User-Defined Gestures for Augmented Reality

Thammathip Piumsomboon, Adrian Clark, Mark Billinghurst, Andy Cockburn

User-Defined Gestures for Augmented Reality

Thammathip Piumsomboon[1,2], Adrian Clark[1], Mark Billinghurst[1]
and Andy Cockburn[2]

[1]HIT Lab NZ, University of Canterbury, Christchurch, New Zealand
[2]Department of Computer Science and Software Engineering, University of Canterbury, Christchurch, New Zealand

Thammathip.Piumsomboon@pg.Canterbury.ac.nz, {Adrian.Clark, Mark.Billinghurst, Andy}@Canterbury.ac.nz

Abstract. Recently there has been an increase in research towards using hand gestures for interaction in the field of Augmented Reality (AR). These works have primarily focused on researcher designed gestures, while little is known about user preference and behavior for gestures in AR. In this paper, we present our guessability study for hand gestures in AR in which 800 gestures were elicited for 40 selected tasks from 20 participants. Using the agreement found among gestures, a user-defined gesture set was created to guide designers to achieve consistent user-centered gestures in AR. Wobbrock's surface taxonomy has been extended to cover dimensionalities in AR and with it, characteristics of collected gestures have been derived. Common motifs which arose from the empirical findings were applied to obtain a better understanding of users' thought and behavior. This work aims to lead to consistent user-centered designed gestures in AR.

Keywords: Augmented reality; gestures; guessability

1 Introduction

By overlaying virtual content onto the real world, Augmented Reality (AR) allows users to perform tasks in the real and virtual environment at the same time [1]. Natural hand gestures provide an intuitive interaction method which bridges both worlds. While prior research has demonstrated the use of hand gesture input in AR, there is no consensus on how this combination of technologies can best serve users. In studies involving multimodal AR interfaces, hand gestures were primarily implemented as an add-on to speech input [2] [3]. 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, meaning that they were often arbitrary and unintuitive [4] [5] [6]. Recent research has integrated hand tracking with physics engines to provide realistic interaction with virtual content [7] [8], but this provides limited support for gesture recognition and does not 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? Is there a taxonomy that can be used to classify gestures in AR? Similar shortcomings were encountered in the fields of surface computing and motion gestures, where Wobbrock et al. [9] and Ruiz et al. [10] addressed absences of design insight by conducting guessability studies [11].

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. Users were also asked to subjectively rate their chosen gestures, based on the perceived "goodness" and ease to perform. By analyzing the results of this study, we were able to create a comprehensive set of user-defined gestures for a range of tasks performed in AR.

This work makes a number of contributions; (1) The first set of user-defined gestures captured from an AR interface, (2) Classification of these gestures based on a gesture taxonomy for AR which was extended from Wobbrock's surface gesture taxonomy [9], (3) Agreement scores of gestures for selected tasks and subjective rating of the gestures, (4) Qualitative findings from the design process, and (5) Discussion of the implications of this work for AR, gesture interfaces, and gesture recognition.

2 Related Work

The topic of hand gesture classification as based on human discourse was excellently covered by the work of Wobbrock et al. [9]. As our work extends this approach to gesture interaction in AR, we focus on related work in bare hand and glove-based unimodal hand gestures interfaces, multimodal interfaces coupled with speech, and recent advancements in AR relevant to interaction. In addition, we briefly discuss previous research that utilized elicitation techniques.

2.1 Hand Gesture Interfaces in AR

Lee et al. [12] designed gloves with conductive fabric on the fingertips and the palm for gesture recognition and vibration motors for haptic feedback. The gloves were tracked using markers placed around the wrist area. A small set of gestures were used to allow selection, gripping, cutting and copying actions.

Lee and Hollerer [5] created Handy AR, a system capable of bare hand interaction using a standard web camera. The supported gestures were limited to an opened/closed hand for object selection and hand rotation for object inspection. Their follow up work allowed objects to be relocated using markerless tracking [13].

Fernandes and Fernandez [6] trained statistical models with hand images to allow bare hand detection. Virtual objects could be translated using the hand in a palm upwards orientation, while rotation and scaling along the marker plane was achieved using two handed pointing.

The main shortcoming of all these interfaces was that they only recognize a small number of gestures, and this gesture set was designed by the researchers for easy recognition. No support was provided for users to define their own gestures which were more comfortable or had contextual meaning.

2.2 Hand Gesture and Speech Interfaces in AR

SenseShapes [14] aimed to find spatial correlation between gestures and deictic terms such as “that”, “here”, and “there” in an object selection task. The user’s hands were tracked using data gloves, and object selection was facilitated by a virtual cone projected out from the users’ fingers. The region of interest was estimated based on speech, gaze projection and the pointing projection.

Heidemann et al. [2] demonstrated an AR interface which identified objects on a tabletop. Skin color segmentation was used to identify the user’s index finger, allowing users to select virtual objects by pointing and make menu selections. Speech could also be used to issue information queries and interact with the 2D menu.

Kolsch et al. [3] created a mobile AR system that supported interaction by hand gesture, speech, trackball and head pose. Gesture recognition was implemented using HandVu, a computer vision-based gesture recognition library. They categorized tasks by dimensionality. For example taking a snapshot was defined as 0D, adjusting the focus region depth was 1D, using a pencil tool was 2D, and orienting virtual objects was 3D. Some actions such as relocating/resizing/orienting could be performed multimodally by speech, gesture or trackball, while other actions such as take/save/discard snapshot could only be performed by speech.

The work most closely related to ours was that of, Lee [15], who conducted a Wizard of Oz study of an AR multimodal interface to measure the types of gestures people would like to use in a virtual object manipulation task. In this study pointing, translation and rotation gestures were captured. She later developed a multimodal gesture and speech interface for a design related task, however speech was used as a primary input as in typical multimodal systems therefore gestures were only mapped to limited number of spatial related tasks for example pointing, grabbing and moving.

