; thus, the output of the Adaptive Filter module is passed to the system feedback module which changes the AR scene according to the outputs.

5.7 Conclusions

In this chapter, we proposed the final AR MMI. A 3D natural hand gesture interface was implemented that recognized three gestures; (a) open hand, (b) closed hand, and (c) pointing. We implemented a simple algorithm to recognize the three difference gestures based on the number of visible fingertips. We developed speech interface using the Microsoft Speech API 5.3 with Microsoft Speech Recognizer 8.0 (Microsoft 2009). We also described a multimodal fusion architecture with adaptive filters. Unlike other multimodal fusion architectures, we designed the filters based on the user observations. As a result, we could implement our fusion module without any neural network or any other complex algorithms for AR applications in real time. Additionally, it includes a scene manager to update the AR scene corresponding to the multimodal input. The speech and gesture recognition results were represented in the semantic form. This helps the multimodal fusion architecture merge the two input in a semantic way.

We are interested in how our MMI improves efficiency and effectiveness of AR interaction by comparing the MMI with unimodal cases: speech-only and gesture-only. We also want to know how users feel using the AR MMI. Thus, we will run a user study to evaluate the usability of the final AR MMI and will describe findings from the user study in the next chapter.

Chapter 6

Usability of the Multimodal Interface

6.1 Introduction

The final goal of our research is on the usability of multimodal input for seamless AR interfaces. Usability is defined by Bevan as "quality in use (Bevan, 1995)." Quality in use measures can be defined with three aspects: effectiveness (accuracy and completeness), efficiency (use of time and resources), and satisfaction (preferences). It is important to account for all three aspects of usability because a subset of the three is often insufficient as an indicator of overall usability (Frøkjær et al., 2000). Thus, in our work we will evaluate the effectiveness, efficiency, and user satisfaction of people interacting with our AR MMI.

To evaluate the usability of the MMI and fusion architecture we conducted a simple user study with the simple AR application described in Chapter 5. The application was a desktop AR interface that allowed users to move virtual objects and change their colour and shape. We used GLUT (2009) to create the AR scene and OpenCV (2009) to implement the gesture recognition module. Speech only and gesture only conditions were also evaluated to

compare with the over usability of the MMI interface. In the next section we describe our experimental set up and user tasks.

6.2 Related Work

There has been little previous research on user evaluation for multimodal interfaces.

Heidemann *et al.* (2004) evaluated the menu control with success rates using their vision algorithm. However, they only conducted user studies for their vision algorithm only. It did not show how multimodal interaction effects to improve accuracy of selecting menus or pointing real objects.

Irawati *et al.* (2006) also conducted a user study, which verified that combined multimodal speech and paddle gesture input is more accurate than using one modality alone. However, the system could not provide a natural gesture interface for users, and required the use of a paddle with computer vision tracking patterns on it. Moreover, they did not fully explore the usability of their MMI system.

However, none of previous research in AR MMIs evaluated the AR MMI with the three aspects of usability; effectiveness, efficiency, and satisfaction. As factors for the three aspects of usability we have the accuracy of the speech

and gesture recognition, the accuracy of fused output commands, and time measurements.

6.3 Proposed Method

The primary goal of the study was to evaluate the usability of the multimodal interface with speech and gesture input. We measured the efficiency of each interface by measuring the task completion time. We measured the effectiveness of each interface by capturing the accuracy of the system input and the user satisfaction by using post-condition questionnaires.

There were twenty five participants in the experiment, twenty-two male and three female, with ages from 25 to 40 years old and all right-handed except one user.

We set up the experimental environment as shown in Figure 6. 1. We used a BumbleBee camera(Point Grey Research Inc, 2009), which has two cameras on a rigid body, to get two synchronized video input (320×240 pixel resolution, 25 fps). The BumbleBee camera was placed on the side of the user to grab the two synchronized images of the user environment to track the user's hand in 3D (according to our algorithm described in Chapter 5). Subjects were asked to wear a headset with a noise cancelling microphone for speech input. A 37-inch LCD screen was placed in front of them for viewing the AR scene. In

between the users and the screen a colour board is placed to get the reference point of the augmentation and the unique background for better skin colour segmentation results.

Figure 6. 1 Experimental setup

6.4 Experimental Task

Users had to complete a number of tasks. For each, the subjects had one sample object at a time that they needed to manipulate. The user was supposed to change the shape or colour of the sample object corresponding to a target object shown on the screen. To let the users easily discern the sample object from the target object, we put a torus under the sample object. We showed the target object as a transparent object with a different colour and shape from the sample object. The typical user tasks are shown inTable 6. 1.

Table 6. 1 Commands list to complete a task

1. Change *the* colour *of the pink cone to the colour of the target.*
2. Change the shape *of the cone to the shape of the target.*
3. Move the object *to the target position.*

Figure 6. 2 shows the AR view of the task at the starting position. Figure 6. 2(a) shows the initial AR scene; (1) represents the sample object which users interact with, (2) is the target object: users have to change the colour and shape of the sample object to this, (3) is the shape change tool, and (4) is the colour change tool. The subjects were asked to perform 10 set tasks of 3 different commands to complete a given task with (1) speech only, (2) gesture only, and (3) multimodal interaction for a total of 90 tasks. Each task involved using a particular interface to interact with a pink cone. There were short questionnaires at the beginning, after each condition, and at the end. In total, the experiment took approximately 45 minutes. For counterbalancing of order effects among different interfaces, we used a 3x3 Latin Squares design.

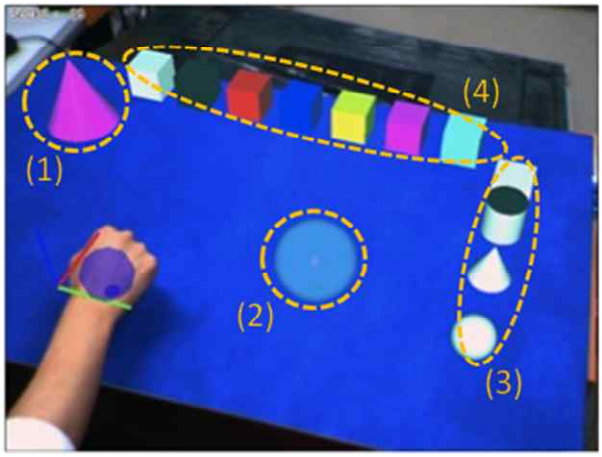


Figure 6. 2 A user doing the task 1 : initial view of the original AR scene; (1) sample purple object to interact with; (2) target blue object representing target shape, colour, and position; (3) shape selection bar; (4) colour selection bar

Video data of user interaction was collected from each of the task and interface conditions for all subjects. Using the video, we compared the fusion result with the actual multimodal commands which users issued. In addition, we also observed the user and system errors while the subjects were interacting with the given objects. Finally, we made further annotations while observing users in the recorded video.

6.5 Pilot Study

From the total subject pool, we ran five participants in an initial pilot study. They were researchers who are familiar with AR applications but had little experience with speech interfaces, gesture interfaces, and MMI.

The experimental task was the same as previously mentioned. A user repeated 10 set tasks of three different commands to complete a given task with speech only, gesture only, and an MMI condition. After running the pilot study, we found several problems, particularly with the lack of depth cues of the object rendering. Additionally, the large number of colour and shape selection conditions made it difficult to remember the correct speech input. This feedback from the users was used to modify the AR scene and available speech commands. We added a ground plane and solved the occlusion problem, which occurred when the augmented virtual object occluded the user's hand, increasing the visual realism. The segmented hand image is used as a mask to filter out the hand region in (1) an AR view image and (2) a real view image. The filtered part in the two images will be drawn in white. After applying an exclusive or (XOR) operation with the two images, we fill the true region with the real video value. As a result, users could have an AR view where their real hands are still shown. This process is shown in Figure 6. 3.

