A Wizard of Oz Study for an AR Multimodal Interface

Minkyung Lee and Mark Billinghurst
HIT Lab NZ, University of Canterbury
Christchurch 8014 New Zealand
+64-3-364-2349
{minkyung.lee, mark.billinghurst}@hitlabnz.org

ABSTRACT
In this paper we describe a Wizard of Oz (WOz) user study of an Augmented Reality (AR) interface that uses multimodal input (MMI) with natural hand interaction and speech commands. Our goal is to use a WOz study to help guide the creation of a multimodal AR interface which is most natural to the user. In this study we used three virtual object arranging tasks with two different display types (a head mounted display, and a desktop monitor) to see how users used multimodal commands, and how different AR display conditions affect those commands. The results provided valuable insights into how people naturally interact in a multimodal AR scene assembly task. For example, we discovered the optimal time frame for fusing speech and gesture commands into a single command. We also found that display type did not produce a significant difference in the type of commands used. Using these results, we present design recommendations for multimodal interaction in AR environments.

Categories and Subject Descriptors
H.5.1 Information Interfaces and Presentation: Multimedia Information Systems - Animations, Artificial, Augmented, and Virtual Realities.

General Terms
Human Factors, Experimentation

Keywords
Augmented Reality, multimodal interaction, multimodal interface, user study, Wizard of Oz, AR, HCI, WOz.

1. INTRODUCTION
Augmented Reality (AR) involves the real time overlay of computer graphics onto the real world. 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 must be used for aligning real and virtual objects, and natural interaction techniques for manipulating virtual content. However, although there is research on interaction techniques in AR, there is often little evaluation of these techniques [1].

ICMI'08, October 20–22, 2008, Chania, Crete, Greece.

In our research we want to develop and evaluate AR interaction techniques based on the user's natural real world behavior. Many current AR applications adopt general Virtual Reality (VR) or GUI interaction techniques [2][3]. However, these methods are designed for fully immersive virtual environments or desktop interfaces and ignore the connection between AR content and the real world. Thus, there is a need to research new interface metaphors ideally suited for AR.

The focus of our research is on multimodal input for AR interfaces. Multimodal interfaces (MMI) that combine speech and hand gesture input have previously been found to be an intuitive way to interact with 2D and 3D graphics desktop applications [4][5][6]. They can support interaction in the real world and on-screen virtual environments at the same time, and so should be ideal for AR interfaces. However, there has been little research on the use of multimodal input in AR interfaces, and especially usability evaluations of AR multimodal interfaces.

One of the most important questions in developing a multimodal interface is what natural speech and gesture commands should be used. In the past researchers have used Wizard of Oz (WOz) techniques to capture natural speech and gesture input [6][7]. These studies simulate perfect speech and gesture recognition to allow subjects to freely use any commands they want. In this way researchers can collect a corpus of multimodal commands.

In this paper we present the first user study exploring a multimodal AR interface with a WOz technique. We have undertaken this research as a first step towards developing intuitive multimodal input for AR applications and our results will be useful for other researchers wanting to develop multimodal AR interfaces. The main contribution is to provide observations on users' behavior and preference when they interact with an AR application using multimodal input.

In the rest of the paper we will first present related work (Section 2), and then we give an overview of our WOz system which combines computer vision based natural hand tracking with simulated speech input (Section 3). Next we present a formal user study conducted with our system and an analysis of the results (Section 4). Finally we use these results to develop design guidelines for AR multimodal interfaces and future research.

2. RELATED WORK
Our work is based on previous research in multimodal interfaces, multimodal AR interfaces, and Wizard of Oz studies.

Multimodal interfaces with gesture and speech input have a long history dating back to the "Put that there" work of Bolt [4]. He used pointing gestures with speech as an interaction channel in a 2D graphics application and showed that combining speech and

gesture input creates an interface that is more powerful than using either input modality alone. Since then, Cohen and Sullivan [5] showed how a mixture of natural language and direct manipulation can overcome the limitations of each modality. Speech and gesture have complementary attributes and they found that combining them provides a more transparent way for interacting with applications than in previous GUI interfaces.