2.3 Recent Advancements in AR Technology

Advancements of markerless tracking algorithms and consumer hardware have enabled greater possibilities for gesture based AR interfaces. Modern vision-based tracking algorithms can robustly register the environment without markers, allowing for higher mobility [16]. Furthermore, an introduction of consumer depth sensors such as the Microsoft Kinect has made real-time 3D processing accessible and have introduced a new interaction paradigm in AR through real-time physically-based natural interaction [7, 8].

2.4 Previous Elicitation Studies

Wobbrock et al. describe prior studies involving elicitation of input from users [9]. The technique is common in participatory design [17] and has been applied in a variety of research areas such as unistroke gestures [11], surface computing [9] and motion gesture for mobile interaction [10]. In AR, a Wizard of Oz study [15] for gestures and speech was conducted and aimed 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 hand pointing due to reliance on speech for command inputs. In this research, our focus is to explore the potential of hand gestures as the unimodal input.

3 Developing a User-defined Gesture Set

To elicit user-defined gestures, we first presented the effect of the task being carried out by showing a 3D animation in AR, and then asked the participants to describe the gestures they would use. Participants designed and performed gestures for forty tasks across six categories, which included gestures for three types of menu. Participants were asked to follow a think-aloud protocol while designing the gestures, and also to rate the gestures for goodness and ease to perform. They were asked to ignore the issue of recognition difficulty to allow freedom during the design process, and to allow us to observe their unrestricted behavior. At the end of the experiment, brief interviews were conducted and preferences of the three types of proposed gesture menus were collected.

3.1 Task Selections

In order for the gesture set to be applicable across a broad range of AR applications [18], we surveyed common operations in previous research e.g. [3, 5, 6, 12, 19], which resulted in forty tasks that included three types of gesture menu, which are horizontal [20], vertical [19], and object-centric that we proposed. These tasks were then grouped into six categories based on context, such that identical gestures could be used across these categories, as shown in Table 1.

3.2 Participants

Twenty participants were recruited for the study, comprising of twelve males and eight females, ranging in age from 18 to 38 with mean of 26 ($\sigma = 5.23$). The participants which were selected had minimal knowledge of AR to avoid the influence of previous experience with gestures interaction. Nineteen of the participants were right handed, and one was left handed. All participants used PCs regularly, with an average daily usage of 7.25 hours ($\sigma = 4.0$). Fifteen owned touchscreen devices, with an average daily usage of 3.6 hours ($\sigma = 4.17$). Eleven had experience with gesture-in-the-air interfaces such as those used by the Nintendo Wii or Microsoft Kinect gaming devices.

3.3 Apparatus

The experimental interaction space, shown in Figure 1 (Left), was the area on and above a 120 x 80cm table. Each participant was seated in front of the table, and a Sony HMZ-T1 head mounted display (HMD) at 1280 x 720 resolutions was used as the display device. A high definition (HD) Logitech c525 web camera was mounted on the front of the HMZ-T1 as a viewing camera, providing a video stream at the display resolution. This HMD and camera combination offered a wide field of view, with a 16:9 aspect ratio, providing a good view of the interaction space and complete sight of both hands while gesturing.

An 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 content. An RGB camera was placed in front of and facing the user to record the users' gestures. A PC was used for the AR simulation and to record the video and audio stream from the user's perspective. A planar image marker was placed in the center of the table, and the OPIRA natural feature registration library [21] was used for registration and tracking of this marker. The 3D graphics, animation and occlusion were handled as described by Piumsomboon et al. [22].

Table 1. The list of forty tasks in six categories.

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

3.4 Procedure

After an introduction to AR and description of how to operate the interface, the researcher described the experiment in detail and showed the list of tasks to the participant. The forty tasks were divided into six categories, as shown in Table 1, and the participant was told they could choose to carry out the categories in any order, providing that there was no conflict between gestures within the same category. For each task, a 3D animation showing the effect of the task was displayed, for example, in the "Move – long distance" task, participants would see a virtual toy block move across the table. Within the same task 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 they felt best suited the task in a think-aloud manner. Participants were free to perform one or two-handed gestures as they saw fit for the task (See Figure 1, Right).

Once the participant had designed a consistent set of gestures 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 goodness and ease of use. At the end of the experiment, a final interview was conducted, where participants were asked to rank the three types of menu presented (horizontal, vertical, and object-centric as shown in Figure 5) in terms of preference and the justification for their ranking. Each session took approximately one to one and a half hours to complete.

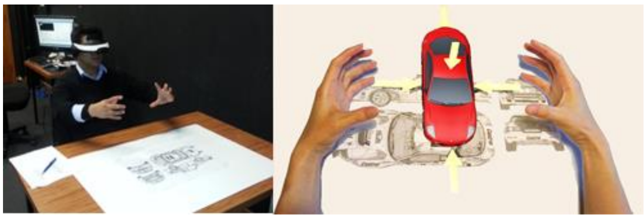


Fig. 1. (Left) A participant performs a gesture in front of the image marker. (Right) The participant sees an AR animation of a shrinking car, and performs their gesture for a uniform scale task

4 Result

A total of 800 gestures were generated from the 20 participants performing the 40 tasks. The data collected for each user included video and audio recorded from the camera facing towards the user and the user's viewpoint camera, the user's subjective rating for each gesture, and transcripts taken from the think-aloud protocol and interview.

4.1 Taxonomy of Gestures in AR

We adapted Wobbrock's surface taxonomy [9] to better cover the AR gesture design space by taking their four-dimensional taxonomy, (*form*, *nature*, *binding*, and *flow)* and extending it with two more dimensions; *symmetry* and *locale*. Each dimension is comprised of multiple categories, as shown in Table 2.

The scope of the *form* dimension was kept unimanual, and in the case of a two-handed gesture, applied separately to each hand. In Wobbrock's original taxonomy, *form* contained six categories including *one-point touch* and *one-point path*, however, these two categories were discarded as they were not relevant to AR gestures that occur in three dimensional space.