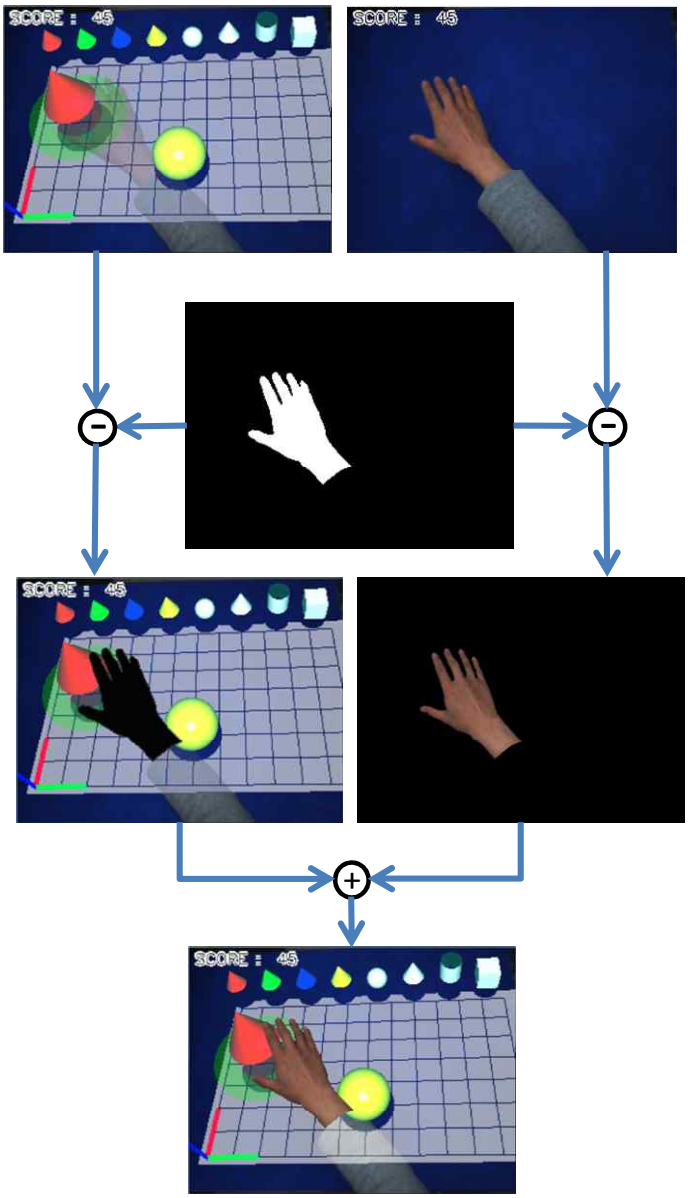


Figure 6. 3 Process to solve the hand occlusion problem.

As a result, we could have the improved application as shown in Figure 6. 4.

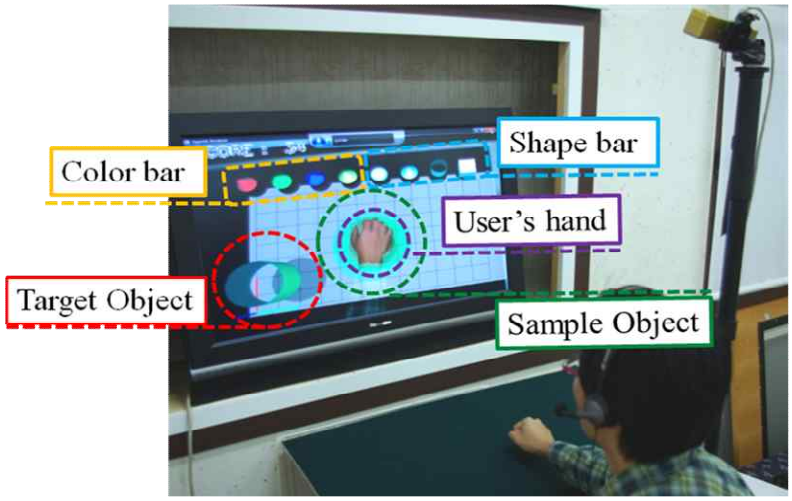


Figure 6. 4 The modified experimental environment

We declared the hypotheses of the study as the following:

- H1: MMI is more efficient than the speech or gesture interface.
- H2: MMI is more effective than the speech or gesture interface.
- H3: MMI is more satisfying than the speech or gesture interface.

6.6 Result and Analysis

In this section we compare the usability factors of (1) efficiency, (2) effectiveness, and (3) satisfaction for each interface. The values to measure

each factors are (1) task completion time, (2) user and system errors, and (3) satisfaction (questionnaire), respectively. In the analysis, we only take into account data from twenty users of the redesigned interface.

6.6.1 Task Completion Time

The task completion time is often used as a factor to measure the efficiency of a system. We measured the time between when users started a task and when they finished the task with the given interface conditions. As a result, we found that the average time that subject spent to complete a task was 52.25 seconds with the speech-only condition and 61.35 seconds with the gesture-only condition. In case of MMI, the subjects spent only 49.30 seconds on average. We used one-way repeated measure ANOVA with post-hoc pairwise comparisons (with Bonferroni correction. Average task completion time is also measured for each interface. There was a significant difference in the task completion time ($F(2,18) = 8.78$, $p < .01$). The gesture interface was different from the speech interface ($p < .01$) and the MMI ($p < .01$). It took a longer time to complete the given tasks with gesture interface (mean = 15.44, std = 4.47) than speech interface (mean = 12.38, mean = 3.15) and MMI (mean = 11.78, std = 2.70).

We also analysed sub-task completion time to check which interface is more efficient to complete sub-tasks. We define each sub-task (SBT) as:

- SBT1: Changing shape of an object
- SBT2: Changing colour of an object
- SBT3: Moving the target object to the destination

In case of speech-only condition, there was significant difference on sub-task completion time ($F(2,18) = 20.542$, $p < .01$). SBT3 completion time was different from SBT1 ($p < .01$) and SBT2($p < .01$). However, we did not find any significant difference between SBT1 and SBT2 ($p = .384$). With speech-only condition, SBT3 (mean = 6.03, std = .476) took longer completion time than SBT1 (3.11, std = 0.23) and SBT2 (mean = 2.723, std = 0.16). This implies that using speech only input to move the target object is less efficient than changing the characteristics of target object. However, we could not find out any significant difference on completing SBT with gesture-only condition ($F(2,18) = 1.19$, $p = .33$).

We also analysed how each modality of interface is used differently to complete the SBTs.

As a result, we observed that speech interaction time in multimodal interface to complete SBTs was different from that in unimodal interface ($F(1,19) = 65.17$, $p < .01$). Speech took shorter SBT completion time when it comes with gesture input (mean = 3.95, std = .21) than in unimodal interface (mean = 2.192, std = .223). Gesture also took shorter SBT completion time when it comes with speech input (mean = 1.01, std = .153) than that as a unimodal interface (mean = 3.06, std = 0.130).

Although it was not statistically validate that gesture interface is more efficient for metaphoric interaction (SBT3), we observed that speech interface is more efficient for detailed interaction related to the characteristics of the interaction object (SBT1, SBT2) ($F(1,19)=80.65$, $p < .01$).

6.6.2 User Errors

A user error represents errors which are caused by user's mistakes. For example, a user was asked to issue a speech command 'red', but he/she said 'yellow'. User errors, as well as system errors, are used as a factor to measure the effectiveness of the system. To measure the user accuracy, we observed how many times users made errors by analyzing the video of them interacting with the system and counting by hand the number of errors made. The average number of user errors with speech input was 0.41 times per task, the average

with gesture input was 0.50 times per task, and for MMI, the average number of user errors was 0.42 times per task. However, we did not find significant difference on the average number of errors with different interfaces. The subjects made 0.44 errors on average per task overall.

Most of the user errors with gestures happened when they did not trigger the proper gestures. For example, when they picked up the object, they had to put a hand on top of the object and close their hand. However, some users closed their hands first and then moved to the sample object to grab it.

6.6.3 System Errors

System errors are dependent on speech recognition accuracy, gesture recognition accuracy, and multimodal fusion accuracy. The accuracy of each interface component was measured using the following equation:

$$A_{Interface} = \frac{N_{Accurate\_respond\_to\_the\_command}}{N_{total\_triggered\_command}}$$

The average accuracy of the speech interface was 94.07%, and of the gesture interface was 85.49%. In a multimodal system, a combined speech and gesture command could fail because of errors in the speech recognition, the gesture recognition, or both the speech and gesture recognition. Thus, when we

assume that the two interfaces are independent, the combined accuracy could be expected to be no smaller than the accuracy of the speech input multiplied by the accuracy of the gesture input. Additionally, the maximum accuracy of the MMI is not larger than the accuracy of the interface that has better accuracy than the other. Thus, the accuracy of our MMI can be represented as follows:

$$A_{speech} \times A_{gesture} < A_{MMI} < Max(A_{speech}, A_{gesture})$$

When we combine two unimodal input without the multimodal fusion architecture, we would expect each of these sources of error to multiply to produce an accuracy of around 81%, and the maximum expected MMI accuracy would be 94%. However, we found that the accuracy of the MMI was 90%, showing that the fusion module helped to increase the system accuracy by capturing related speech and gesture input and compensating for error. However, when we consider the fact that MMI requires two unimodal input, the accuracy of the MMI may deviate greatly according to the combination of two modalities.