Many previous multimodal interfaces are map- or screen-based applications [8][9][10]. In this case it is easy to use pen input or a touch screen for stable gesture recognition. However, in our AR applications we wanted to support natural 3D object interaction. Previously other researchers have used speech input for descriptive commands and used hand tracking devices or DataGloves [11][12][13] to explore gesture input in 3D graphics environments. Alternatively, computer vision based hand tracking techniques have been used in systems such as "VisSpace" [14] to estimate where users were pointing. Raushcert et al. [9] also demonstrated a 3D graphics multimodal interface with speech and vision-based gesture input. However, their system did not support natural 3D object interaction as they were only concerned with where users were pointing.

There has been little research on multimodal input in AR interfaces. One of the first multimodal AR interfaces, SenseShapes [11], used volumetric regions of interest that were attached to the user's gaze direction or hand to provide visual information about interaction with virtual objects. Object selection was available with a data glove to detect user's gestures and with trackers to monitor hand position for interaction with objects. Speech recognition provided information about where the user wanted to move an object, interpreting "this" or "that" spoken commands with deictic pointing gestures. The user had to wear a data glove and the researchers did not conduct user studies to explore the effectiveness of SenseShapes. Irawati et al. [15] has developed a computer vision based multimodal AR system by adding speech input to the VOMAR furniture arranging application [16]. 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. Irawati et al. conducted a pilot user study on the benefits of multimodal interaction [17]. However, their system did not support natural free hand input and users had to memorize or refer a list of commands to interact with virtual objects.

Several researchers have also explored computer vision input in multimodal AR interfaces. Kölsch et al [18] developed a multimodal information visualization system with natural hand tracking in a wearable AR environment. Similarly, HandVu [19] was an AR application that recognized users' hand gestures from texture and colour. However, the output in both cases was the user's hand location in 2D image coordinates which could not be easily used to manipulate augmented virtual objects in 3D space.

To provide a natural multimodal interface we need to know what speech and gesture commands users would like to use if there were no technical limitations. 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. Salber and Coutaz [20] provide a good overview of how WOz techniques can be applied to a multimodal interface. Their NEIMO system [21] uses these methods in a multimodal usability lab, although they have not explored AR and VR systems. There are many examples of how WOz techniques can be used for system prototyping in various research areas. For example, Oviatt et al. [22][23] 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 graphics applications. For example, Hauptmann [7] provides 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 gesture input was the preferred method for manipulating the blocks. Corradini and Cohen [24] describe using a WOz technique for navigating through a virtual environment. Molin [25] also made a WOz prototype for cooperative interaction design of graphical interfaces. After this WOz study, Molin concludes that the WOz experience triggers an analysis of the interaction which produces new design ideas that can be tested, and the recordings of screen and video can 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 interaction. 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.

Our research 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 will be useful for others trying to develop multimodal AR interfaces and lays the foundation for a significant amount of future work.

3. AR WIZARD OF OZ SYSTEM

From previous research we can learn that 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/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.

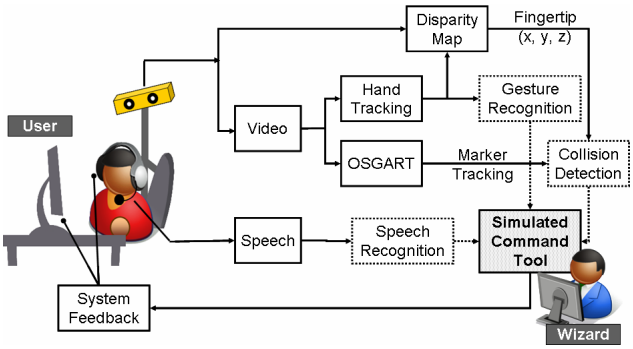


Figure 1: Software components of our AR WOz system.

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 additional tools for supporting the WOz experiment. In this section we describe our system in more detail. Figure 1 shows how the system components are connected.