The *nature* dimension was divided into *symbolic*, *physical*, *metaphorical* and *abstract* categories. Examples of symbolic gestures are thumbs-up and thumbs-down for *accept* and *reject*. *Physical* gestures were classified as those that would act physically on the virtual object as if it was a real object for instance grabbing a virtual block and relocating it for a *move* task. *Metaphorical* gestures express actions through existing metaphor e.g. pointing an index finger forward and spinning it clockwise to indicate *play* or *increase speed* as if one was playing a roll film. Any arbitrary gestures were considered *abstract*, such as a double-tap on the surface to *deselect* all objects.

The *binding* dimension considered relative location where gestures were performed. The *object-centric* category covered *transform* tasks such as *move*, *rotate*, and *scale*, as these are defined with respect to the objects being manipulated. Opening and closing horizontal or vertical menus were classified in the *world-dependent* category as they are located relative to the physical workspace. Gestures in the *World-independent* category could be performed anywhere, regardless of the relative position to the world, such as an open hand facing away from one's body to indicate *stop* during a simulation. Gestures performed across multiple spaces, such as *insert* where selection is *object-centric* and placement is *world-dependent,* fell into the *mixed dependencies* category.

In the *flow* dimension, gestures were categorized as *discrete* when the action is taken only when the gesture is completed, for example an index finger must be spun clockwise in a full circle to perform the *play* command. The gestures were considered *continuous* if the simulation must respond during the operation, such as manipulating an object using the *transform* gestures.

The first additional dimension we developed, *symmetry*, allowed classification of gestures depending on whether they were one handed (*unimanual*) or two handed (*bimanual*). The *unimanual* category was further split into *dominant* and *nondominant,* as some participants preferred to use their nondominant hand to perform gestures that required little or no movement, leaving their dominant hand for gestures requiring finer motor control. An example of this would be to use the dominant hand to execute a *selection,* and then use the non-dominant hand to perform a scissor pose for a *cut* operation. The *bimanual* category also subdivided, *symmetric* gestures representing two-handed gestures where both hands executed the same *form*, such as scaling, where both hands perform a pinch moving toward or away from each other. Two handed gestures, where the *form*s of the hands are different, fall into the *asymmetric* *bimanual* category. An example of this is the *copy (1)* gesture where one hand is used to select the target object (*static pose*) while the other hand drags the copy into place (*static pose and path*).

The other dimension we introduce is *locale*. If a gesture required physical contact with the real surface, they are considered *on-the-surface* as opposed to *in-the-air*. Gestures that require both are considered *mixed locales*. For example, an index finger tapped *on-the-surface* at a virtual button projected on the tabletop to perform horizontal menu selection task, as opposed to an index finger pushed *in-the-air* at a floating button to execute vertical menu selection. An example of a *mixed locales* gesture is, the *box selection (1)*, where one hand indicated the area of the bottom surface of the box by dragging an index finger diagonally along the table's surface, while another hand lifted off the surface into the air to indicate the height of the box (See Figure 5).

Table 2. Taxonomy of gestures in AR extended from taxonomy of surface gestures.

Taxonomy of Gestures in AR		
Form	static pose	Hand pose is held in one location.
	dynamic pose	Hand pose changes in one location.
	static pose and path	Hand pose is held as hand relocates.
	dynamic pose and path	Hand pose changes as hand relocates.
Nature	Symbolic	Gesture visually depicts a symbol.
	physical	Gesture acts physically on objects.
	metaphorical	Gesture is metaphorical.
	abstract	Gesture mapping is arbitrary.
Binding	object-centric	Gesturing space is relative to the object.
	world-dependent	Gesturing space is relative to the physical world.
	world-independent	Gesture anywhere regardless of position in the world.
	mixed dependencies	Gesture involves multiple spaces.
Flow	Discrete	Response occurs after the gesture completion.
	continuous	Response occurs during the gesture.
Symmetry	dominant unimanual	Gesture performed by dominant hand.
	nondominant unimanual	Gesture performed by nondominant hand.
	symmetric bimanual	Gesture using both hands with the same form.
	asymmetric bimanual	Gesture using both hands with different form.
Locale	on-the-surface	Gesture involves a contact with real physical surface.
	in-the-air	Gesture occurs in the air with no physical contact.
	mixed locales	Gesture involves both locales.

4.2 Findings from Classification

Classification was performed on the 800 gestures as shown in Figure. 2. Within the six dimensional taxonomy, the most common characteristics of gestures were *static pose and path*, *physical*, *object-centric*, *discrete*, *dominant unimanual*, and *in-the-air*.

Within the *form* dimension, there was a slightly higher number of *static poses* (3%) performed with a *non-dominant* hand and lower for *static poses with path* gestures

(2.5%) over a *dominant* hand. This slight discrepancy was contributed by some participants preferring to use their *dominant* hand for gestures with movement while using a *non-dominant* for a static pose.

In the *nature* dimension, overall the gestures were dominantly *physical* (39%) and *metaphorical* (34.5%). The gestures chosen to perform *transform*, *selection*, and *menu* tasks were predominantly *physical*, with the percentage of 76.1%, 50%, and 57.8% respectively. The *browsing* and *editing* task gestures were mainly *metaphorical* (100% and 40.9% respectively), while the *simulation* task, gestures were split across *symbolic* (37%), *metaphorical* (34%), and *abstract* (29%) categories. For the *binding* dimension, the majority of gestures for the *transform* and *selection* tasks were *object-centric* (100% and 75% respectively). *Simulation* (93%) and *browsing* (100%) task gestures were mainly *world-independent* (93% and 100%), while *editing* tasks gestures were *world-independent* (39.5%) and *object-centric* (32.3%). *Menu* tasks gestures were *object-centric* (50%) and *world-dependent* (45.6%).

For the remaining dimensions including *flow*, *symmetry*, and *locale*, the gestures chosen across all tasks were primarily *discrete* (77.5%), *dominant unimanual* (67.8%) and *in-the-air* (78%).

