User-Defined Gestures for Augmented Reality

Thammathip Piumsomboon
HITLab NZ, University of Canterbury
thammathip.piumsomboon
@pg.canterbury.ac.nz

Adrian Clark
HITLab NZ, University of Canterbury
adrian.clark@canterbury.ac.nz

Mark Billinghurst
HITLab NZ, University of Canterbury
mark.billinghurst@canterbury.ac.nz

CHI'13, April 27 – May 2, 2013, Paris, France.
ACM 978-1-XXXX-XXXX-X/XX/XX.

Abstract
Recent developments in Augmented Reality (AR) have utilized hand gestures for interaction. However, little is known about user's preference and behavior gesturing in AR. In this paper, we present the results of a guessability study for hand gestures in AR. A total of 800 gestures have been elicited for 40 selected tasks from 20 participants. Using the agreement found among gestures, a user-defined gesture set has been created to guide designers to achieve consistent user-centered gestures in AR.

Author Keywords
Augmented reality; gestures; guessability.

ACM Classification Keywords
H.5.2. User Interfaces: Interaction styles, style guides, user-centered design.

General Terms
Design

Introduction
By overlaying virtual content onto the real world, Augmented Reality (AR) technology [1] allows users to perform tasks in the real and virtual environment at the same time. Natural hand is a promising medium that bridges the interaction in both worlds. Our research

Category		Tasks
Transforms	*Move*	1. Short distance
		2. Long distance
	Rotate	3. Roll (X-axis)
		4. Pitch (Y-axis)
		5. Yaw (Z-axis)
	Scale	6. Uniform scale
		7. X-axis
		8. Y-axis
		9. Z-axis
Simulation		10. Play/Resume
		11. Pause
		12. Stop/Reset
		13. Increase speed
		14. Decrease
Browsing		15. Previous
		16. Next
Selection		17. Single selection
		18. Multiple selection
		19. Box selection
		20. Select all
Editing		21. Insert
		22. Delete
		23. Undo
		24. Redo
		25. Group
		26. Ungroup
		27. Accept
		28. Reject
		29. Copy
		30. Cut
		31. Paste
Menu	*Horizontal (HM)*	32. Open
		33. Close
		34. Select
	Vertical (VM)	35. Open
		36. Close
		37. Select
	Object-centric (OM)	38. Open
		39. Close
		40. Select

Table 1. The list of forty AR tasks under six categories.

explored how natural hand input can be used for AR interaction. Past AR researchers have demonstrated the use of hand input, but shortcomings still exist. For example, in the studies of multimodal AR interfaces, hand gestures have primarily been studied as an add-on to speech [4, 5]. In cases of unimodal gesture interfaces, only a limited number of gestures have been used and the gestures were designed by researchers for optimal recognition rather than for naturalness, which meant that they could be arbitrary and unintuitive [4]. Recent research has integrated hand tracking with physics engines to provide realistic interaction with virtual content [3], but this system has limited support for gesture recognition and doesn't take into account the wide range of expressive hand gestures that could potentially be used for input commands.

To develop truly natural gesture based interfaces for AR applications, there are a number of unanswered questions that must be addressed. For example, for a given task is there a suitable and easy to perform gesture? is there a common set of gestures among users that would eliminate the need for arbitrary mapping of commands by designers? Similar shortcomings were also encountered in surface computing and motion gestures where Wobbrock et al. [10] and Ruiz et al. [7] addressed an absence of design insight in each area, by conducting guessability studies [9].

In this study, we focus explicitly on hand gestures for unimodal input in AR. We follow Wobbrock's approach and employ a guessability method, first showing a 3D animation of the task and then asking participants for their preferred gesture to perform the task. The result is the first comprehensive set of user-defined gestures for a range of different selected tasks in AR.

Previous Elicitation Studies

Past studies that elicited input from users had been described by [10]. The technique has been applied in a variety of research area such as unistroke gestures [9], surface computing [10] and motion gesture for mobile interaction [7]. In AR, only one Wizard of Oz study [5] has ever been conducted. This was to capture the type of speech and gesture input that users would like to use. In an object manipulation task it was found that the majority of gestures used was hand pointing due to reliance on speech for command inputs.

Developing a User-defined Gesture Set

To elicit user-defined gestures, we first presented the effects of the tasks being carried out by showing 3D animations in AR view and then requested the twenty participants for the gestures. Participants were asked to follow a think-aloud protocol while designing and also provided the rating for goodness and ease for each gesture. Participants were informed to ignore the issue of recognition for freedom in design and allowed us to observe their unrevised behavior.