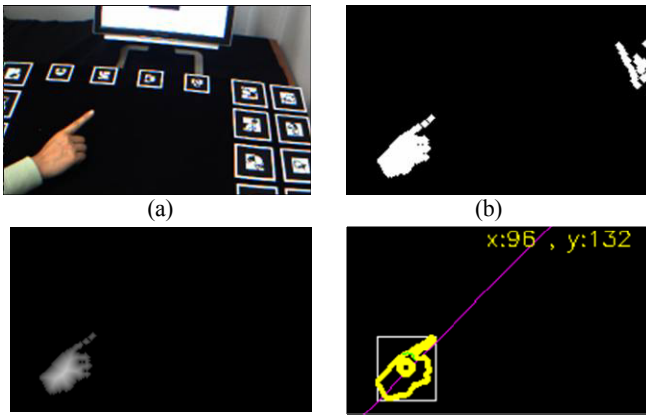
3.1 3D Natural Hand Interaction

It is not easy to simulate 3D natural hand interaction in real time in a WOz application, so we have implemented a 3D vision-based hand tracking system. Our hand tracking is based on three methods: (1) segmenting skin colour, (2) finding feature points for the center of the palm and fingertips, and (3) finding the hand direction. We used a BumbleBee2 stereo camera and our software is based on the OpenCV library [26].

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. 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.

Following the skin colour segmentation, we find the biggest contour [27] of the segmented area to extract the user's hand more accurately. Afterwards, a distance transformation [28] 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 center of palm and the fingertip, are mapped to a disparity map to estimate the 3D information of each point for interaction with AR objects.

Figure 2 shows the results from the hand tracking algorithm. We were able to track the user's fingertip with accuracy from 3mm up to 20mm depending on display type and 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.



(a) (b) (c) (d)

Figure 2: Hand segmentation results: (a) reference image, (b) skin colour segmented image, (c) finding the hand centre, and (d) fingertip and hand direction finding.

3.2 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 shows the dialog menu used by the Wizard to quickly input commands. It has three functions for replacement of gesture commands ("pick-up", "drop", and "delete"), and two groups for speech: "change colour" and "change shape".

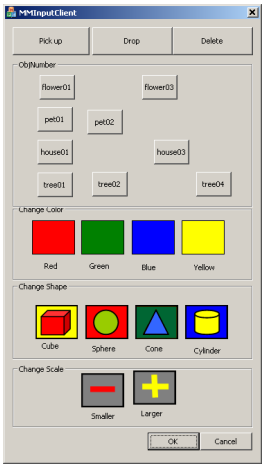


Figure 3: The simulated command menu.

3.3 Augmented Reality View

To provide an AR view we used the OSGART rendering and interaction library [29] with the ARToolKit [30] computer vision input to track the user's real camera position relative to square fiducial markers. Once the camera position is known, OSGART can create a realistic 3D graphics rendering which is overlaid on the live video view to create an AR view. We added shade and shadows to improve the realism of the AR scene.

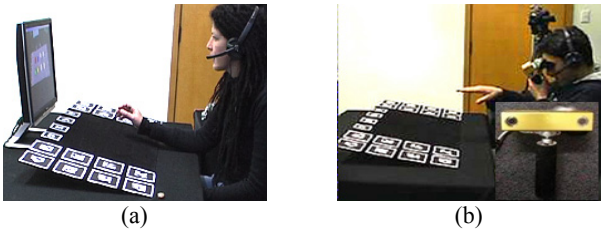
4. USER STUDY SETUP

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, and also if the display conditions would have any effect on the multimodal input. In this section we describe our experimental set up and tasks, while in the next section we present the results.

4.1 Experiment Set Up

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, how easy they found it to complete the task, etc.

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 for each of the two display conditions; a screen display (Figure 4(a)) and a Hand Held Display (HHD) (Figure 4(b)).



(a) (b)

Figure 4: System hardware configurations: (a) Screen-based AR system and (b) HHD-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 screen while the BumbleBee camera was fixed to show a view of the workspace in front of it. This view was combined with 3D virtual image overlay to create an AR view shown on the screen. The simulated command menu (see Figure 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.