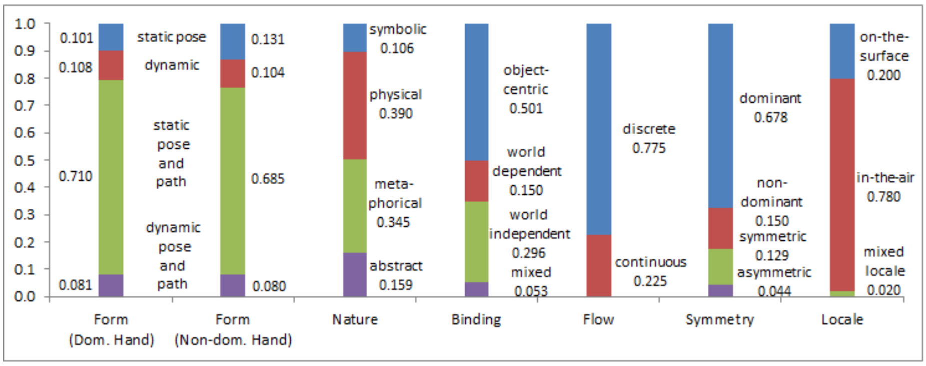


Fig. 2. The proportion of gestures in each category in the six dimensional taxonomy. *Form* has been calculated for each hand separately.

4.3 A User-defined Gesture Set

As defined in prior work by Wobbrock et al. [9] and Ruiz et al. [10], the user defined gesture set, known as the "consensus set", is constructed based on the largest groups of identical gestures that are performed for the given task. In our study, each gesture valued at one point; therefore there were 20 points within each task and a total of 800 points for all tasks.

We found that participants used minor variations of similar hand poses, for example a swiping gesture with the index finger or the same swipe with the index and middle fingers, and therefore chose to loosen the constraints from "gestures must be identical within each group" to "gestures must be similar within each group". We defined "similar gestures" as *static pose and path* gestures that were identical or having consistent directionality although the gesture had been performed with different *static* hand poses.

We had classified the major variants of observed hand poses into 11 poses with the codes, *H01* to *H11*, as illustrated in Figure 4. For tasks where these variants existed, the variant poses could be used interchangeably, as indicated by the description under each user-defined gesture's illustration (Figure 5).

Exercising the "similar gesture" constraint, we were able to reduce the original 800 gestures into 320 unique gestures. The top 44 highly scored gestures were selected to make the consensus set, while the remaining 276 lowest scored gestures were discarded, defined by Wobbrock et al. [9] as the discarded set. The selected gestures of the consensus set represented 495 (61.89%) of the 800 recorded gestures (495 of 800 points). The consensus set of gestures comprised the overall task gestures in the following percentage *transform* (19.38%), *menu* (17.75%), *editing* (11.75%), *browsing* (5.00%), *selection* (4.63%), and *simulation* (3.38%), which sum up to 61.89%.

Level of Agreement. To compute the degree of consensus among the designed gestures, an agreement score *A* was calculated using Equation 1 [11]:

$$A = \sum_{P_s} \left(\frac{|P_s|}{|P_t|} \right)^2 \qquad \textbf{(1)}$$

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 gestures with scores of 8, 6, 4, 1, and 1 points. The calculation for A_{pitch} is as follows:

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

The agreement scores for all forty tasks are shown in Figure 3. While there is low agreement in the gestures set for tasks such as *all select*, *undo*, *redo* and *play*, there were notable groups of gestures that stood out with higher scores.

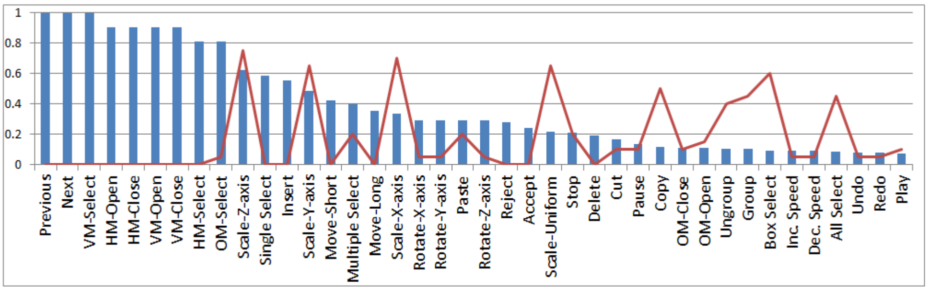


Fig. 3. Agreement scores for forty tasks in descending order (bars) and ratio of two-handed gestures elicited in each task (line).

User-defined Gesture Set and Its Characteristics. As mentioned in Section 3.4, we allowed users to assign the same gesture to different tasks as long as the tasks were not in the same category. In addition to this, there were some tasks where there were two or more gestures commonly assigned by the participants. This non one-to-one mapping resulted in a consensus set of 44 gestures for a total of 40 tasks, which resulted in improved guessability [11].

When mapping multiple gestures to a single task, there was one task which had three gestures assigned to it (*uniform-scaling*), seven tasks had two gestures (*x, y, z scaling, box select, stop, delete, and copy*), and 23 tasks only had gesture. On the contrary, for the nine remaining tasks, two gestures were assigned to four tasks (*short, long move, insert, and paste*), one gesture assigned to three tasks (*play, increase speed, and redo*), and one gesture assigned to two tasks (*decrease speed and undo*).

When creating the consensus set, we only found one conflict between gestures within the same category. This was between the *pause* and *stop* tasks, where the gesture of an open-hand facing away from the body was proposed for both with scores of 4 and 7 points respectively. To resolve this, we simply assigned the gesture to the task with the higher score, in this case *stop*.

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 with no conflict. For *play* and *increase speed*, the participants intention was to use the number of spin cycles of the index finger to indicate the speed of the simulation i.e. 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 felt the two tasks served a similar purpose; *insert* allowed a user to select the object from menu and placed it in the scene, whereas *paste* allowed a user to place an object from the clipboard into the scene. In the follow up interviews, participants suggested a simple resolution to this would be to provide unique selection spaces for the *insert* menu and *paste* clipboard.

With the minor ambiguities resolved, we were able to construct a consistent set of user-defined gestures which contained 44 gestures, where 34 gestures were unimanual and 10 were bimanual. The complete gesture set is illustrated in Figure 5.

