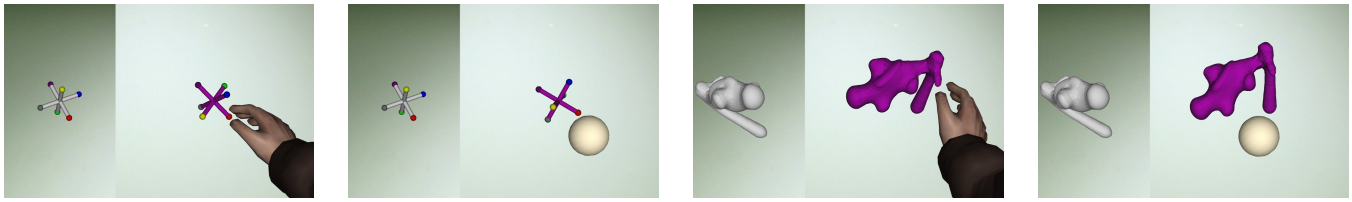
Kinect Based 3D Object Manipulation on a Desktop Display

Mukund Raj* [1] Sarah H. Creem-Regehr [2] Kristina M. Rand [2] Jeanine K. Stefanucci [2] William B. Thompson [1]
[1]School of Computing [2]Department of Psychology
University of Utah



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

Figure 1: *A training trial with (a) self-avatar display and (b) sphere display. A test trial with (c) self-avatar display and (d) sphere display.*

Abstract

Gesture-based controllers such as the Microsoft Kinect are low cost devices that allow a user to interact with complex, three-dimensional simulations using an interface argued to be more natural than game controllers, joy sticks, or a mouse and keyboard. This paper presents a controlled experimental evaluation of the use of Microsoft Kinect to support a 3D object manipulation task. Users were asked to match the orientation of objects with a manipulation interface that displayed either a self-avatar hand and arm or a sphere, both corresponding to users' arm gestures and wrist rotation. Our results show that while there was no overall difference in performance between the self-avatar and sphere visual display conditions, there were clear differences in the two visual display conditions as a function of gender and video-game experience.

CR Categories: I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Virtual Environments H.5.2 [Information Systems]: Information Interfaces and Presentation—User Interfaces;

Keywords: virtual environments, human perception and performance, self avatars, object rotation, kinect, user interfaces

1 Introduction

The ability of the user to efficiently manipulate a 3D object in a virtual world viewed on a desktop display is a challenging task for many users. 3D object rotation is particularly difficult, and has been shown to be slow and relatively difficult for users in comparison to translation of 3D objects [Ware and Rose 1999]. In this paper, we focus on evaluating a gesture-based method of interaction which uses a conventional Microsoft Kinect sensor. A controlled perceptual experiment is used to compare user performance in an object rotation task where two different methods are used to provide feedback about how the user's hand motions correspond to the virtual world being manipulated. One method uses a generic marker to signal the point on a virtual object on which the user's gesture is operating. The second method provides this information using a self-avatar rendered to move with the user's real hand and arm.

Jacob et al. [1994, p. 6] argue for the importance of "matching the perceptual structure of the task and the control structure of the input device" (see also [Darken and Durost 2005]). In other words, in the context of 3D object orientation matching, performance should be best when the control space of the input device is as close as possible to real rotations of 3D objects. Methods of interaction with a virtual world aimed at aiding rotations have included devices such as a virtual sphere, arcball and magnetic orientation tracker with which the user can rotate an object in the virtual world to any desired orientation [Hinckley et al. 1997]. Others have used hybrid environments in which a real object is manipulated in order to control a virtual display, such as a doll's head used to manipulate brain images [Hinckley et al. 1994]. Another approach for accomplishing rotations is by automatically rotating the object to certain prespecified orientations using a ViewCube [Khan et al. 2008]. The ViewCube provides an easy way to rotate to some key orientation, but does not allow rotation to arbitrary angles and, along with other methods mentioned earlier, has the inherent drawback of not relating well to how we rotate objects with our hands in the real world.