4.2 Experimental Tasks

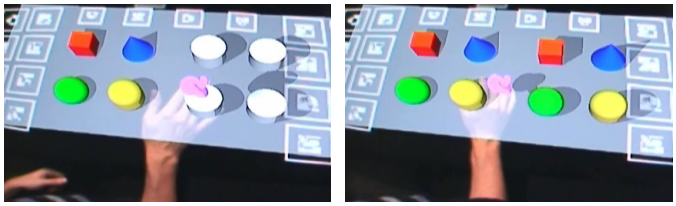
The experiment consists of subjects performing three simple tasks involving virtual object manipulation. Most interaction in an AR environment involves one or more of; moving virtual objects, rotating or translating virtual objects, or changing object colour or shape. Thus, we designed our tasks to include these interactions. The available interaction sub-tasks are shown in Table 1.

Table 1: Task types and available interaction modes.

	Task 1	Task 2	Task 3
Changing colour	√	√	X
Changing shape	√	X	X
Selecting object	2D	3D	2D/3D
Moving object	2D	3D	2D/3D

4.2.1 Simple Task I

For the first task (see Figure 5) 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. The users were supposed to change the colour and shape of four white cylinders to the same shape and colour of target objects. Subjects needed to let the system know the color or shape of which object they wanted to change. However, they could not change the position of any object displayed. In this case, the virtual objects were positioned on a table so gesture input was a largely 2D task.

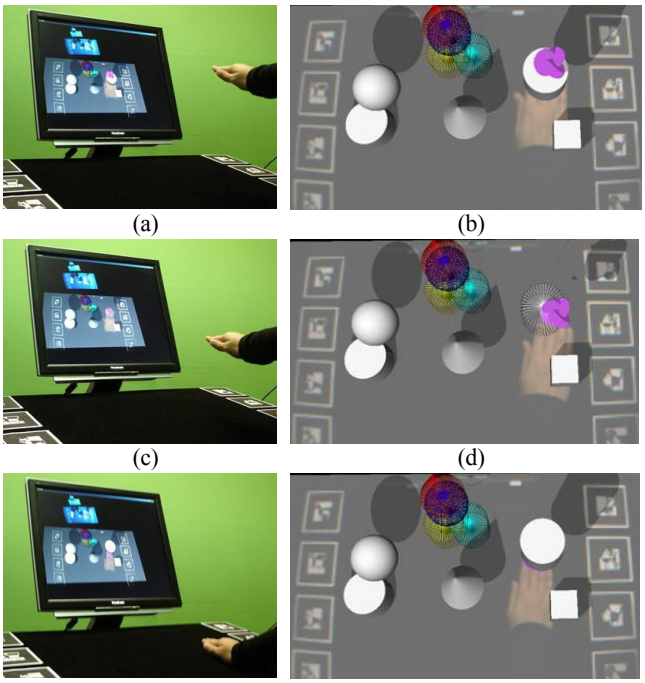


(a) (b)

Figure 5: Simple task I: (a) an initial scene and (b) final scene.

4.2.2 Simple Task II

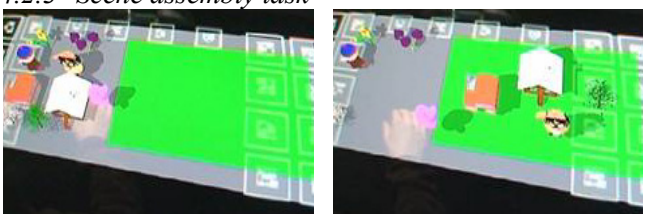
The second task involved moving sample objects distributed in 3D space into a final target 3D arrangement of objects (see Figure 6). The subjects needed to move their hands in all three directions to select and move objects. Figure 6 shows the system recognizing a user's hand in 3D. When the user's hand is located within the object, then the system recognizes it as a collision (Figure 6 (c) and (d)) 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.



(a) (b) (c) (d) (e) (f)

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

4.2.3 Scene assembly task



(a) (b)

Figure 7: Scene assembly task.

The final task was to create a scene with prebuilt 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. Figure 7(a) shows the initial AR scene and Figure 7(b) shows one user's final result.

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 a 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.

5. RESULT AND ANALYSIS

Video data of user interaction was collected from each of the task conditions for all subjects. 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 the 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.

5.1 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 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 (sd=0.66) words long and the sentences used were 2.94 (sd=1.08) words long. There was no significant change in speech patterns over time.

Table 2: The numbers of words used for speech input.