The Subjective Rating on Goodness and Ease. After the participants had finished designing gestures for a task category, they were asked to subjectively rate their gestures for goodness and ease to perform on a 7-point Likert scale. By comparing these subjective ratings between the consensus set (user-defined set) and the discarded set, we found that the average score for gestures that users believed were a good match for the tasks was 6.02 ($\sigma = 1.00$) for the consensus set and 5.50 ($\sigma = 1.22$) for the discarded set, and the average score for the ease of performance was 6.17 ($\sigma = 1.03$) for the consensus set and 5.83 ($\sigma =1.21$) for the discarded set. The consensus set was rated significantly higher than the discarded set for both goodness ($F_{1,\ 798} = 43.896$, $p<.0001$) and ease of performance ($F_{1,\ 798}=18.132$, $p<.0001$). Hence, we could conclude that, on average, gestures in the user-defined set were better than those in the discarded set in terms of goodness and ease of performance.

4.4 Findings from the Design Process

Participants were asked to think-aloud when designing their gestures, and a follow-up interview was conducted after the experiment was complete. Analysis of the resulting empirical data showed recurring thought processes. We present seven motifs which describe the mutual design patterns encountered in designing gestures for AR, which we describe as *reversible and reusable*, *size does matter*, *influence from existing UI*, *the obvious and the obscure*, *feedback backfired*, *menu for AR, axes and boxes,* and *variation of hand poses*.

Reversible and Reusable. The consensus set included reversible and reusable gestures. We defined reversible gestures as those when performed in an opposite direction yielded opposite effects e.g. *rotation*, *scaling, increase/decrease speed* etc. We defined reusable gestures as those which were used commonly for tasks which were different, but participants felt had common attributes e.g. increase speed/ redo, decrease speed/undo and insert/paste. In the experiment there were several dichotomous tasks that are separate tasks which perform the exact opposite operation. Participants used reversible gestures for both tasks where the opposite effect was presented in the single animation, such as *rotation* and *scaling,* as well as tasks where the opposite effects were shown separately, such as *increase/decrease speed*, *previous/next*, *undo/redo*, *group/ungroup*, and *open/close menus*. All two-handed dichotomous tasks were *symmetric bimanual* with the gestures performed on both hands being the same *form*.

Size Does Matter. We found that the virtual object's size influenced the design decision of some participants, especially with regards to the number of hands that they would use to manipulate the object for example the majority of gestures performed for *scale* tasks were bimanual. This was due to scaling involving shrinking and enlarging the target object within and beyond the palm size. Some comments are as follows:

"Instinctively, I would use two hands to adapt to the size of the model but it's cool if I can use just the two fingers (one-handed) for something as large." – P04

"Depending on the size of the piece, I can use two hands when it's big but in the case of small piece, it's enough to use the two fingers (thumb and index)." – P12

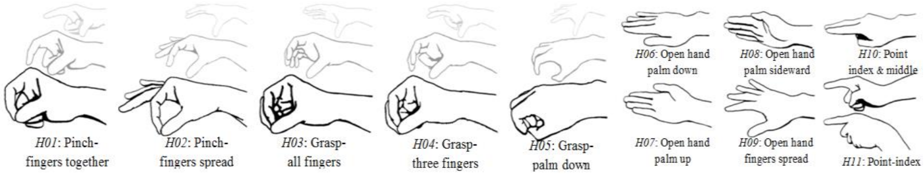
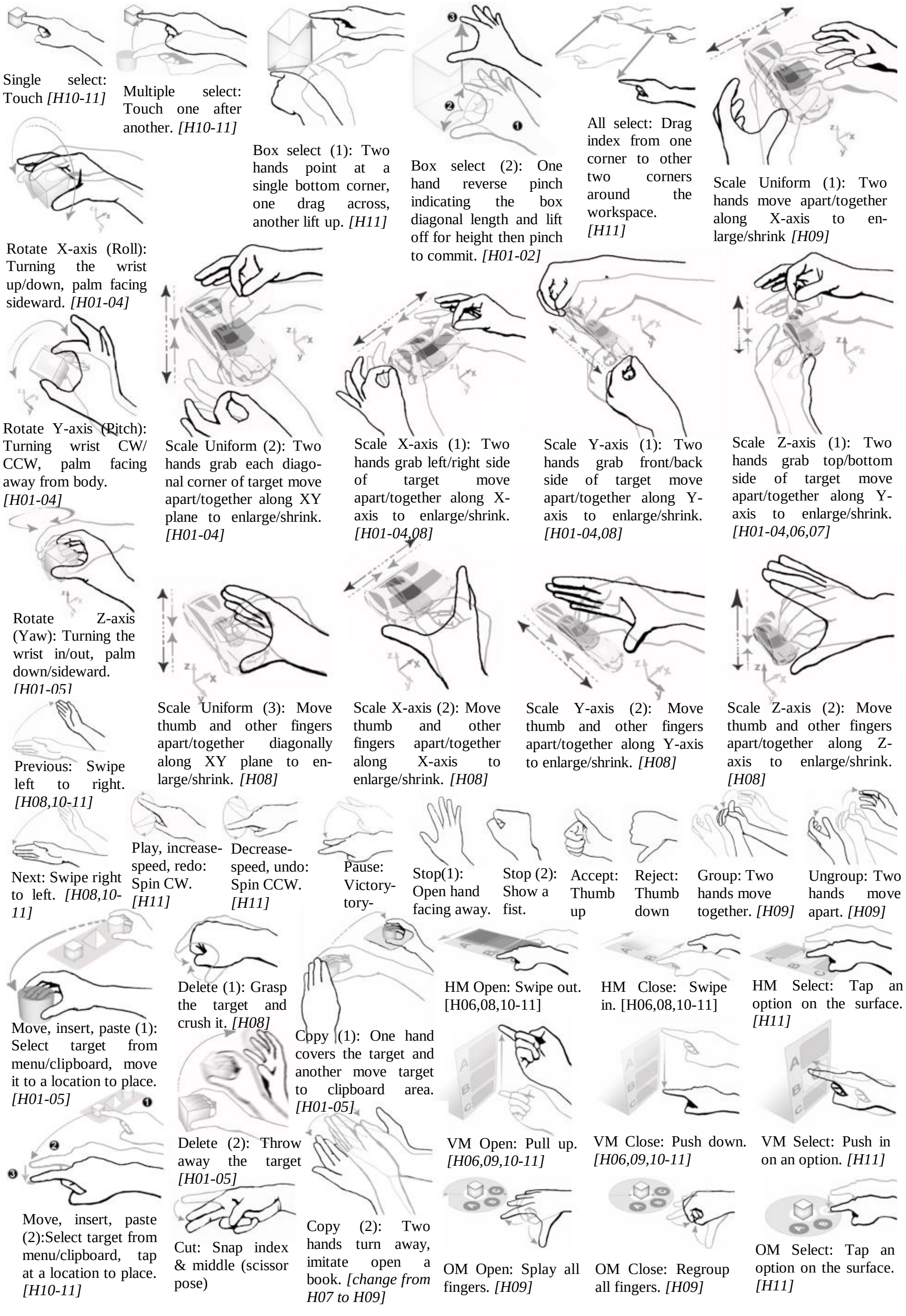