6.6.4 Satisfaction

We also collected user feedback to observe user satisfaction with each modality, and especially the MMI. The subject answered questions on a Likert scale from 1(very low) to 7 (very high). The users' English fluency was 4.2 out of 7. Their average experiences with speech and gesture interfaces were 3.45 and 3.55 respectively. Their average experiences with MMI were 3.05. They scored their experience with AR as 5.1.

We used one-way repeated measure ANOVA with post-hoc pair wise comparisons (with Bonferroni correction) to see how different type of interfaces affected user satisfaction.

6.6.4.1 Naturalness of the Interfaces

To see how natural the interfaces were for the users, we asked

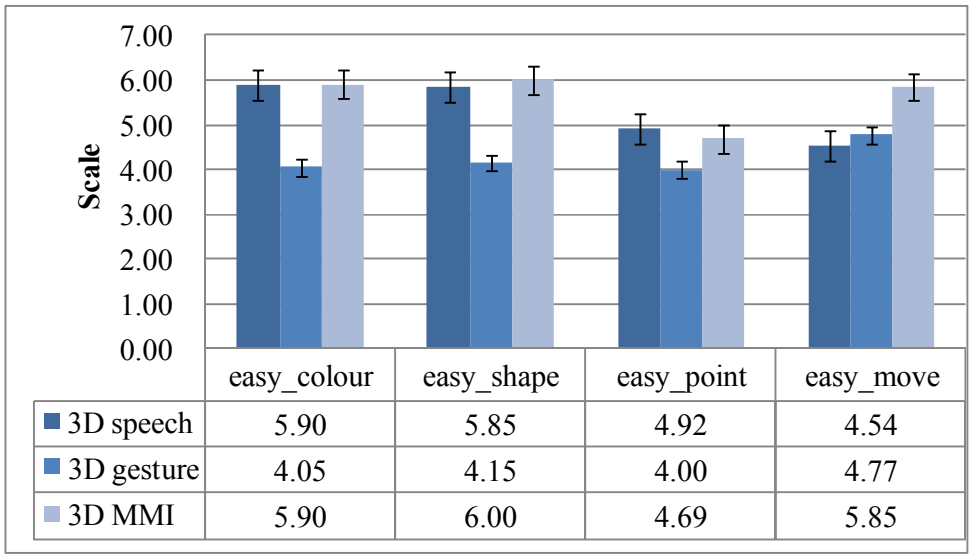
- *Q1: How natural it was to manipulate the object?*

There was a significant difference in the naturalness of the interface ($F(2, 18) = 9.62$, $p < .01$). The naturalness of the gesture interface was rated as significantly different from that of the speech ($p < .01$) and that of the MMI ($p < .01$). The subjects felt that using the speech interface (mean = 5.60, std = 1.10) and the MMI (mean = 5.80, std = 0.83) were more natural than using the gesture only interface (mean = 4.60, std = 1.14).

6.6.4.2 Ease of Use of the Interfaces

We asked users how easy it was to change the colour and shape, of the objects and to point to and move the object with the following questions:

- *Q2: How easy was it to change the colour of the objects?*
- *Q3: How easy was it to change the shape of the objects?*
- *Q4: How easy was it to point to the objects?*
- *Q5: How easy was it to move the objects?*



	easy_colour	easy_shape	easy_point	easy_move
3D speech	5.90	5.85	4.92	4.54
3D gesture	4.05	4.15	4.00	4.77
3D MMI	5.90	6.00	4.69	5.85

Figure 6. 5 Users' feedback on the ease of use of the interfaces

Figure 6. 5 shows the average results for each question. We found a significant difference in the ease of the changing colour task ($F(2, 18) = 11.68$, $p < .01$). Using speech was significantly easier than using gestures ($p < .01$). The subjects felt that using the speech interface (mean = 5.90, std = 1.02) and the MMI (mean = 5.90, std = 1.02) to change the object colour were easier than using the gesture interface (mean = 4.05, std = 1.32).

There was a significant difference in the ease of the changing shape task ($F(2, 18) = 14.52$, $p < .01$). Using gesture to change the shape of the object was different from using speech ($p < .01$) and MMI ($p < .01$). The subjects also indicated that using the speech interface (mean = 5.90, std = 1.02) and the MMI (mean = 6.00, std = 0.97) were easier to change the shape of the object than using the gesture interface (mean = 4.00, std = 1.34).

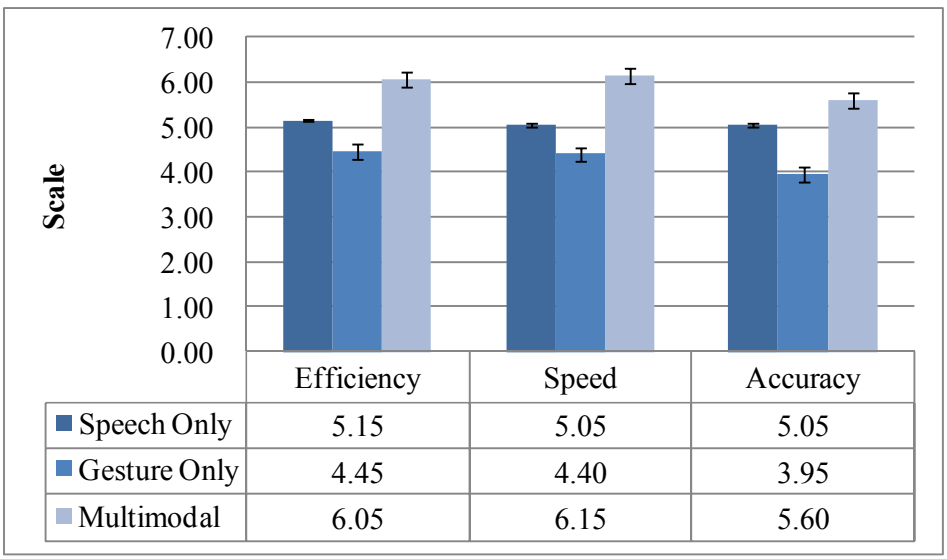
We also found a significant difference in the ease of moving tasks ($F(2, 18) = 7.54$, $p < .01$). The MMI was different from the gesture ($p < .03$) and the speech ($p < .04$). According to the result, the MMI (mean = 5.70, std = 0.98) was easier than the gesture (mean = 4.70, std = 1.53) and the speech (mean = 4.75, std = 1.29) for moving object tasks. However, there was no significant difference in the easiness of the pointing gestures ($F(2,18) = 1.83$, $p = .19$).

6.6.4.3 Interface Performance

We found a significant difference in the efficiency, speed, and accuracy of the interfaces. Overall, users felt that the MMI was the most efficient, fastest, and most accurate interface compared to the gesture and speech only interfaces.

We asked the subjects how they felt about the usability of each interface with the following questions:

- *Q6: I could perform the task efficiently.*
- *Q7: I performed the task quickly with this interface.*
- *Q8: I performed the task accurately with this interface.*



	Efficiency	Speed	Accuracy
Speech Only	5.15	5.05	5.05
Gesture Only	4.45	4.40	3.95
Multimodal	6.05	6.15	5.60

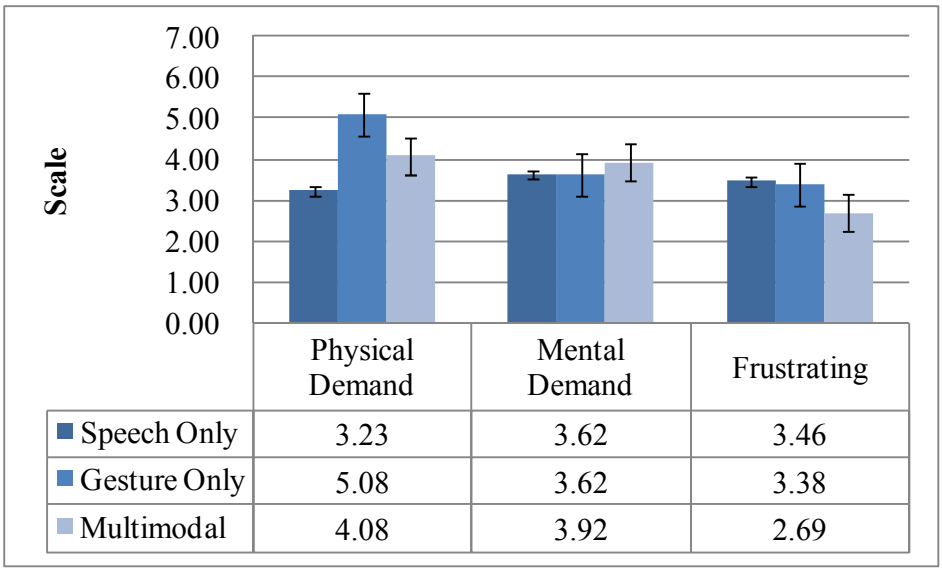
Figure 6. 6 User Feedback on efficiency, speed, and accuracy