We introduced a self-avatar as a means of visual feedback to test whether it may facilitate the task of 3D object manipulation. Research on the effects of self-avatars on spatial task performance has been mixed. Some previous work suggests that the portrayal of a self-avatar in immersive virtual environments results in more accurate estimates of distance in the virtual world [Mohler et al. 2010]. However, this research involved very different spatial judgments within a region of space that extends well beyond that which can be reached to and manipulated with the hand. In an object-manipulation task comparing real, purely virtual, and hybrid virtual environments with real objects, Lok et al. [2003] found that the realism of an avatar hand had little effect. Behavioral and neuroscience-based studies of *mirror systems* show close parallels between the processes involved in the observation of action with that of overt action [Wilson 2002]. Thus, it is possible that observation of a virtual limb while using a gesture interface will facilitate performance on a rotation task because the interaction with the virtual object becomes more *embodied*.

An inherent problem in working with complex 3D graphical objects presented on desktop displays is maintaining an understanding of an object when viewed in different orientations. In many applica-

*e-mail:mraj@cs.utah.edu

SAP 2012, Los Angeles, CA, August 3 – 4, 2012.
© 2012 ACM 978-1-4503-1431-2/12/0008 $15.00

tions such as mechanical CAD or medical visualizations, objects are presented on a display without the frames of reference present in the real world [Ziemek et al. 2012]. Our current interface may address this problem by providing visual feedback during the object manipulation task. One possibility is that the virtual hand provides an additional familiar egocentric frame of reference to aid in the rotation of an unfamiliar object. Thus, our goal was to test whether we provide not only a more natural means of object manipulation, but also an additional visual frame of reference that could be used to understand changes in orientation.

Our system was built using a commodity sensor Microsoft Kinect. Following the argument in Jacob et al. [1994], we provided an extra degree of freedom of rotation about the user's wrist, which makes object handling as intuitively close to the real world as possible. While we used an Intersense InertiaCube3 (IC3) for this purpose, inexpensive commodity orientation sensors such as those in smart phones and tablets could also be used. We examined whether there would be an advantage in response time for the self-avatar versus the generic sphere display as a representation of the user's rotational device, whether there would be differences in the users' preference to either use an arm gesture or a wrist rotation to manipulate the objects, and whether gender and/or gaming experience would influence task performance.

2 Experiment

An orientation matching task was used to examine rotation performance for participants receiving visual feedback in the form of a self-avatar or a generic sphere. All participants performed 12 trials in which they were presented with two identical objects at varying 3D orientations on a desktop display, and were asked to manipulate the object on the right to match that of the object on the left. Twenty-three University of Utah students (13 male, 10 female) participated for compensation of $5 dollars.

Participants interacted with our system, which used the joint orientation data from the Kinect to animate either the partial self-avatar of the user, or a sphere to provide visual feedback to the participant. The virtual environment was rendered using WorldViz Vizard and FAAST [Suma et al. 2011] was used to read and process Kinect data feed to get the user's joint orientations. In the self-avatar condition, participants viewed a virtual arm closely following the movements of their real arm. For those in the sphere condition, in place of the hand was a sphere. Thirteen objects were presented to the participants: one for the practice trials, and 12 additional objects for the experimental test trials. The practice object was in the shape of a jack with 6 bars with different colored ends. The objects used in the test trials were a subset of the anatomical objects used in Ziemek et al. [2012] created using digital embryos [Brady and Kersten 2003] (see Figure 1). The radius of the bounding sphere of the objects ranged from 0.16m to 0.29m and were rendered 1.4m in front of the avatar's position. All objects were displayed on an Asus ProArt Series 24.1 inch display. The distance of the participant from the display was 1.81m and the geometric field of view of the participant was matched with the rendered display field of view. The wireless IC3 was attached to a strap which was then fastened over the right hand of the participant such that the orientation sensor was situated over the back of the participant's right hand. See Figure 2.