H01: Pinch-fingers together
H02: Pinch-fingers spread
H03: Grasp-all fingers
H04: Grasp-three fingers
H05: Grasp-palm down
H06: Open hand palm down
H07: Open hand palm up
H08: Open hand palm sideward
H09: Open hand fingers spread
H10: Point index & middle
H11: Point-index

Fig. 4. Variants of hand poses observed among gestures where the codes, H01-H11, were assigned for ease of reference.



Single select: Touch *[H10-11]*

Multiple select: Touch one after another. *[H10-11]*

Box select (1): Two hands point at a single bottom corner, one drag across, another lift up. *[H11]*

Box select (2): One hand reverse pinch indicating the box diagonal length and lift off for height then pinch to commit. *[H01-02]*

All select: Drag index from one corner to other two corners around the workspace. *[H11]*

Scale Uniform (1): Two hands move apart/together along X-axis to enlarge/shrink *[H09]*

Rotate X-axis (Roll): Turning the wrist up/down, palm facing sideward. *[H01-04]*

Rotate Y-axis (Pitch): Turning wrist CW/ CCW, palm facing away from body. *[H01-04]*

Scale Uniform (2): Two hands grab each diagonal corner of target move apart/together along XY plane to enlarge/shrink. *[H01-04]*

Scale X-axis (1): Two hands grab left/right side of target move apart/together along X-axis to enlarge/shrink. *[H01-04,08]*

Scale Y-axis (1): Two hands grab front/back side of target move apart/together along Y-axis to enlarge/shrink. *[H01-04,08]*

Scale Z-axis (1): Two hands grab top/bottom side of target move apart/together along Y-axis to enlarge/shrink. *[H01-04,06,07]*

Rotate Z-axis (Yaw): Turning the wrist in/out, palm down/sideward. *[H01-05]*

Scale Uniform (3): Move thumb and other fingers apart/together diagonally along XY plane to enlarge/shrink. *[H08]*

Scale X-axis (2): Move thumb and other fingers apart/together along X-axis to enlarge/shrink. *[H08]*

Scale Y-axis (2): Move thumb and other fingers apart/together along Y-axis to enlarge/shrink. *[H08]*

Scale Z-axis (2): Move thumb and other fingers apart/together along Z-axis to enlarge/shrink. *[H08]*

Previous: Swipe left to right. *[H08,10-11]*

Next: Swipe right to left. *[H08,10-11]*

Play, increase-speed, redo: Spin CW. *[H11]*

Decrease-speed, undo: Spin CCW. *[H11]*

Pause: Victory-tory-

Stop(1): Open hand facing away.

Stop (2): Show a fist.

Accept: Thumb up

Reject: Thumb down

Group: Two hands move together. *[H09]*

Ungroup: Two hands move apart. *[H09]*

Move, insert, paste (1): Select target from menu/clipboard, move it to a location to place. *[H01-05]*

Delete (1): Grasp the target and crush it. *[H08]*

Copy (1): One hand covers the target and another move target to clipboard area. *[H01-05]*

HM Open: Swipe out. [H06,08,10-11]

HM Close: Swipe in. [H06,08,10-11]

HM Select: Tap an option on the surface. *[H11]*

Delete (2): Throw away the target *[H01-05]*

VM Open: Pull up. *[H06,09,10-11]*

VM Close: Push down. *[H06,09,10-11]*

VM Select: Push in on an option. *[H11]*

Move, insert, paste (2):Select target from menu/clipboard, tap at a location to place. *[H10-11]*

Cut: Snap index & middle (scissor pose)

Copy (2): Two hands turn away, imitate open a book. *[change from H07 to H09]*

OM Open: Splay all fingers. *[H09]*

OM Close: Regroup all fingers. *[H09]*

OM Select: Tap an option on the surface. *[H11]*

Fig. 5. The user-defined gesture set for AR. The number shown in the parenthesis indicates multiple gestures in the same task. The codes in the square bracket indicate the hand pose variants (Figure 4) that can be used for the same gesture.

Influence from Existing UI. When participants found it difficult to come up with a gesture for a particular task, they would often resort to using metaphors from familiar UI. For example when designing a gesture for the *delete* task, several participants imagined having a recycle bin that they could move the target object to. For other arbitrary tasks, users would often resort to *double-tapping*. Some examples of how participants explained these actions were:

"I would select and double-click… I'm thinking too much like Microsoft. It's just the thing that I'm used to." – P10

"The way I do it on my phone is that I would scale like this and then tap it once." – P14

The Obvious and the Obscure. Gesturing in 3D space allows for higher expressiveness, which in turn led to use of commonly understood gestures in the real-world. For example, there was a high level of agreement on the *symbolic* gestures thumbs up/down for *accept*/*reject* with scores of 9 and 10 respectively (out of 20). This was also the case for *metaphoric* gestures such as a scissor gesture for the *cut* task with the score of 7. User's liked the idea of using these gestures from the real world, resulting in higher than average goodness and ease scores, with averages of 6.87/6.75 (σ =.35/.71) for thumbs up, 6.5/6.5(σ =.71/.85) for thumbs down and 6.5/6.67(σ =.84/.82) for scissor pose.