Figure 6. 6 shows the average results across study conditions. The efficiency of the MMI was significantly different from that of the gesture interface and the speech interface ($F(2, 18) = 12.61$, $p < .01$). From the analysis, we found that users felt that the MMI (mean = 6.05, std = 1.05) was more efficient than the gesture only interface (mean = 4.45, std = 1.28) and the speech only interface (mean = 5.15, std = 1.14). For fast interaction, they indicated that the interaction with the MMI (mean = 6.15, std = 0.93) was quicker than with the gesture-only (mean = 4.40, std = 1.35) or the speech only interface (mean = 5.05, std = 1.19). We found a significant difference in the speed of interaction ($F(2, 18) = 14.83$, $p < .01$). The MMI was different from the speech input ($p< .01$) and the gesture input ($p< .01$). The users felt that they interacted with the MMI faster than with the speech input or gesture input. There was a significant difference in the accuracy of the interaction ($F(2, 18) = 9.03$, $p < .01$). For the accuracy of the interaction, we found no significant differences between the MMI and speech input or between the speech and gesture input. However, the MMI was significantly different from gesture input ($p< .01$). For the accuracy of the interface, the users felt that they interacted more accurately with the MMI (mean = 5.60, std = 1.19) than with the gesture interface (mean = 3.95, std = 1.39).

6.6.4.4 Physical and Mental Demands of Interfaces

We also asked the subjects how they felt about the physical and mental demands of each interface with the following questions:

- *Q9: I found that using this interface was physically demanding.*
- *Q10: I found that using this interface was mentally demanding.*
- *Q11: I found that using this interface was frustrating.*



	Physical Demand	Mental Demand	Frustrating
Speech Only	3.23	3.62	3.46
Gesture Only	5.08	3.62	3.38
Multimodal	4.08	3.92	2.69

Figure 6. 7 User feedback on physical demand, mental demand, and frustration.

Figure 6. 7 shows the average results across the study conditions. The physical demand of the speech interface was significantly different from that of the gesture interface and the MMI ($F(2, 18) = 7.28$, $p < .01$). From the analysis,

we found that users felt that the speech interface (mean = 3.23, std = 1.64) was less physically demanding than the gesture only interface (mean = 5.08, std = 1.04) and the MMI (mean = 4.08, std = 1.50). However, there was no significant difference on the mental demand ($F(2,18) = 0.89$, $p = .43$) and frustrations ($F(2,18) = 2.10$, $p=.15$) of the interface.

6.6.5 Interviews

After the experiments, we asked users to rank the most preferred interface. In total 14 users preferred the MMI over the unimodal interfaces. 5 users preferred speech only and only one user preferred the gesture only interface. We asked users why the MMI was the most preferred interface. They preferred the MMI because they could use two different modalities simultaneously or sequentially and, as a result, they could complete the task more quickly and efficiently. Another reason was that the two different modalities compensated each other and let the users do the task more accurately. For example, speech was used to change the colour of the sample object, and gesture was useful in changing the position of the object. They also mentioned that the MMI was more intuitive way of interaction to complete the given tasks than the unimodal interfaces. The users said that they could feel that the using the MMI was becoming easier as they proceeded with the given tasks. We also asked users why they least preferred the gesture input. They indicated that it was

physically demanding to use the gesture interface and that they had to search the control menu to find the target colours and shapes. The users also mentioned that gesture was least preferred because it took a longer time than using other interfaces to complete the given tasks and less accurate than other interfaces.

6.6.6 Observations

We observed the number of commands used to complete each task. On average, the subjects issued 5.23 commands (std = 1.08) with the speech interface, 6.14 commands (std = 0.57) with the gesture interface, but with the MMI, users issued only 4.93 commands (std = 1.11) per task. We found significant difference between the number of commands used with each interface ($F(2,18) = 11.58$, $p < .01$) after applying one-way repeated measures ANOVA with post-hoc pair wise comparisons (with Bonferroni correction). The number of gesture commands was different from that of speech commands ($p < .01$) and that of MMI commands ($p < .01$). However, there was no significant difference between the speech and MMI.

We classified types of commands into two groups: (1) characteristic and (2) movement. The characteristic commands include shape and colour of an object, while the movement commands include grabbing (selecting), moving, and

dropping. We divided the MMI commands into two modalities: (1) gesture and (2) speech. As a result, a simultaneously or sequentially triggered MMI command was not recognized as a single command; instead, the MMI command was considered as multiple combinations of gesture and speech commands.

There was no significant difference on characteristic commands with the different types of interfaces. However, we found significant difference in the number of movement commands with each interface ($F(2,18) = 3.82$, $p < .04$). The speech (mean = 28.50, std = 2.07) required less commands than the gesture input (mean = 35.25, std = 1.42) for moving the objects ($p < .03$). We did not find difference between number of speech and MMI commands (mean = 33.15, std = 1.33) or between number of gesture and MMI commands for moving the objects.

We analysed the percentage of simultaneous and sequential multimodal commands during the study. Most of multimodal integrations were done sequentially. On average, 20.69% of MMI were triggered simultaneously and 79.31% were triggered sequentially. There were four users who triggered the multimodal input only sequentially. Speech command precedes gesture in only 1.12% of simultaneously integrated multimodal input.

We also observed that the subjects' reactions to the system errors differed for each interface. The subjects repeated the same commands which gave them system errors with unimodal interfaces. Most of the subjects did the same when they had system errors while using MMI. However, we found that three subjects switched modality when they have a system error while interacting with MMI.

For example, a user triggered a pointing gesture to change the shape of the sample object to cylinder and it did not work. The user spoke to the system *"cylinder."*, rather than repeating the same command.

6.7 Discussion

By comparing task completion time for three different interfaces, we found that the MMI significantly reduced interaction time in completing the given tasks than the gesture interface. This is partly because using the speech input for changing colour or shape of the objects took less time than pointing gesture. Additionally, in a MMI case, speech and gesture commands were triggered simultaneously. From the results, we observed that the MMI was more efficient than gesture interface for this AR application. We did not find any significant difference between the speech interface and the MMI.

From the video analysis result, we found that the speech interface produced the smallest number of user errors (0.41 errors/ task) compared with the gesture input (0.50 errors/ task) and the MMI input (0.42 errors/ task). Although, in the interview, the users said they felt MMI is getting easier than other interfaces, there was no significant difference on the number of user errors.

These two findings also correlated with the users' feedback that we received. They preferred the MMI to the gesture interface. Although speech recognition produced slightly fewer errors, they felt that the MMI was overall more accurate than the speech input. This is because performing the tasks well typically required a combination of speech and gesture input.

Previous research (Kaiser *et al.*, 2003) has found that speech input is helpful for descriptive commands, such as changing the colour or shape of an object, and that gesture input is useful for spatial commands, such as pointing or moving an object. The MMI takes advantage of the complementary nature of the two input modalities, combining descriptive and spatial commands. Although the users had little experience with speech interfaces, gesture interfaces, and MMIs, we could see that the users found that speech commands were more useful for descriptive commands and gesture commands were more helpful for spatial commands.

6.8 Conclusions

In this chapter, we described a pilot user study and a full user study exploring the usability of the seamless AR MMI for object manipulation, compared with speech-only and 3D hand gesture-only conditions. After running the pilot study, we found several problems, particularly with the lack of depth cues in the object rendering. Additionally, the large number of colour and shape selection conditions made it difficult to remember the correct speech input commands. This feedback from the users was used to modify the AR scene and available speech commands. In the full user study, we compared the three usability factors of (1) efficiency, (2) effectiveness, and (3) satisfaction for each interface. The values to measure each factors were (1) task completion time, (2) user and system errors, and (3) satisfaction (questionnaire), respectively.