A between-subjects design was used for the visual feedback condition (self-avatar $n = 11$, sphere $n = 12$). Two randomized orders of experimental trials were also manipulated between subjects. Participants performed 6 practice trials, and 12 experimental trials, for a total of 18 trials. Objects presented on each trial differed from the goal orientation by 28 to 177 degrees using direct quaternion rotation.

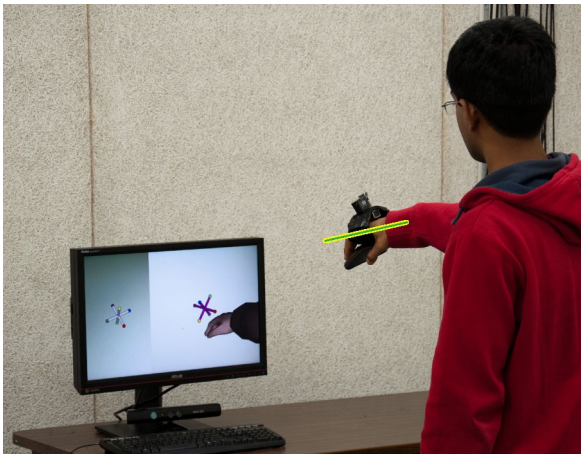


Figure 2: *The experimental setting and interface. The overlay is a representation of the axis of wrist-rotation.*

Participants were instructed that the goal of the experiment was to manipulate the object on the right side of the screen to match the orientation of the object on the left side of the screen as quickly as possible. Participants were equipped with the InertiaCube and mouse on their right hand, and led to the starting location. They were then informed of the two modes in which they would be able to manipulate the object: *swipe* or *twist*. The experimenter then demonstrated the distinction between the modes on a real object with their hands. The *swipe* mode used a drag motion performed with the arm similar functionally to methods such as the virtual sphere or arcball, where the object rotates in the direction of the drag. The *twist* mode referred exclusively to changes in the orientation due to rotation of the about the wrist axes, closely resembling what would be expected in the real world. In both conditions, the rotation of the wrist about the axis joining the elbow to the wrist was also accurately mapped using orientation data from the wireless IC3. Visual feedback for wrist rotation was evident in the hand condition but absent in the sphere condition. Finger joints were not animated. In both rotation modes, manipulation was only possible if the hand was close to the object, conveyed by a change in the color of the object. There was also the option of ratcheting to accomplish large rotations as a sum of smaller motions and scaling of the user action based on the speed of movement to maintain precision as well as range of manipulation.

Participants were informed that a left or right mouse click would perform the swipe or twist, respectively, and were asked to demonstrate their understanding of the distinction between the modes. Following the instructions, participants performed the practice and test trials. All trials began with the presentation of two objects at differing orientations. After three seconds, a prompt on the screen instructed the participant to begin. At this point participants could use the two modes of object manipulation to match the objects. When the orientation of the object was within 15 degrees of the orientation of the target object, a "match detected" prompt appeared on the screen. During experimental trials, if 90 seconds elapsed between the onset of the trial before a match was detected, a prompt indicated that the allotted time expired. This time interval was chosen to allow most of the trials to be completed but also to reduce the chance of frustration or fatigue on an individual trial level. After completion of the trials, participants were given a brief video game experience questionnaire, with rating scales ranging from 1-7 on first-person video game, gesture-based game, and third-person game (such as racing, sports) experience.

3 Results

Time to rotate each object was averaged across the 12 trials for each participant. Rotation time was slightly faster overall for the self-avatar (30.36 sec) versus the sphere conditions (34.81 sec). A 2 (visual display) x 2 (gender) ANOVA was performed on average response time. While there was not a significant difference overall between the self-avatar and sphere conditions ($p = .43$), there was a significant visual display x gender interaction, $F(1, 19) = 4.49$, $p < .05$, $\eta_p^2 = .19$, showing that while males and females showed no difference in performance on the self-avatar task ($p = .84$), males outperformed females on the sphere task ($t(9) = 3.32$, $p < .01$). See Figure 3. It is important to note that gender and gaming experience were highly related in our sample (see Figure 5), as discussed further in Section 4.