The majority of participants found it challenging to come up with metaphors to design gestures for 3D tasks that they referred to as "abstract", such as *box selection*. In this task, users' had to design a gesture to define the width, depth and height of a 3D bounding box around target objects for selection. There was little agreement upon a common gesture, with a low agreement score of 0.095. In cases where the agreement score is below 0.1, we recommend further rigorous studies and usability tests to select the best gesture for the task. One participant expressed an opinion which was shared by many others:

"I don't think that it's unsuitable (the proposed gesture) but it's just very arbitrary and there is not a lot of intrinsic logic to it. If somebody told me that this is how to do it then I would figure it out but it's not obvious. It's just an arbitrary way of selecting a 3D area." - P11

Feedback Backfired. Our experimental design included the use of a 3D camera to support hand occlusion, which gave users some concept of the relative position between the virtual contents and their hands, however some participants found it to be obtrusive. We present ideas on how to improve this component of the experience in Section 5.1. One example of this criticism was as follows:

"Your hand gets in the way of the object so it can be hard to see how you're scaling it." – P11

Menus for AR. There was no significant difference in menu ranking. Some participants favored the horizontal menu because it was simple, familiar, easy to use/understand, supported on-the-surface gestures for touch sensing and did not interfere with virtual content. Others disliked the horizontal menu and felt it did not take advantage of 3D space with some options being further away and hard to reach.

The majority of participants found the vertical menu novel, and some found it to be appealing, easy to understand and that it made a good use of space with the distance to all options was evenly distributed. However, some found it harder to operate as they needed to lift their hands higher for options at the top if the buttons were arranged vertically.

Finally, some participants liked the object-centric menu because it was unique and object-specific so they knew exactly which object they were dealing with. However, some participants thought that it was unnatural and harder to operate in a crowded workspace. Furthermore, the open/close gestures for the object-centric menu were not as obvious, as indicated by the low agreement score of 0.11, as opposed to horizontal and vertical that scored 0.905.

Axes and Boxes. The *rotation* and *scaling* tasks, allowed for three possible coordinate systems, local, global, and user-centric, which corresponded to the *object-centric*, *world-dependent*, and *world-independent* categories in the *binding* dimension. In practice, we found that the transformations were mostly *object-centric*; the participant would perform gestures based on the direction of the transformation presented on the object. This was expected because people would naturally perform these tasks physically and adapted their bodies and gestures to suit the operation.

To perform a *rotation*, participants would grasp the object with at least two contact points and would move their hand or turn their wrist accordingly. For *scaling* on 3 axes, participants would grasp or use open-hands to align with the sides of object and increased or decreased the distance between them to enlarge or shrink in the same direction as the transformation. *Uniform scaling* was less obvious, for example some participants preferred using open hands moving along a single axis in front of them, as shown in Figure 5 *uniform scale (1)*. Others preferred grasping the objects' opposing diagonal corners and moving along a diagonal line across the local plane parallel to the table surface as shown in Figure 5 *uniform scale (2)*. Some user's expressed concern about how to perform the task for a round object, and suggested that bounding volumes must be provided for these models for manipulation.

Variation of Hand Poses. Variants of a single hand pose were often used across multiple participants, and sometimes even by a single participant. We clustered common hand poses into eleven poses, as shown in Figure 4. Multiple hand poses can be used interchangeably for each gesture in a given task.

5 Discussion

In this section, we discuss the implications of our results for the fields of AR, gesture interfaces, and gesture recognition.

5.1 Implications for Augmented Reality

While our experiment was conducted in a tabletop AR setting, the majority of the user-defined gestures are equally suitable to be performed in the air. Only four gestures were *on-the-surface*, *select all* and *open/close/select horizontal menu*, with three *mixed locale*, *box select (1)* and *insert/paste (2)*. This opens up our gesture set to other AR configurations, including wearable interfaces.

For our experiment, we implemented hand occlusion to give better understanding of the relative positions of the users' hands and virtual content. However we found that this could hinder user experience when virtual objects are smaller than the user's hand, occluding the object completely. We recommend that the hands should be treated as translucent rather than opaque, or occluded objects are rendered as outlines to provide some visual feedback of the objects' location.

As discussed in *axes and boxes* motif, a clear indicator of axes and bounding boxes should be provided during object manipulation tasks. Due to an absence of haptic feedback, visual feedback should be provided to inform users of the contact points between hands and objects.

5.2 Implications for Gesture Interfaces

We found most of the gestures elicited were *physical* (39%). Wobbrock et al. reached a similar outcome for surface gestures and suggested using a physics engine for handling these gestures. This approach was implemented by Hilliges et al. [7] and Benko et al. [8], who introduced "physically-based interaction", however only basic manipulations were demonstrated, with limited precision and control over the virtual contents. We believe that better control can be achieved by manipulation of the dynamical constraints imposed by the engine. Many gestures can make use of the collision detection component without the dynamics for tasks such as object selection, scaling etc.

In the *size does matter* motif we described how object size influences the number of hands used for manipulation. Since the resulting user-defined gesture set contains both one-handed and two-handed gestures for tasks such as scaling, we suggest taking an advantage of this fact to provide different levels of control. For example, in scaling tasks, by combining a snap-to feature for different granularities, unimanual scaling could offer snap-to in millimeter steps and bimanual in centimeter steps, as users tend to use one hand for an object smaller than their palm's size and two when it is larger.

As mentioned in *the obvious and the obscure* motif, care must be taken when choosing gestures for tasks with low agreement scores. We recommend follow up studies to determine usability by comparing these gestures, designer-refined gestures, menu options and even alternative modalities in case of multimodal interface.

5.3 Implications for Gesture Recognition

High degree of freedom hand pose recognition is achievable, however it is computationally expensive. In the *variation of hand poses* motif, we found a limited number of common poses (Figure 4), reducing the search space. Furthermore, the majority of the resulting gestures were *static pose and path*, which are simpler to recognize than *dynamic pose and path* gestures.