We found that the MMI produced shorter task completion times and required a smaller number of commands to complete the task than gesture interface. Although the MMI produced more user errors and was less accurate than the speech input, 70% of users preferred the MMI overall. The subjects felt that the MMI was easier to use and more effective than the other two unimodal interfaces. We also observed that our multimodal fusion architecture was more accurate than using a simple combination of speech and gesture input. These

results imply that the multimodal interfaces may be more useful for AR applications than speech-only or gesture-only interfaces.

The usability of the interface also can be different from the easiness of the tasks. Thus, in future work, we need to count the effect of task levels as well. In addition, performance may be improved by adding a feedback channel to give the user information about the fusion result. We could also use this to build a learning module into the multimodal fusion architecture which would improve the accuracy of the MMI based on the users' behaviour. Finally, we should explore the value of MMI in a wider range of AR applications in addition to object manipulation, for example, an AR world navigation or an AR game application.

Chapter 7

Conclusions and Future Work

Augmented Reality (AR) provides an enhanced user experience by seamlessly superimposing computer generated information onto the real environment. Early research in AR was mainly focused on vision-based tracking or registration techniques. Those techniques are essential to seamlessly align the computer generated information and the real environment. Moreover, a natural interface which supports computer generated world and the real world at the same time is also necessary to provide a seamless connection between two worlds.

As we observed earlier, the combination of hand interface and speech would be useful for interactions in AR environments. However, there is no research which provides natural hand interface in 3D with a corresponding speech input in an AR environment. Additionally, user studies on AR MMI are not fully explored yet. Finally, a fusion architecture which is designed according to the users' interaction behaviour in an AR MMI environment has not previously been studied.

As a first user observation, we described a Wizard of Oz study for an AR multimodal interface and virtual model manipulation tasks that allowed users to use natural speech and gesture input. We found the frequencies of multimodal input and the optimal size of the multimodal input time window. Deictic gestures (65%) and metaphoric gestures (35%) were the main types of gestures used. We also found that subjects used same gestures with meanings that varied depending on how they moved and which speech command they used. Thus, we need to consider a context-based multi-signal fusion architecture to analyze them more accurately.

From the speech input analysis, we found that most of speech commands were given in phrases with a few discrete words (74%), and not full sentences (26%). Overall, in 94% of the multimodal commands, gesture commands came earlier than the corresponding speech commands.

After the formal study with the exploratory data, we found that the MMI used depended on task types, but not on display types; users, however, preferred the screen display over the handheld display. Thus, for the multimodal system interaction in AR environments, a screen display may be preferable. The size of time window for combining speech and gesture input depends on the tasks as well. Moreover, although users felt gesture input alone was a more natural

interface than speech or the combination of speech and gesture, 68% of the input involved combined speech and gesture commands.

In the second user observation study, we have developed a gesture classification method based on the hand distance from the user's body. We normalized the distance of the users hand based on the initial distance from a subject's hand to a target object with normalized time to exclude effects of object position in gesture pattern. Using this we observe that touching and moving gestures (metaphoric gestures) were triggered further away from the subjects' body than the pointing gesture (deictic gesture). We also compared the gesture pattern curves from the Mixed environment with the combined (Real and AR) gesture pattern curves.

From the subjective user study, we found the subjects felt that using gestures in the Real environment was more natural, easier, quicker, more accurate, and less mentally demanding than in the Mixed or in the AR environment. In addition, all the users answered that the Real environment was the easiest one to complete the given tasks compared to the other two environments.

We found a consistent pattern from the normalized pointing gesture curves. We also found that metaphoric gestures were triggered further away from the subject's body than the pointing gesture. However, we did not find a common

pattern from the touching or moving gesture curves. Additionally, although there is a certain pattern on the pointing gesture curves, it would not be easy to apply the pattern curves to predict the pointing gesture in real time.

Based on the two user observations, we analysed users' behaviour using MMI in an AR environment. These findings were adopted to design a multimodal fusion architecture with adaptive filters, and showed how this could be applied in an AR MMI application. We also showed the AR MMI application which is connected to the fusion architecture.

Finally, a simple AR MMI application was evaluated by running a pilot user study and a full user study exploring the usability of the seamless AR MMI for object manipulation, compared with speech-only and 3D hand gesture-only conditions. We found that the MMI produced shorter task completion time, and required a smaller number of commands to complete the task. Although the MMI produced more user errors and was less accurate than the speech input, users preferred the MMI overall. The subjects felt that the MMI was easier to use and more effective than the other two unimodal interfaces. We also observed that our multimodal fusion architecture was more accurate than using a simple combination of speech and gesture input. These results imply that the multimodal interfaces may be more useful for AR applications than

speech-only or gesture-only interfaces. The usability of the interface also can be different from the easiness of the tasks.

7.1 Design Recommendations

From the previous user studies, we observed users' behaviour using gestures or MMIs in AR environments. As a result, we have measured the user preference and usability of our AR MMI. In this subsection, we propose design guidelines for MMIs in AR environments which will be helpful for researchers who want to develop AR MMIs in various applications. These include:

- Phrase-based speech command
- Fast gesture recognition module
- Gesture triggered multimodal fusion.
- Audiovisual feedback
- Learning module in the multimodal fusion architecture

We recommend the researchers adopt speech commands for their MMI in a phrase form, not in a full sentence. This was also seen in the Chapter 3. When we observed users' speech patterns, we found that most of speech commands were in phrase, the combination of two or three words, not complete sentences. Not only was the phrase-based speech commands natural to users, but it was

also useful for more accurate speech recognition. In case of the speech commands in a full sentence, we normally need to have a keyword spotting algorithm to find key words in the full sentence. However, we only have speech commands in a phrase. As a result, we can exclude any possibilities which would cause speech recognition errors.

A fast gesture recognition module is another aspect which is necessary for implementing AR MMIs. From the Wizard of Oz Study in Chapter 3, we recommended that researchers have an accurate gesture recognition for providing a better MMI experience to users. However, although there were differences between the interface accuracy, we could not find significant differences between the effectiveness of different modalities from the usability test in Chapter 6. This implies that the accuracy of gesture interface did not affect the effectiveness of the MMI. On the other hand, we found significant differences in efficiency and satisfaction of the interfaces. The gesture interface had a longer task completion time than speech or MMIs. Thus the gesture interface was less efficient than the speech interface or MMIs. To improve the efficiency of the gesture interface, we need to have a fast gesture recognition module.

In the final usability test (Chapter 6), we observed that only in 1.12% of simultaneously integrated multimodal input commands did speech input preceded the gesture input. We found a similar pattern in Chapter 3 where

94% of gesture commands preceded speech input. Thus, we need to have a gesture-triggered MMI fusion architecture.

After the pilot study prior to the full usability test, we learned that users wanted to have proper audiovisual feedback for each gesture and speech command. Although the users had visual feedback after triggering each command by looking at the changed shape or colour of the sample object, users still want to have sound feedback, such as 'ding' or 'beep' sound, when the system change the colour or shape of the sample object. Additionally, we needed to provide enough visual cues for better depth perception. A virtual plane was used as a background for the AR scene and users said it was helpful to perceive the 3D position of the augmented virtual objects.

We designed our multimodal fusion architecture by observing users' behaviour while they were interacting with the MMI. As a result, we found that the meaning of the gesture could vary depending on the corresponding speech command. From the last user study, we found that the way a user made errors with multimodal interface were different from the speech and gesture only error patterns. We also found that the pattern of the user error was similar for each user. As a result, we assume that if we have a learning module in a multimodal fusion architecture, the accuracy of the multimodal integration would be improved after all.

There are also some limitations of the recommendations. We only considered gesture-speech combination as a multimodal input. Thus, the other combination of multimodal input would require other design recommendations. The given task was also limited to interaction with computer generated graphics; other types of AR contents could lead to different results and different recommendations. Additionally, we did not explore the effect from the misalignment of user's hand in a real environment with the one in an AR environment.