A second 2 (visual display) x 2 (gender) ANOVA was performed on average number of timeouts (in which the participant did not complete the rotation task within the allotted time of 90 sec). Similar to the rotation time analysis, the only significant effect found was a visual display x gender interaction, $F(1, 19) = 6.76$, $p < .02$, $\eta_p^2 = .26$, showing that there was no difference between male (1.43) and female (.80) performance for the self-avatar condition ($p = .52$), but significantly more timeouts for females (2.80) compared to males (.17) in the sphere condition ($t(9) = 3.30, p < .01$).

Given the two different modes to rotate (swipe and twist performed with the left and right mouse clicks, respectively), we also analyzed total number of left and right mouse clicks with a 2 (visual display) x 2 (gender) x 2 (click: left vs. right) mixed ANOVA with left/right click as a within-subjects variable. The analysis showed a greater number of total left (87.32) versus right (66.28) clicks, $F(1, 19) = 4.52$, $p < .05$, $\eta_p^2 = .19$, as well as a click x gender interaction, $F(1, 19) = 4.44$, $p < .05$, $\eta_p^2 = .19$. The interaction revealed that males used swipe (left click) and twist (right click) modes equally ($p = .98$), but females used the swipe more than the twist mode, $t(9) = 2.27$, $p < .05$. See Figure 4. The visual display condition did not influence total number of left or right clicks.

Finally, we examined the influence of self-reported video game experience on rotation time by performing separate bivariate correlations between average rotation time and the three video game rating scales for both the self-avatar and sphere conditions. Gesture-gaming experience did not correlate with rotation time for either display conditions. Both first- and third-person game experience correlated with rotation time across the display conditions (see Figure 5), but effects were greater for the sphere versus the self-avatar display. (First-person: Self-avatar $R = -.55$, $p < .06$, Sphere $R = -.71$, $p < .02$; Third-person: Self-avatar $R = -.42$, $p < .16$, Sphere $R = -.77$, $p < .01$).

4 Discussion

The results of our experiment show the feasibility of using off the shelf technology, such as the Microsoft Kinect, to drive user interfaces that aid in manipulating 3D objects on a desktop display. Our findings suggest that care should be taken to understand the individual differences among users that could interact with display type. Here we showed that the gender and video game experience of the participant influenced performance, specifically in the sphere display condition which presented less information about the relationship between the orientation of the participants' real hands and the orientation of the displayed interface. Given the individual differences found only in the sphere condition, we suggest that the self-avatar provided additional body-based information that may be beneficial to a broader population of users.

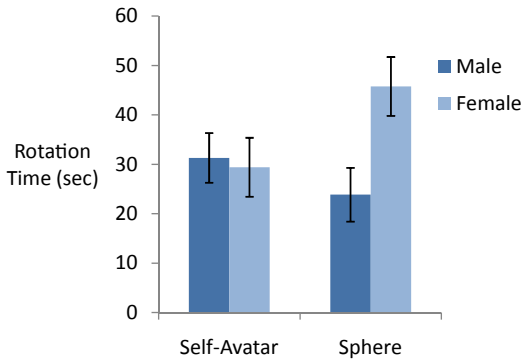


Figure 3: *Mean rotation time (± 1 SE) for the self-avatar and sphere conditions by gender.*