6 Conclusion and Future Work

We have presented an experiment and the results of a guessability study for natural hand gestures in AR. Using the agreement found among the elicited gestures, 44 user-defined gestures were selected as a "consensus set". Although gestures were found for all 40 tasks, agreement scores varied, suggesting that some gestures are more universally accepted than others. We are conducting a further study to validate our gestures, where a different group of participants will be shown the elicited gestures from both consensus and discarded sets and their preferences determined for each task to confirm our result.

References

1. Azuma, R.: A Survey of Augmented Reality. Presence 6, 355-385 (1997)
2. Heidemann, G., Bax, I., Bekel, H.: Multimodal interaction in an augmented reality scenario. In: ICMI'04 - Sixth International Conference on Multimodal Interfaces, October 14, 2004 - October 15, 2004, pp. 53-60. Association for Computing Machinery, (2004)
3. Kolsch, M., Bane, R., Hollerer, T., Turk, M.: Multimodal interaction with a wearable augmented reality system. IEEE Computer Graphics and Applications 26, 62-71 (2006)
4. Kaiser, E., Olwal, A., McGee, D., Benko, H., Corradini, A., Li, X., Cohen, P., Feiner, S.: Mutual disambiguation of 3D multimodal interaction in augmented and virtual reality. In: ICMI'03: Fifth International Conference on Multimodal Interfaces, November 5, 2003 - November 7, 2003, pp. 12-19. Association for Computing Machinery, (2003)
5. Lee, T., Hollerer, T.: Handy AR: Markerless inspection of augmented reality objects using fingertip tracking. In: 11th IEEE International Symposium on Wearable Computers, ISWC 2007, October 11, 2007 - October 13, 2007, pp. 83-90. IEEE Computer Society, (2007)
6. Fernandes, B., Fernandez, J.: Bare hand interaction in tabletop augmented reality. In: SIGGRAPH 2009: Posters, SIGGRAPH '09, August 3, 2009 - August 7, 2009. Association for Computing Machinery, (2009)
7. Hilliges, O., Kim, D., Izadi, S., Weiss, M., Wilson, A.: HoloDesk: direct 3d interactions with a situated see-through display. Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems, pp. 2421-2430. ACM, Austin, Texas, USA (2012)
8. Benko, H., Jota, R., Wilson, A.: MirageTable: freehand interaction on a projected augmented reality tabletop. Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems, pp. 199-208. ACM, Austin, Texas, USA (2012)
9. Wobbrock, J.O., Morris, M.R., Wilson, A.D.: User-defined gestures for surface computing. Proceedings of the 27th international conference on Human factors in computing systems, pp. 1083-1092. ACM, Boston, MA, USA (2009)

10. Ruiz, J., Li, Y., Lank, E.: User-defined motion gestures for mobile interaction. Proceedings of the 2011 annual conference on Human factors in computing systems, pp. 197-206. ACM, Vancouver, BC, Canada (2011)
11. Wobbrock, J.O., Aung, H.H., Rothrock, B., Myers, B.A.: Maximizing the guessability of symbolic input. CHI '05 extended abstracts on Human factors in computing systems, pp. 1869-1872. ACM, Portland, OR, USA (2005)
12. Lee, J.Y., Rhee, G.W., Seo, D.W.: Hand gesture-based tangible interactions for manipulating virtual objects in a mixed reality environment. International Journal of Advanced Manufacturing Technology 51, 1069-1082 (2010)
13. Lee, T., Hollerer, T.: Multithreaded Hybrid Feature Tracking for Markerless Augmented Reality. Visualization and Computer Graphics, IEEE Transactions on 15, 355-368 (2009)
14. Olwal, A., Benko, H., Feiner, S.: SenseShapes: using statistical geometry for object selection in a multimodal augmented reality. In: Mixed and Augmented Reality, 2003. Proceedings. The Second IEEE and ACM International Symposium on, pp. 300-301. (2003)
15. Lee, M.: Multimodal Speech-Gesture Interaction with 3D Objects in Augmented Reality Environments. Department of Computer Science and Software Engineering, vol. Doctor of Philosophy. University of Canterbury, Christchurch, New Zealand (2010)
16. Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohli, P., Shotton, J., Hodges, S., Fitzgibbon, A.: KinectFusion: Real-time dense surface mapping and tracking. In: 2011 10th IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2011, October 26, 2011 - October 29, 2011, pp. 127-136. IEEE Computer Society, (2011)
17. Schuler, D., Namioka, A.: Participatory Design: Principles and Practices. Lawrence Erlbaum, Hillsdale, NJ (1993)
18. van Krevelen, D.W.F., Poelman, R.: A Survey of Augmented Reality Technologies, Applications and Limitations. The International Journal of Virtual Reality 9, 1-20 (2010)
19. Broll, W., Lindt, I., Ohlenburg, J., Wittkamper, M., Yuan, C., Novotny, T., Fatah, g., Mottram, C., Strothmann, A.: ARTHUR: A Collaborative Augmented Environment for Architectural Design and Urban Planning. Journal of Virtual Reality and Broadcasting 1, (2004)
20. Lee, G.A., Billinghurst, M., Kim, G.J.: Occlusion based interaction methods for tangible augmented reality environments. In: Proceedings VRCAI 2004 - ACM SIGGRAPH International Conference on Virtual Reality Continuum and its Applications in Industry, June 16, 2004 - June 18, 2004, pp. 419-426. Association for Computing Machinery, (2004)
21. Clark, A. and Green, R. Optical-Flow Perspective Invariant Registration. In Proc. International Conference on Digital Image Computing: Techniques and Applications, 117-123. (2011)
22. Piumsomboon, T., Clark, A., Umakatsu, A., Billinghurst, M.: Poster: Physically-based natural hand and tangible AR interaction for face-to-face collaboration on a tabletop. In: 3D User Interfaces (3DUI), 2012 IEEE Symposium on, pp. 155-156. (2012)