7.2 Future Research

In the future, we will compare user behaviours interacting with a MMI in AR and VR environments. It would be useful to compare how user interactions are varied in different environments. Additionally, we should explore the value of MMI in a wider range of AR applications in addition to direct object manipulation. For example, building plan design or product design can be considered as a domain to apply AR MMIs. Both of them include the decision process before building a house or selecting the final product design. It is required to provide 3D free hand gesture and speech commands to interact with virtual models. In most experiments, we adopted a screen-based AR environment. However, we did not consider peripheral perception and proprioception while the users interacted in the environment. This has been

left as a future work. For the field study, we need to extend our AR MMI for multiple users. For two or more users we should consider different work spaces, since a collaboration task is totally different from a single user task. Prior to building a user-centred collaborative AR MMI, we also need to observe users' behaviour in the particular domain. Additionally, an algorithm to recognize multiple user commands via voice should be considered because existing speech recognition engines in Microsoft Speech API 5.3 (2009) only assume there is a single user in a quiet environment. Thus, if there is more than one user, the speech recognition engine cannot recognize which command was given by which user. With the current gesture interface, we cannot have more than one hand in the AR view. For the collaborative AR MMI, we need to extend our gesture recognition algorithm to support two or more users interacting in the same collaborative space. Moreover, the architecture for multimodal signal fusion should be reconsidered as well to support two or more users. With the collaborative application, user studies will definitely follow to verify how the AR MMI works properly in a practical domain. We can also extend the application domain from the desk-top AR to mobile AR.

References

Azuma, R. T.: 1997, A Survey of Augmented Reality, *Presence: Teleoperators and Virtual Environments*, 6(4), 355-385.

Bærentsen, K.: 2001, Intuitive User Interfaces, Scandinavian Journal of Information Systems archive, 12(1-2), 29 – 60.

Bolt, R. A.: 1980, "put-that-there": Voice and gesture at the graphics interface, *Proceedings of the 7th annual conference on Computer graphics and interactive techniques,* 262-270.

Borgefors, G.: 1986, Distance Transformations in Digital Images, *Computer Vision, Graphics and Image Processing*, 34, 344-371.

Chai, D. and Bouzerdoum, A.: 2000, A Bayesian approach to skin color classification in YCbCr color space, *Proceedings of IEEE TENCONO '00,* 2, 421-424.

Chu, C., Dani, T., 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.

Cohen, P. R., Dalrymple, M., Moran, D. B., Pereira, F. C., Sullivan, J. W., Jr, R. A. G., Schlossbert, J., and Tyler, S. W.: 1989, Synergistic use of direct manipulation and natural language, *CHI '89: Proceedings of the SIGCHI conference on Human factors in computing systems: Wings for the mind*, 227-233.

Cohen, P. R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L., and Clow, J.: 1997, Quickset: Multimodal interaction for distributed applications, *Multimedia '97*: *Proceedings of the International Multimedia Conference*, 31-40.

Cohen, P.R., Coulston, R. and Krout, K.: 2002, Multimodal interaction during multiparty dialogues: Initial results, *ICMI 2002: Proceedings of IEEE International Conference on Multimodal Interfaces*, 448-452.

Corradini, A. and Cohen, P.: 2002, On the Relationships among Speech, Gestures, and Object Manipulation in Virtual Environments: Initial Evidence, *Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems*, 52-61.

Coutaz, J., Salber, D., Carraux, E., and Portolan, N.: 1996, NEIMO, a multiworkstation usability lab for observing and analyzing multimodal

interaction, *Conference companion on Human factors in computing systems: common ground*, 402-403.

Dumas, B., Lalanne, D., and Oviatt, S.: 2009, Multimodal Interfaces: A Survey of Principles, Models and Frameworks, *Lecture Notes In Computer Science, Human Machine Interaction: Research Results of the MMI Program.* Springer-Verlag, Berlin/Heidelberg, 3-26.

Freeman, H.: 1974, Computer processing of line-drawing images, *Computing Surveys*, 6, 157-97.

GLUT: 2009, The OpenGL Utility Toolkit: http://www.opengl.org/ resources/libraries/glut/.

Hansson, P., Wallberg, A., and Simsarian, K.: 1997, Techniques for "natural" interaction in multi-user CAVE-like environments, ECSCW '97: *Proceedings of the European conference on computer Supported Cooperative Work.*

Hartley, R. and Zisserman, A.: 2004, *Multiple View Geometry in Computer Vision*. Cambridge University Press, UK.

Hauptmann, A. G.: 1989, Speech and gestures for graphic image manipulation, *CHI '89: Proceedings of the SIGCHI conference on Human factors in computing systems: Wings for the mind*, 241-245.

Heidemann, G., Bax, I., and Bekel, H.: 2004, Multimodal Interaction in an Augmented Reality Scenario, *ICMI '04: Proceedings of International Conference on Multimodal Interfaces*, 53-60.

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, *Proceedings of International Conferences on Artificial Intelligence and Teleextistance*, 272-283.

Irawati, S., Green, S., Billinghurst, M., Duenser, A., and Ko, H.: 2006, "move the couch where?": Developing an augmented reality multimodal interface, *ISMAR '06: Proceedings of IEEE/ACM International Symposium on Mixed and Augmented Reality*, 183-186.

Irawati, S., Daniela C., and Ko, H.: 2006b, Spatial ontology for semantic integration in 3D multimodal interaction framework, Proceedings of the 2006 ACM International Conference on Virtual Reality Contimuum and its Applications, 129-135.

Ishii, H. and Ullmer, B.: 1997, Tangible bits: Towards seamless interfaces between people, bits and atoms, *Proceedings of the SIGCHI conference on Human factors in computing systems*, 22-27.

Johnston, M., Cohen, P.R., McGee, D., Oviatt, S.L, Pittman, J.A., and Smith, I. : 1997, Unification-based Multimodal Integration, *Proceedings of the 35th annual Meetings of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, 281-288.

Kaiser, E., Olwal, A., McGee, D., Benko, H., Corradini, A., Li, X., Cohen, P., and Feiner, S.: 2003, Mutual disambiguation of 3D multimodal interaction in augmented and virtual reality, *ICMI '03: International Conference on Multimodal Interfaces (Aug 2003)*, 12-19.

Kato, H., and Billinghurst, M.: 1999, Marker tracking and HMD calibration for a video-based augmented reality conferencing system, *IWAR '99: Proceedings of International Workshop on Augmented Reality*, 85-94.

Kato, H., Billinghurst, M., Poupyrev, I., Imamoto, K., and Tachibana, K.: 2000, Virtual object manipulation on a table-top AR environment, *ISAR 2000: Proceedings of IEEE and ACM International Symposium on Augmented Reality*, 111~119.

Kölsch, M.: 2004, Vision based hand gesture interfaces for wearable computing and virtual environments, *Ph. D. Dissertation*, University of California, Santa Barbara.

Kölsch, M., and Turk, M.: 2004, Analysis of rotational robustness of hand detection with a Viola-Jones detector, *ICPR 2004: Proceedings of International Conference on Pattern Recognition,* 3, 107-110.

Kölsch, M., and Turk, M.: 2004a, Fast 2D hand tracking with flocks of features and multi-cue integration, *Proceedings of 2004 Conference on Computer Vision and Pattern Recognition Workshop*, 158-158.

Kölsch, M., Turk, M., and Tobias, H.: 2006, Multimodal Interaction with a Wearable Augmented Reality System. *IEEE Computer Graphics and Applications*, 26(3), 62-71.

Koons, D. and Sparrell, C.: 1994, ICONIC: speech and depictive gestures at the human-machine interface, *Proceedings of CHI '94: Conference companion on Human factors in computing systems*, 453-454.

Krum, D. M., Omotesto, O., Ribarsky, W., Starner, T., and Hodges, L. F. : 2002, Speech and gesture control of a whole earth 3D visualization environment, *Proceedings of Joint Eurographics-IEEE TCVG Symposium on visualization*, 195-200.

LaViola, J.: 2000, MSVT: A virtual reality-based multimodal scientific visualization tool, *Proceedings of the Third IASTED International Conference on Computer Graphics and Imaging*, 221-225.

Lee, T. and Hoellerer, T.: 2007, Handy AR: Markerless inspection of augmented reality objects using fingertip tracking, *ISWC 2007: Proceedings of International Symposium on Wearable Computing*, 83-90.

Looser, J., Grasset, R., Seichter, H., and Billinghurst, M.: 2006, OSGART - A Pragmatic Approach to MR, *Proceedings of Industrial AR Workshop at International Symposium on Mixed and Augmented Reality (ISMAR) 2006*.

Lucente, M., Zwart, G. J., and George, A. D.: 1998, Visualization Space: A Testbed for Deviceless Multimodal User Interface, *In AAAI Spring Symposium on Intelligent Environments*, AAAI TR SS-98-02.

Looser, J.:2007, AR Magic Lenses: Addressing the Challenge of Focus and Context in Augmented Reality, *PhD Dissertation*, University of Canterbury.