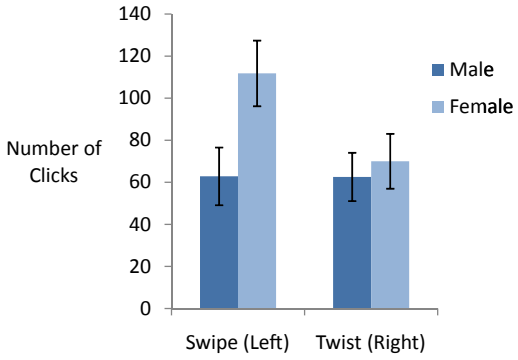


Figure 4: *Average total number of left and right button clicks (± 1 SE) by gender.*

There are several experimental design issues to consider when interpreting our results. First, we did not employ a condition in which no visual feedback was displayed on the screen, so we cannot make conclusions about the presence versus absence of the visual display of the user's movements. Second, we limited the time that participants were able to manipulate the objects in order to reduce frustration and increase motivation. Future work could give users unlimited time to more fully assess reaction time for successful manipulation. Finally, while there were strong correlations between gaming experience and task performance, it is important to note that in our current sample, gender differences highly overlapped with video-game experience (see Figure 5). Thus, it is unknown whether the differences seen in the two display conditions as a function of individual differences will replicate with a greater range of female gamers or male non-gamers.

However, the individual differences affecting performance in the sphere condition are not surprising given previous work demonstrating gender differences in *spatial abilities*, defined as a range of skills involving the mental representation and manipulation of information about geometric entities [Hegarty and Waller 2005]. The best documented gender-related performance difference in spatial abilities is a male advantage in mental rotation tasks, in which viewers are asked to determine the congruence between images of two static objects presented at differing orientations [Linn and Petersen 1985]. Some individuals find it very challenging to determine correspondence between multiple views of 3D objects, particularly when they are irregular shapes [Ziemek et al. 2012].

The present gender differences are also consistent with some prior

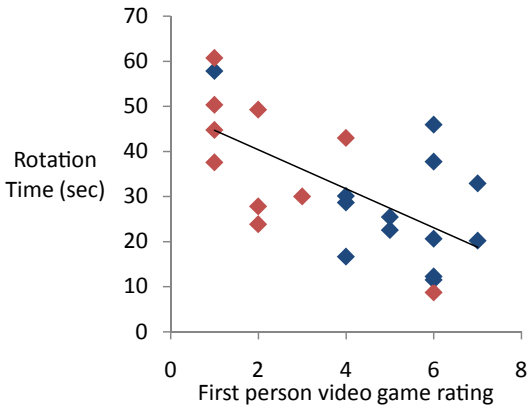


Figure 5: *Correlation ($R = -.63$, $p < .001$) between first person video game experience and rotation time. There is high overlap between gaming experience and gender (red symbols = females, blue symbols = males).*

work on the use of 3D user interfaces. Hinckley et al. [1997] found overall faster performance for males compared to females across the 3D interfaces tested, but they did not further test explanations for this effect. An explicit investigation of gender differences in 3D user interfaces concluded that the poorer performance of women compared to men in navigating virtual environments disappeared if users were provided with a wide field of view display [Tan et al. 2003]. In another applied study of gender differences, Hubona and Shirah [2006] examined performance on several spatial tasks relevant to visual interfaces and found a male advantage on mental rotation of abstract objects, the use of motion-related cues, and on a task that involved moving and positioning objects, but a female advantage on a size perception task. These results suggest the importance of adapting user interfaces in ways that make them accessible for all users. Work to date on designing user interface software with an awareness of the effects of individual differences has been limited, although some progress has been made [Ziemek et al. 2012].

5 Conclusion

With the advent of new low-cost gesture-based interfaces, such as the Microsoft Kinect, users are able to more easily interact with 3D objects in a desktop display. Here, we show the feasibility of two different visual displays for interaction, portraying feedback to the user's actions as either a self-avatar or a sphere. We find that the time taken to rotate the objects to match the target and the number of successful trials within the allotted time did not differ across display conditions when averaging across all users. However, gender differences that were also related to video gaming experience did influence performance in the sphere display condition, which provided less of an egocentric frame of reference and was less anthropomorphic. Thus, when designing interfaces for object manipulation, individual differences in users' spatial abilities and experience should be taken into account in order to determine the interface that is most advantageous for the highest number of users.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1116636. We thank Jennifer Bishoff for help in data collection.