Task Selection

From a broad range of applications in AR [8], we intended to keep the selected tasks generic and applicable across various applications. We had surveyed for common operations based on the previous researches e.g. [3-5]. This resulted in forty tasks which were grouped into six categories such that identical gestures could be used across these categories as shown in Table 1.

Participants

Twenty students were recruited for the study, comprising of twelve males and eight females, ranging in age

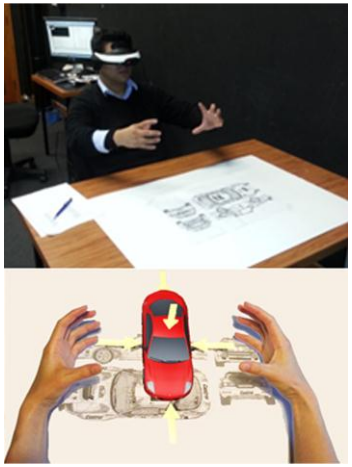

Figure 1. (Top) A participant was performing a gesture in front of the image marker. (Bottom) Showing the AR view from the HMD while the user gestured for a uniform scale task on an animation showing a shrinking car.

from 18 to 38 years with mean of 26 (SD =5.23). Participants had minimal knowledge of AR in order to avoid the influence of previous experience in designing input gestures. All participants used PC regularly with average usage in a day of 7.25 hours (SD = 4.0). Fifteen owned devices with touchscreen with average usage of 3.6 hours (SD = 4.17) a day. Eleven had experienced with gesture-in-the-air interface such as Nintendo's Wii or Microsoft's Kinect.

Apparatus

The interaction space was setup on and above a table's surface of size 120 x 80 cm where an image-based marker was placed in the middle. Each participant was seated in front of the marker and used Sony HMZ-T1 head mounted display (HMD) as the display device at 1280 x 720 resolutions. A high definition (HD) Logitech c525 web camera, was mounted in front of the HMZ-T1 as a viewing camera, providing a video stream at the display resolution. The combination of these HMD and HD camera offered a wider field of view at 16:9 aspect ratios, providing excellent coverage of the interaction space and complete sight of both hands while gesturing so as to improve user experience.

The Asus Xtion Pro Live depth sensor was placed 100 cm above the tabletop facing down onto the surface to provide reconstruction and occlusion between the user's hands and virtual contents. Another RGB camera was placed in front of the user for recording the frontal view of the users' gestures. The simulation PC was also used for monitoring and recording video and audio stream from the user's viewpoint so that the researchers can see what the user see. The OPIRA natural feature registration library [2] was used for registration and tracking of the marker. The method for providing 3D graphics animation and occlusion was described by [6].

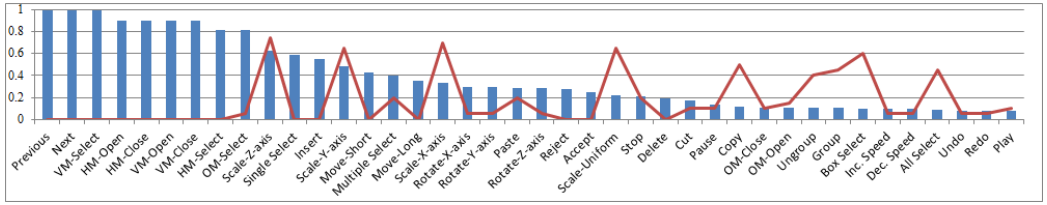

Figure 2. Agreement scores for forty tasks in descending order (blue bar) and ratio of two-handed gestures elicited in each task (red line).

Procedure

After a brief introductory to AR, the researcher described the experiment in details and showed the list of tasks to the participant. The forty tasks were divided into six categories and the participant could choose to carry out each category in any order, providing that there was no conflict between gestures within the same category. For each task, the 3D animation showing the effect of the task was displayed. For example, in the "Move – long distance" task, the participants would see a virtual toy block moves across the table from one location to another. Within the same category, the participant could view each task as many times as she/he needed. Once the participant understood the function of the task, she/he was asked to design the gesture that best suited for the task in a think-aloud manner. Participants are free to perform one or two-handed gestures as they seem fit for the task (See Figure 1).