McNeil, D.:1992, *Hand and Mind: What gestures reveal about thought,* University of Chicago Press, Chicago, IL, USA.

Man, W. T., Qui, S. H., and Hong, W. K.: 2005, Thumbstick: A novel virtual hand gesture interface, *Proceedings of IEEE International Workshop on Robots and Human Interactive Communication*, 300-305.

Medl, A., Marsic, I., Andre, m., Liang, Y., Shaikh, A., Burdea, G., Wilder, J., Kulikowski, C., and Flanagan, J.: 1998, Multimodal Man-Machine Interface for Mission Planning, *Proceedings of the AAAI Spring Symposium on Intelligent Environments*, 41-47.

Microsoft Speech API (SAPI) 5.3: 2009, http://msdn.microsoft.com/en-us/library/ms723627(VS.85).aspx.

Molin, L.: 2004, Wizard-of-Oz Prototyping for Cooperative Interaction Design of Graphical User Interfaces, *Proceedings of NordiCHI'04*, 425-428.

Olwal, H. Benko, and S. Feiner.: 2003, SenseShapes: Using Statistical Geometry for Object Selection in a Multimodal Augmented Reality System. *Proceedings of International Symposium on Mixed and Augmented Reality*, 300-301.

OpenCV Library: 2009, http://sourceforge.net/projects/opencvlibrary/.

Oviatt, S. L., Cohen, P. R., Fong, M. W., and Frank, M. P.: 1992, A rapid semi-automatic simulation technique for interactive speech and handwriting,

Proceedings of International Conference on Spoken Language Processing, 2, 1351-1354.

Oviatt, S. L., Cohen, P. R., and Wang, M.: 1994, Toward interface design for human language technology: Modality and structure as determinants of linguistic complexity, *Speech Communication*, 15(3-4), 283-300.

Oviatt, S., DeAngeli, A., and Kuhn, K.: 1997, Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction. *CHI '97: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Atlanta, Mar. 22-27, 1997)*. ACM Press, New York, 415-422.

Oviatt, S., Cohen, P., Wu, L, Vergo, J., Duncan, L., Suhm, B., Bers, J., Holzman, T., Winograd, T., Landay, J., Larson, J., and Ferro, D.: 2000, Designing the user interface for multimodal speech and gesture applications: State-of-the-art systems and research directions, *Human Computer Interaction,* 15(4), 263-322.

Oviatt, S.: 2003, Advances in Robust Multimodal Interface Design. *IEEE Computer Graphics and Applications (Sep. 2003)*, 23(5), 62-68.

Oviatt, S. L., Coulston, R., Lunsford, R.: 2004, When do we interact multimodally?: cognitive load and multimodal communication patterns, *ICMI*

2004: Proceedings of International Conference on Multimodal Interfaces, 129-136.

Poupyrev, I., Berry, R., Billinghurst, M., Kato, H., Nakao, K., Baldwin, L., and Kurumisawa, J.:2001, Augmented reality interface for electronic music performance, *Proceedings of International Conference on Human-Computer Interaction*, 805-808.

Poupyrev, I., Tan, D., Billinghurst, M., Kato, H., and Tetsutani, H.: 2001a, Tiles: A mixed reality authoring interface. *Proceedings of INTERACT 2001*, 334~341.

Piekarski, W. and Thomas, B.:2001, Tinmith - augmented reality with wearable computers running linux, *Proceedings of the 2nd Australian Linux Conference*.

Piekarski, W. and Thomas, B. Tinmith-hand: 2002, Unified user interface technology for mobile outdoor augmented reality and indoor virtual reality, *Proceedings of IEEE Virtual Reality Conference*.

Point Grey Research Inc: 2009, http://www.ptgrey.com/.

Pfleger, N.: 2006, Context Based Multimodal Fusion, *ICMI 2004: Proceedings of International Conference on Multimodal Interfaces (State*

College, PA, USA, October 13-15, 2004), ACM Press, New York, NY, 265-272.

Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X. Kirbas, C., McCullough, K.E., and Ansari, R.: 2002, Multimodal human discourse: gesture and speech. *TOCHI: ACM Transactions on Computer-Human Interaction*, 9(3), 171-193.

Rauschert, I., Agrawal, P., Sharma, R., Fuhrmann, S., Brewer, I., MacEachren, A., Wang, H., and Cai, G.: 2002, Designing a human-centered, multimodal GIS interface to support emergency management, *GIS '02: Proceedings of the 10th ACM international symposium on Advances in geographic information systems*, (Sep. 2002, McLean, USA), 119-124.

Salber, D. and Coutaz, J.: 1993, Applying the Wizard of Oz Technique to the Study of Multimodal Systems, *Lecture Notes in Computer Science: Selected papers from the International Conference on Human-Computer Interaction,* 753, 219-230.

Sharma, R., Pavlovic, V.I., and Huang, T.S.: 1998, Toward Multimodal Human-Computer Interface, *Proceedings of IEEE* (May 1998), 86(5), 853-860.

Sharma, R., Yeasin, M., Krahnstoever, N., Rauschert, I., Cai, G., Brewer, I., Maceachren, A. M., and Sengupta, K.: 2003, Speech-Gesture Driven

Multimodal Interfaces for Crisis Management. *Proceedings of the IEEE,* 91(9), 1327-1354.

Sowa, T. and Wachsmuth, I.: 2003, Coverbal Iconic Gestures for Object Descriptions in Virtual Environments: An Empirical Study. *Proceedings of "Gestures. Meaning and Use"*, 365-376.

Swan II, J.E. and Gabbard, J. L.: 2005, Survey of User-Based Experimentation in Augmented Reality. *Proceedings of the first International Conference on Virtual Reality, HCI International 2005*.

Thomas, B., Demczuk, V., Piekarski, W., Hepworth, D., and Gunther, B.: 1998, A wearable computer system with augmented reality to support terrestrial navigation. *ISWC '98: Proceedings of International Symposium on Wearable Computers (1998)*, 168~171.

Tse, E., Greenberg, S., Shen, C.: 2006, GSI demo: multiuser gesture/speech interaction over digital tables by wrapping single user applications, *ICMI '06: Proceedings of the 8th International Conference on Multimodal Interfaces*, 76-83.

Wang, J.: 1995, Integration of eye-gaze, voice and manual response in multimodal user, *Proceedings of IEEE International Conference on Systems, Man and Cybernetic*, 5, 3938-3942.

Wark, T., Sridharan, S., and Chandran, V.: 1999, Robust speaker verification via fusion of speech and lip modalities, *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 6, 3061-3064.

Weimer, D. and Ganapathy, S. K.: 1989, A synthetic visual environment with hand gesturing and voice input, *CHI '89: Proceedings of the SIGCHI conference on Human factors in computing systems: Wings for the mind*, 235-240.

Zhang, Z.: 2000, A Flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1330-1334.

Zhou, Z., Cheok, A. D., Li, Y., and Kato, H.: 2005, Magic cubes for social and physical family entertainment, *CHI '05: Proceedings of International Conference on Human Factors in Computing Systems: extended abstracts on Human factors in computing systems*, 1156~1157.

Zhou, Z., Cheok, A. D., Pan, J., and Li, Y.:2004, An interactive 3D exploration narrative interface for storytelling, *Proceedings of the 2004*

Conference on Interaction Design and Children: Building a Community, 155~156.

Zhu, X., Yang, J., and Waibel, A.: 2000, Segmenting hands of Arbitrary Color, *Proceedings of International Conference on Automatic Face and Gesture Recognition*, 446-453.

Appendix A

Wizard of Oz Study Questionnaire

Pre-Experiment Questionnaire

Subject: _____

Sex: M / F

Age: _____ years

Are you native English speaker?

☐ Yes or ☐ No

How familiar are you with speech interfaces and speech recognition?

1	**2**	**3**	**4**	**5**
Not Very Familiar				Very Familiar

How familiar are you with gesture interfaces?