References

BRADY, M. J., AND KERSTEN, D. 2003. Bootstrapped learning of novel objects. *Journal of Vision 3*, 6, 413–422.

DARKEN, R. P., AND DUROST, R. 2005. Mixed-dimension interaction in virtual environments. In *Proc. ACM Symposium on Virtual Reality Software and Technology*, 38–45.

HEGARTY, M., AND WALLER, D. 2005. Individual differences in spatial abilities. In *Handbook of Higher-level Visuospatial Thinking*, P. Shah and A. Miyake, Eds. Cambridge University Press, New York, 121–169.

HINCKLEY, K., PAUSCH, R., GOBLE, J. C., AND KASSELL, N. F. 1994. Passive real-world interface props for neurosurgical visualization. In *Proc. SIGCHI Conf. on Human Factors in Computing Systems: Celebrating Interdependence*, ACM, 452–458.

HINCKLEY, K., TULLIO, J., PAUSCH, R., PROFFITT, D., AND KASSELL, N. 1997. Usability analysis of 3D rotation techniques. In *Proc. 10th ACM Symp. on User Interface Software and Technology*, 1–10.

HUBONA, G. S., AND SHIRAH, G. W. 2006. The paleolithic stone age effect?: Gender differences performing specific computer-generated spatial tasks. *International Journal of Technology and Human Interaction 2*, 2, 24–48.

JACOB, R. J. K., SIBERT, L. E., MCFARLANE, D. C., AND MULLEN, J. M. P. 1994. Integrality and separability of input devices. *ACM Trans. on Comput-Hum Interaction 1*, 1, 3–26.

KHAN, A., MORDATCH, I., FITZMAURICE, G., MATEJKA, J., AND KURTENBACH, G. 2008. Viewcube: A 3D orientation indicator and controller. In *ACM Symp. on Interactive 3D Graphics*, 17–25.

LINN, M. C., AND PETERSEN, A. C. 1985. Emergence of characterization of sex differences in spatial ability: A meta-analysis. *Child Development 56*, 6, 1479–1498.

LOK, B., NAIK, S., WHITTON, M., AND BROOKS, J. F. P. 2003. Effects of handling real objects and self-avatar fidelity on cognitive task performance and sense of presence in virtual environments. *Presence: Teleoperators and Virtual Environments 12*, 6 (Dec.), 615–628.

MOHLER, B. J., CREEM-REGEHR, S. H., THOMPSON, W. B., AND BUELTHOFF, H. H. 2010. The effect of viewing a self-avatar on distance judgments in an HMD-based virtual environment. In *Presence: Teleoperators and Virtual Environments*, 230–242.

SUMA, E. A., LANGE, B., RIZZO, A., KRUM, D., AND BOLAS, M. 2011. FAAST: the flexible action and articulated skeleton toolkit. In *IEEE Virtual Reality*, 247–248.

TAN, D. S., CZERWINSKI, M., AND ROBERTSON, G. G. 2003. Women go with the (optical) flow. In *Proc. SIGCHI Conf. on Human Factors in Computing Systems*, 209–215.

WARE, C., AND ROSE, J. 1999. Rotating virtual objects with real handles. *ACM Trans. on Comput-Hum Interaction 6*, 2, 162–180.

WILSON, M. 2002. Six views of embodied cognition. *Psychonomic Bulletin and Review 9*, 625–636.

ZIEMEK, T., CREEM-REGEHR, S. H., THOMPSON, W. B., AND WHITAKER, R. 2012. Evaluating the effectiveness of orientation indicators with an awareness of individual differences. *ACM Transactions on Applied Perception. 9*, 2.