Once the participant designed a consistent set of gestures designed for all tasks within the same category, they were asked to perform each gesture three times. After performing each gesture, they were asked to rate the gesture on a 7-point Likert scale in term of good-

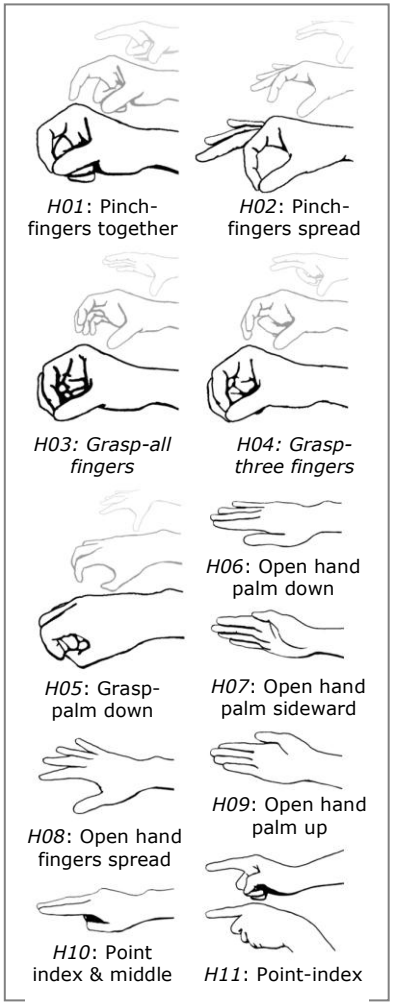

Figure 3. Variants of hand poses observed among gestures where the codes, H01-H11, were assigned for ease of reference.

ness and ease of use and a final interview was conducted. Each session took approximately one to one and a half hour to complete.

Result

A total of 800 gestures were generated from 20 participants performing the 40 tasks. The collected data includes the video recording from the front facing camera toward the user and the user's viewpoint camera, the subjective ratings for each gesture, and lastly, transcripts taken from the think-aloud protocol and the interview.

A User-defined Gesture Set

As demonstrated in the prior works by Wobbrock et al. [10] and Ruiz et al. [7], the user defined gesture set known as the consensus set was constructed based on the largest groups of identical gestures that were performed for the given task. For this study, we discovered that minor variation of hand poses existed and the fact that some groups within the same task were closely scored. Therefore, we had to loosen the constraints from "gestures must be identical within each group" to "gestures must be similar within each group" and multiple groups with the top scores could be selected in each task to improve guessability [9].

By similar, we meant that the gestures were identical or only differed by the variants of hand poses used with consistent directionality. For example, in the *previous* and *next* tasks, participants might have used an open hand, an index finger or two fingers to swipe from left to right or vice versa to perform these two tasks. These gestures were variants in term of the hand pose but the relevant characteristic in these tasks was the swiping direction that distinguished the task from *previous* and *next*. Therefore they were considered members from the same group. From 800 gestures, we had clustered similar gestures into 320 groups. Only 44 highly scored groups were included in the user-defined set, this is the consensus set, while 276 groups with low scores were discarded, this is referred to as the discarded set. The selected gestures of the consensus set made up of 495 gestures or 61.89% of all gestures collected, which comprises of gestures from six categories in the following percentage *transform* (19.38%), *menu* (17.75%), *editing* (11.75%), *browsing* (5.00%), *selection* (4.63%), and *simulation* (3.38%).

Level of Agreement

To compute the degree of consensus among gestures designed by twenty participants, we calculate an agreement score A from the following equation by [9], which has also been applied in prior guessability studies [7, 10].

$$A = \sum_{P_s} \left(\frac{|P_s|}{|P_t|} \right)^2$$

Where P_t is the total number of gestures within the task, *t*. P_s is a subset of P_t containing similar gestures and the range of *A* is [0, 1]. Consider the *rotate-pitch (y-axis)* task that contained five groups of 8, 6, 4, 1, and 1. The calculation for A_{pitch} is as follows:

$$A_{pitch} = \left(\frac{|8|}{|20|} \right)^2 + \left(\frac{|6|}{|20|} \right)^2 + \left(\frac{|4|}{|20|} \right)^2 + \left(\frac{|1|}{|20|} \right)^2 + \left(\frac{|1|}{|20|} \right)^2 = 0.295$$