1	**2**	**3**	**4**	**5**
Not Very Familiar				Very Familiar

Post-Condition Questionnaire for a simple task

Subject: _____

	With respect to this simple task…	**Natural hand + speech + monitor**
Q1	It was natural to use speech input in this task	**1** **2** **3** **4** **5** Disagree … Agree
Q2	It was natural to use gesture input in this task	**1** **2** **3** **4** **5** Disagree … Agree
Q3	I felt that it was natural to manipulate the virtual object with combined speech and gesture input.	**1** **2** **3** **4** **5** Disagree … Agree
Q4	I found this technique was enjoyable	**1** **2** **3** **4** **5** Disagree … Agree
Q5	It was easy to manipulate the virtual object	**1** **2** **3** **4** **5** Disagree … Agree
Q6	I felt that I performed the task quickly	**1** **2** **3** **4** **5** Disagree … Agree
Q7	I felt that I performed the task accurately	**1** **2** **3** **4** **5** Disagree … Agree
Q8	i think the use of speech helped me communicate descriptively with the system	**1** **2** **3** **4** **5** Disagree … Agree
Q9	I think the use of gestures helped me communicate spatially with the system	**1** **2** **3** **4** **5** Disagree … Agree
Q10	I found using this technique was physically demanding	**1** **2** **3** **4** **5** Disagree … Agree
Q11	I found using this technique was mentally demanding	**1** **2** **3** **4** **5** Disagree … Agree
Q12	I found this technique frustrating	**1** **2** **3** **4** **5** Disagree … Agree
Please add any comments about the condition.		

Post-Condition Questionnaire for Overall Conditions

Subject: _____

Q1. Which interface was the easiest to use?

_____ Multimodal Interface with Monitor

_____ Multimodal Interface with Handheld display

Q2. Which interface was the most enjoyable?

_____ Multimodal Interface with Monitor

_____ Multimodal Interface with Handheld display

Q3. Which interface do you prefer overall?

_____ Multimodal Interface with Monitor

_____ Multimodal Interface with Handheld display

Q4. Why the highest ranked interface is the most preferred? Please list three reasons.

Q3. Why the lowest ranked interface is the least preferred? Please list three reasons.

Q4. Do you have other comments for the improvement of the application?

Appendix B

Gesture Classification Questionnaire

Pre-Experiment Questionnaire

Subject: _____

Gender: M / F

Age: _____ years

Which handed are you?

☐ Left or ☐ Right

How familiar are you with AR applications?

1	2	3	4	5	6	7
Not Very Familiar						Very Familiar

How familiar are you with gesture interfaces?

1	2	3	4	5	6	7
Not Very Familiar						Very Familiar

Post-Condition Questionnaire for an interaction with augmented objects

Subject: _____

	With respect to the task…	
Q1	It was natural to use gesture input in this task	1 2 3 4 5 6 7 Disagree Agree
Q2	It was easy to point the objects	1 2 3 4 5 6 7 Disagree Agree
Q3	It was easy to touch the objects	1 2 3 4 5 6 7 Disagree Agree
Q4	It was easy to move the objects	1 2 3 4 5 6 7 Disagree Agree
Q5	I think wearing the thimbles affected my concentration when performing gestures.	1 2 3 4 5 6 7 Disagree Agree
Q6	I think putting hand in the preparation space is uncomfortable or unnatural.	1 2 3 4 5 6 7 Disagree Agree
Q7	How quickly did you perform the tasks?	1 2 3 4 5 6 7 Not very much very much
Q8	How accurately did you perform the tasks?	1 2 3 4 5 6 7 Not very much very much
Q9	How physically demanding was it to perform the task?	1 2 3 4 5 6 7 Not very much very much
Q10	How mentally demanding was it to perform the task?	1 2 3 4 5 6 7 Not very much very much
Q11	How frustrating was it to perform the task?	1 2 3 4 5 6 7 Not very much very much
Please add any comments about the condition.		

Post-Experiment Questionnaire for Overall condition

Subject: _____

Q1. Which environment was the easiest to use the gesture input?

_____ Augmented environment

_____ Real environment

_____ Mixed environment

Q2. Which environment was the most enjoyable to use the gesture input?

_____ Augmented environment

_____ Real environment

_____ Mixed environment

Q3. Which gesture was the most enjoyable to use the gesture input?

_____ Pointing

_____ Touching

_____ Moving

Q4. Which environment do you prefer overall?

_____ Augmented environment

_____ Real environment

_____ Mixed environment

Q5. Do you have other comments for the improvement of the application?

Appendix C

MMI Usability Questionnaire

Pre-Experiment Questionnaire

Subject: _____

Gender: M / F

Age: _____ years

Are you Right or Left handed?

☐ Left or ☐ Right

How fluent an English speaker are you?

1	2	3	4	5	6	7
Not Very Fluent						Very Fluent

How familiar are you with speech interfaces and speech recognition?

1	2	3	4	5	6	7
Not Very Familiar						Very Familiar

How familiar are you with gesture interfaces?

1	2	3	4	5	6	7
Not Very Familiar						Very Familiar

How familiar are you with multimodal interfaces?

1	2	3	4	5	6	7
Not Very Familiar						Very Familiar

How familiar are you with Augmented Reality applications?

1	2	3	4	5	6	7
Not Very Familiar						Very Familiar

Post-Condition Questionnaire for an interaction with speech interface : 3D

Subject: _____

For each of the following statements please mark how much you agree or disagree with the statement

	With respect to the task…	
S1	It was natural to manipulate the object.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree
S2	It was easy to change the colour of the objects.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree
S3	It was easy to change the shape of the objects.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree
S4	It was easy to point to the objects.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree
S5	It was easy to move the objects.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree
S6	I could perform the task efficiently with the interface.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree
S7	I performed the task quickly with this interface.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree
S8	I performed the task accurately with this interface.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree
S9	I found that using this interface was physically demanding.	**1 2 3 4 5 6 7** None … All
S10	I found that using this interface was mentally demanding.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree
S11	I found using this interface frustrating.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree

Q1	How much of the task did you complete? (0% - 100%)	
Q2	How did you feel about the task?	

Post-Condition Questionnaire for an interaction with gesture interface : 3D

Subject: _____

For each of the following statements please mark how much you agree or disagree with the statement

	With respect to the task…	
S1	It was natural to manipulate the object.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree
S2	It was easy to change the colour of the objects.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree
S3	It was easy to change the shape of the objects.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree
S4	It was easy to point to the objects.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree
S5	It was easy to move the objects.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree
S6	I could perform the task efficiently with the interface.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree
S7	I performed the task quickly with this interface.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree
S8	I performed the task accurately with this interface.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree
S9	I found that using this interface was physically demanding.	**1 2 3 4 5 6 7** None … All
S10	I found that using this interface was mentally demanding.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree
S11	I found using this interface frustrating.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree

Q1	How much of the task did you complete? (0% - 100%)	
Q2	How did you feel about the task?	

Post-Condition Questionnaire for an interaction with the multimodal interface : 3D

Subject: _____

For each of the following statements please mark how much you agree or disagree with the statement

	With respect to the task…	
S1	It was natural to manipulate the object.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree
S2	It was easy to change the colour of the objects.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree
S3	It was easy to change the shape of the objects.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree
S4	It was easy to point to the objects.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree
S5	It was easy to move the objects.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree
S6	I could perform the task efficiently with the interface.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree
S7	I performed the task quickly with this interface.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree
S8	I performed the task accurately with this interface.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree
S9	I found that using this interface was physically demanding.	**1 2 3 4 5 6 7** None … All
S10	I found that using this interface was mentally demanding.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree
S11	I found using this interface frustrating.	**1 2 3 4 5 6 7** Strongly Disagree … Strongly Agree

Q1	How much of the task did you complete? (0% - 100%)	
Q2	How did you feel about the task?	

Only for Multimodal Interfaces

SS1	I think there was no system delay in response to the multimodal input.	**1** **2** **3** **4** **5** **6** **7** None All
SS2	I think the integration of gesture and speech input was done accurately.	**1** **2** **3** **4** **5** **6** **7** Strongly Disagree Strongly Agree
SS3	I think the multimodal interface worked very well.	**1** **2** **3** **4** **5** **6** **7** Strongly Disagree Strongly Agree

Post-Experiment Questionnaire for Overall condition 3D

Subject: _____

For each of following questions please rank the interfaces in order.

Q1. Which interface was the easiest to perform the given tasks?

_____ Speech only
_____ Gesture only
_____ Multimodal

Q2. Which interface was the most efficient to perform the given tasks?

_____ Speech only
_____ Gesture only
_____ Multimodal

Q3. Which interface was the most natural to use?

_____ Speech only
_____ Gesture only
_____ Multimodal

Q4. Which interface do you prefer overall?

_____ Speech only
_____ Gesture only
_____ Multimodal

Q5. Why is the highest ranked interface the most preferred? Please list three reasons.

Q6. Why is the lowest ranked interface the least preferred? Please list three reasons.

Q7. How can the multimodal interface be improved?

Q8. Do you have other comments for improving the application?