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

5.2 Gesture Frequency

From the experiment video we analyzed users' gestures according to the gesture classification scheme of McNeill [31] (Deictic, Metaphoric, Iconic, and Beat-like gestures). The classifications of the gesture are the following:

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

Table 3 shows 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: The numbers of gestures used.

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

5.3 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.

The Multimodal window, a time frame to combine gesture and speech input is shown in Figure 8, and is made up of the following:

- *Gesture Window:* how long the users holds a particular gesture for
- *Speech Window:* how long it takes 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 finished.

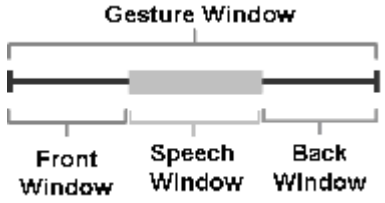


Figure 8: The definition of Multimodal window.

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. 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.

Table 4: Overall window sizes (Seconds)

	Display	Gesture Window	Speech Window	Front Window	Back Window
Screen	Task1	7.64 (1.67)	3.09 (1.70)	4.18 (1.33)	2.73 (0.79)
	Task2	8.33 (1.97)	2.58 (1.56)	4.75 (1.29)	3.83 (1.34)
	Task3	7.40 (1.27)	1.90 (0.88)	4.70 (0.95)	3.90 (1.20)
HHD	Task1	7.20 (1.55)	3.30 (1.42)	2.80 (1.03)	3.40 (1.17)
	Task2	8.91 (2.47)	2.73 (1.56)	5.27 (1.62)	3.55 (0.93)
	Task3	7.80 (1.23)	2.10 (0.74)	5.10 (1.20)	3.90 (0.99)

The mean size of the gesture time window which covers up to 98% of gesture time windows was 7.9 seconds (sd=1.20), the mean size of the speech time window was 2.6 seconds (sd=1.41), the mean size of the front window was 4.5 seconds (sd=1.46), and the mean size of the back window was 3.6 seconds (sd=1.13). Each window size with different task and display conditions is shown in Table 4.

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 inputs 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.

5.4 Dependences on task or display type

We used a two factor (task type, display type) repeated measures ANOVA with post-hoc pairwise 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.

5.4.1 Dependences of speech input

The numbers of words for colour ($F(2,10)=7.212$, $p=0.012$), shape ($F(2,10)=19.843$, $p<0.001$), and miscellaneous commands ($F(2,10)=9.520$, $p=0.005$) 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 ($p=0.010$). 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.

5.4.2 Dependences of gesture input

A two factor (task type, display type) repeated measures ANOVA with post-hoc pairwise 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 ($F(2,10)=10.023$, $p=0.004$). Task 1 was significantly different from task 2 ($p=0.003$) because gestures in task 1 were all pointing gestures. Therefore, compared with task 2 which included more other gestures, task 1 had more deictic gestures than task 2. In case of metaphoric gestures, there was a significant difference across task type ($F(2,10)=13.676$, $p=0.001$). Task 1 was significantly different from task 2 ($p=0.001$) 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 ($F(2,10)=119.207$, $p<0.001$). Task 1 was different from task 2 ($p<0.001$) and task 3($p<0.001$). 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.

5.4.3 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 ($F(2,6)=8.145$, $p=0.020$). Task 1 was different from task 2 ($p=0.041$) and task 3($p=0.025$). Task 1 had a longer speech timing window (mean=3.50, sd=0.34) than task 2 (mean=2.69, sd=0.35) and task 3 (mean=2.00, sd=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 back time window among task types ($F(2, 6) = 9.297$, $p = 0.015$). Task 1 showed a smaller size of the back time window than task 3.

5.5 Interview

We asked users to pick one display type based on (1) their preference, (2) enjoyableness, and (3) ease of use. In total 66.7% of people both 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, when users were using the screen display, there was nothing special about it and the AR experience it provided was not as immersive or compelling.

On the other hand, the HHD provided a natural AR view to users because the viewpoint of the camera was exactly the same as where the users were looking. The novelty of the HHD was also attractive to users. However, the HHD did have disadvantages compared with the screen display. Holding the HHD for the whole task was physically demanding, and the tracking was not as good as the screen display because the camera moved around according to the users' view. The users' interaction area also 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.