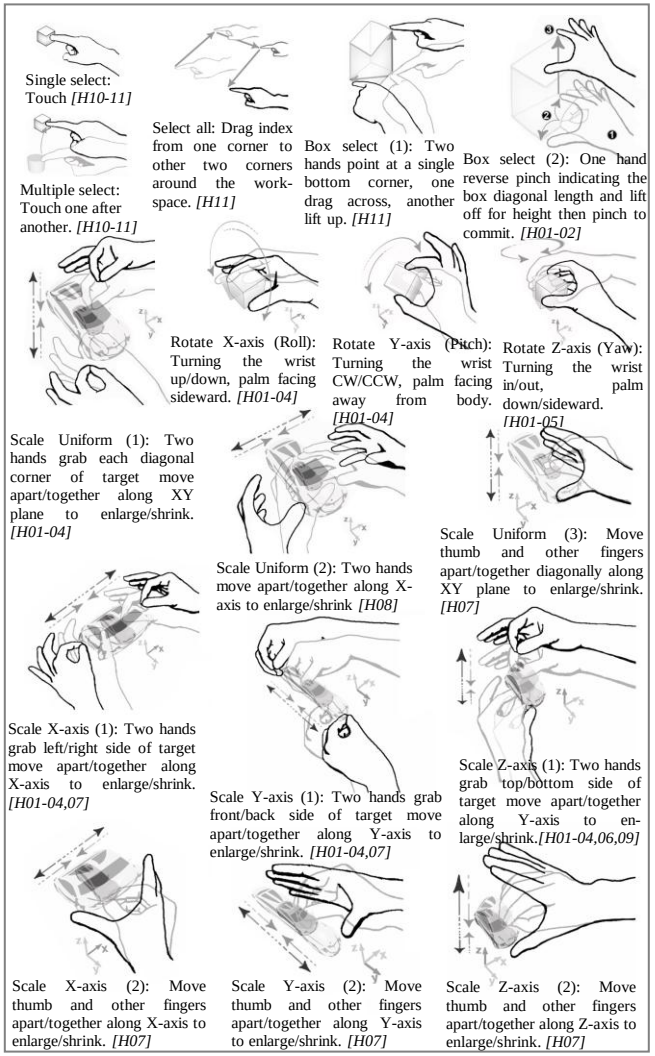

Figure 4. The user-defined gesture set for AR Part 1. The number shows in the parenthesis indicates multiple gestures in the same task. The codes in the square bracket indicate the hand pose variants (See Figure 3) for the same gesture.

The agreement scores for all forty tasks are shown in Figure 2. Despite the low agreement scores for *all select*, *undo*, *redo* and *play* tasks, there were notable groups of gestures that stood out with more votes over the other as well as recurring gestures patterns observed among various groups, which yielded us the user defined gesture set for all forty tasks.

User-defined Gesture Set and Its Characteristics

There were 40 tasks but a total of 44 selected groups of gestures. The greater number of gestures per task was due to aliasing where multiple gestures were allowed to map to a single task as well as multiple tasks across different categories could be mapped to a single gesture, which could improve guessability [9]. One task was assigned with 3 gestures (*uniform-scaling*), 7 tasks with 2 gestures (*x, y, z scaling, box select, stop, delete, and copy*), 23 tasks with 1 gesture. Since duplicate gestures were allowed across categories of tasks with certain exceptions that are explained in the following paragraphs, 2 gestures were assigned for 4 tasks (*short, long move, insert, and paste*), 1 gesture for 3 tasks (*play, increase speed, and redo*), and another for 2 tasks (*decrease speed and undo*). We had classified the major variants of observed hand poses into 11 poses with the codes, *H01* to *H11*, as illustrated in Figure 3. For tasks where these variants existed, multiple poses could be used interchangeably as indicated by the description under each user-defined gesture's illustration (See Figure 4 and 5).

There was only one conflict between gestures within the same category. This was between pause and stop where the gesture of an open-hand facing away from the body was proposed for both with the votes of 4 and 7, respectively. In this circumstance, the task with greater number of votes got assigned the gesture and so *stop* won over *pause*. The two gestures from four tasks, *short* and *long move*, *insert,* and *paste* were identical and therefore shared. *Play*, *increase speed* and *redo* shared one gesture while *decrease speed* and *undo* also shared another.

Play and *increase speed* as well as *insert* and *paste* were the exceptions where a single gesture was assigned to two tasks within the same category. For *play* and *increase speed*, the reason was due to the participants intention to use the number of spin cycles by the index finger to indicate the speed of the simulation e.g. a single clockwise spin to indicate *play*, two clockwise spin to indicate *twice* the speed and three spins for *quadruple* speed. For *insert* and *paste*, the participants found the two tasks serving a similar purpose where *insert* allowed a user to select the object from menu and placed it in the scene, whereas, user could imagine selecting the target object from a clipboard and also placed in the scene. Therefore as long as unique selection spaces were provided for an insert menu and a clipboard for paste, no conflict would occur. Thus reus-

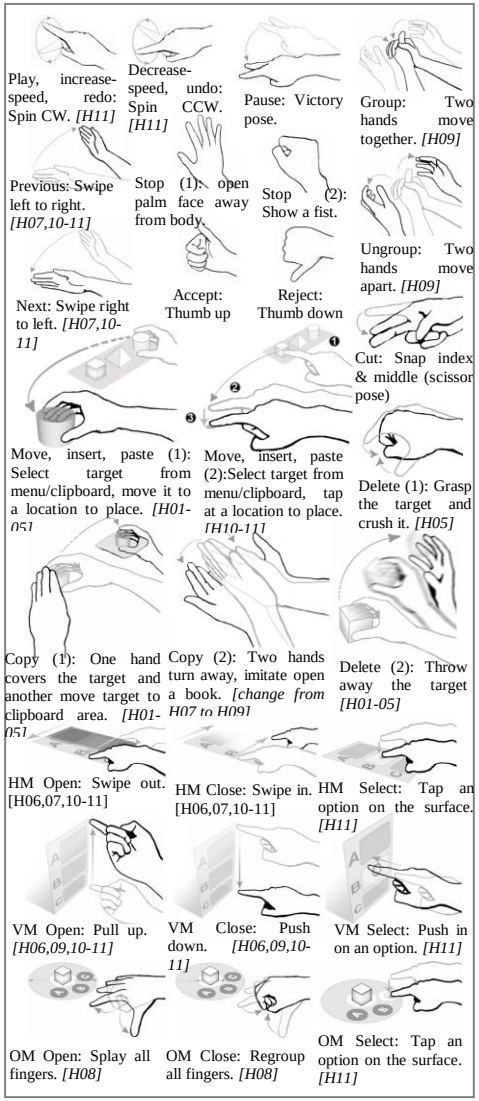

Figure 5. The user-defined gesture set for AR Part 2.

ing the same gestures were only natural and did not cause conflict.

The result is a consistent set of user-defined gestures that contains 44 gestures, where 34 gestures are unimanual and 10 are bimanual. The complete gesture set is illustrated in Figure 4 and 5.

The Subjective Rating on Goodness and Ease

By comparing the subjective rating for goodness and ease between the consensus set and the discarded set, we found that the average scores on gestures being good match for their tasks were 6.02 (SD=1.00) and 5.50 (SD=1.22) and the average scores for ease of performance were 6.17 (SD = 1.03) and 5.83 (SD=1.21), respectively. The user-defined set were rated significantly higher than the discarded set in both ratings of goodness ($F_{1, 798}$ = 43.896, p<.0001) and ease ($F_{1, 798}$=18.132, p<.0001). Hence, we could conclude that on average, gestures in the user-defined set were better than the discarded one in term of goodness and ease.

Discussion and Conclusions

We have presented the results of a guessability study for hand gestures in AR. Using the agreement found among the elicited gestures, 44 user-defined gestures have been selected. Although the gestures have been found for all 40 tasks but the agreement score varied where lower score indicates less confidence in the gesture selected. This requires a further study to validate our gestures. In the follow up experiment, another group of participants will be shown the elicited gestures from both consensus and discarded sets and determine their preference for each task to confirm our result.

References

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

[2] Clark, A. J., Green, R. D. and Grant, R. N. Perspective correction for improved visual registration using natural features. In *Proc. IVCNZ 2008,* (2008), 1-6.

[3] Hilliges, O., Kim, D., Izadi, S., Weiss, M. and Wilson, A. HoloDesk: direct 3d interactions with a situated see-through display. In *Proc. CHI 2012,* ACM (2012), 2421-2430.

[4] Kolsch, M., Bane, R., Hollerer, T. and Turk, M. Multimodal interaction with a wearable augmented reality system. *IEEE Computer Graphics and Applications*, *26*, Compendex (2006), 62-71.

[5] Lee, M. *Multimodal Speech-Gesture Interaction with 3D Objects in Augmented Reality Environments*. University of Canterbury, Christchurch, New Zealand, 2010.

[6] Piumsomboon, T., Clark, A., Umakatsu, A. and Billinghurst, M. Poster: Physically-based natural hand and tangible AR interaction for face-to-face collaboration on a tabletop. In *Proc. 3DUI 2012, IEEE* (2012), 155-156.

[7] Ruiz, J., Li, Y. and Lank, E. User-defined motion gestures for mobile interaction. In *Proc. CHI 2011*, ACM (2011), 197-206.

[8] van Krevelen, D. W. F. and Poelman, R. A Survey of Augmented Reality Technologies, Applications and Limitations. *The International Journal of Virtual Reality*, *9*, 2 (2010), 1-20.

[9] Wobbrock, J. O., Aung, H. H., Rothrock, B. and Myers, B. A. Maximizing the guessability of symbolic input. In *Proc. CHI 2005 extended abstracts*, ACM (2005), 1869-1872.

[10] Wobbrock, J. O., Morris, M. R. and Wilson, A. D. User-defined gestures for surface computing. In *Proc. CHI 2009*, ACM (2009), 1083-1092.