In interviews after the experiments, 75% of users said they did not feel it was natural to talk to the computer. Moreover, all of the users did not want to talk with the computer in the same conversational way as they did with other people.

5.6 Observations

We have several observations from watching subjects complete the experiment. First of all, when the wizard did not react to their gesture commands properly, most users repeated the same command again. However, in case of speech commands, they tried other commands for the system. If the wizard made a mistake simulating the users' command, the users thought they did something wrong, not the system. We also found that not having a fixed command set made some users initially frustrated. For example, one user said *"What can I say?",* and then tried to figure out which commands were available by saying *"Move the target. Does it work?*" However, when they learned how the system worked, they interacted more quickly with it. In this case, although the user changed the target object to a box, they still tried to change other object shapes with other commands, for example, *"Change it to a dice. Change this to a cube. Oh, it works as well!"*

Although subjects used a small number of gestures, the gestures had 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. The meaning of the gesture changed according to the corresponding speech input or with certain movement of user's hand. We also observed that users hold the same gesture while they were moving objects, as shown in Figure 9.

Figure 9: Hand gesture while moving the object.

We also observed the user's head movement while they were using handheld display device. As shown in Figure 10, the users changed their head pose to change the AR view depending on their viewpoint or to move in closer to the AR scene.

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

6. DISCUSSION

Subjects felt that using gestures was the most natural input technique for them. However, when we looked at the usage of speech and gesture, combined speech and gesture input was the most used input modality. Counting the number of commands issued, commands that combined speech and gesture input were 63% of the total (49% combined word command and gesture, and 14% combined sentence command and gesture), whereas gesture input only commands were 34%, and speech only input was 3.7% (0.4% of word command and 3.3% of sentence command). This implies that multimodal AR interfaces for object manipulation will rely heavily on accurate recognition of input gestures, as almost 97% of commands involved gesture input.

We expected that the display type would affect the way users interacted with the virtual content since the size of the interaction area varied according to display type. From the analysis results, none of factors showed a significant difference due to display type. However, 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 users use similar multimodal speech and gesture patterns in an AR interface regardless of display type.

6.1 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, including:

- Use a gesture-triggered MMI system to reduce delay
- Use dynamic gesture recognition algorithms
- Make sure that the gesture recognition input is as accurate as possible, and is particularly good at recognizing deictic and metaphoric gestures.
- Use key word spotting for better speech recognition
- 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, 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.

We need to have dynamic gesture recognition algorithms. From observing the users, we found that almost all the gestures were either deictic or metaphoric. The users used the same static gesture in different conditions which meant the meaning of gesture changed depending upon the context of use.

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.

A keyword spotting algorithm for speech commands is necessary to improve speech recognition results. This is particularly the case because 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 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.

7. CONCLUSIONS AND FUTURE WORK

In this paper we described a Wizard of Oz study for an AR multimodal interface and model manipulation tasks. We found the frequencies of multimodal inputs 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 users used same gestures with meanings that varied depending on how the users 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 dynamic 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.

8. REFERENCES

[1] Swan II, J.E. and Gabbard, J. L. 2005. Survey of User-Based Experimentation in Augmented Reality. In Proc. of 1st Int'l Conf. on Virtual Reality. HCI International 2005, 2005.

[2] Broll, W., Stoerring, M., and Mottram, C. 2003. The Augmented Round Table – a New Interface to Urban Planning and Architectural Design, In Proc. INTERACT'03, pp. 1103-1104.

[3] Nakashima, K., Machida, T., Kiyokawa, K., and Takemura, H., 2005. A 2D-3D Integrated Environment for Cooperative Work, In Proc. VRST'05, pp. 16-22.

[4] Bolt, R. A. 1980. "Put-That-There": Voice and Gesture at the Graphics Interface. In Proc. Annual Conf. on Computer Graphics and Interactive Techniques, 262-270.

[5] Cohen, P. R. and Sullivan, J. W. 1989. Synergistic User of Direct Manipulation and Natural Language. In Proc. CHI'89, 227-233.

[6] Oviatt, S., Coulson, R., and Lunsford, R. 2004. When Do We Interact Multimodally? Cognitive Load and Multimodal Communication Patterns. In Proc ICMI'04, 129-136.

[7] Hauptmann, A. G. 1989. Speech and gestures for graphic image manipulation. In Proc CHI'89, 241-245.

[8] Cohen, P. R., Johnston, M., McGee, D., and Oviatt, S. 1997. QuickSet: Multimodal Interaction for Distributed Applications. In Proc. Int'l Conf. on Multimedia, 31-40.

[9] Rauschert, I., Agrawal, P., Sharmar, R., Fuhrmann, S., Brewer, I., MacEachren, A., Wang, H., and Cai, G. 2002. Designing a Human-Centered, Multimodal GIS Interface to Support Emergency Management. In Proc. GIS'02, 119-124.

[10] Tse, E., Greenberg, S., and Shen, C. 2006. GSI DEMO: Multiuser Gesture/Speech Interaction over Digital Tables by Wrapping Single User Applications. In Proc. ICMI'06, 76-83.

[11] Olwal, A., Benko, H., and Feiner, S. 2003. SenseShapes: Using Statistical Geometry for Object Selection in a Multimodal Augmented Reality System. In Proc. ISMAR'03, 300-301.

[12] Weimer, D., and Genapathy, S.K. 1989. A Synthetic Visual Environment with Hand Gesturing and Voice Input. In Proc. CHI'89, 235-240.

[13] Koons, D. B., and Sparrell, C. J. 1994. ICONIC: Speech and Depictive Gestures at the Human-Machine Interface. In Proc. CHI'94, 453-454.

[14] 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.

[15] Irawati, S., Green, S., Billinghurst, M., Duenser, A., and Ko, H. 2006. "Move the Couch Where?": Developing an Augmented Reality Multimodal Interface". In Proc. ISMAR'06, 183 – 186.

[16] Billinghurst, M., Kato, H., Poupyrev, I., Imamoto, K., and Tachibana, K. 2000. Virtual Object Manipulation on a Table-Top AR Environment. In Proc. ISAR'00, 111-119.

[17] Irawati, S., Green, S., Billinghurst, M., Duenser, A., and Ko, H. 2006. An Evaluation of an Augmented Reality Multimodal Interface Using Speech and Paddle Gestures. In Proc. ICAT'06, 272-283.

[18] 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.

[19] Kölsch, M., Turk, M., and Tobias, H. 2004. Vision-Based Interfaces for Mobility. In Proc. MobiQuitous'04, 86- 94.

[20] Salber, D., and Coutaz, J. 1993. Applying the Wizard of Oz Technique to the Study of Multimodal Systems. In Proc. HCI'93, 219-230.

[21] Coutaz, J., Salber, D., Carraux, E., and Portolan, N. 1996. NEIMO, a multiworkstation usability lab for observing and analyzing multimodal interaction. In Proc. Conf. Companion on Human Factors in Computing Systems, 402-403.

[22] 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. In Proc. Int'l Conf. on Spoken Language Processing, 2, 1351-1354.

[23] 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.

[24] Corradini, A. and Cohen, P. R. 2002. On the Relationships among Speech, Gestures, and Object Manipulation in Virtual Environments: Initial Evidence. In Proc. Int'l CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems, 52-61.

[25] Molin, L. 2004. Wizard-of-Oz Prototyping for Cooperative Interaction Design of Graphical User Interfaces. In Proc. NordiCHI'04, 425-428.

[26] OpenCV Library. 2008. http://sourceforge.net/projects/opencvlibrary/

[27] Freeman, H. 1974. Computer processing of line-drawing images. Computing Surveys, 6, 157-97.

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

[29] Looser, J., Grasset, R., Seichter, H., and Billinghurst, M. 2006. OSGART - A Pragmatic Approach to MR. In Proc. Industrial AR Workshop, ISMAR'06.

[30] Billinghurst, M., Campbell, S., Chinthammit, W., Hendrickson, D., Poupyrev, I., Takahashi, K., and Kato, H. 2000. Magic book: Exploring transitions in collaborative AR interfaces. Emerging Technologies Proposal, SIGGRAPH'00.

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