The Costs and Benefits of Combining Gaze and Hand Gestures for Remote Interaction

Yanxia Zhang[1(✉)], Sophie Stellmach[2], Abigail Sellen[3], and Andrew Blake[3]

[1] Lancaster University, Lancaster, UK
yazhang@lancaster.ac.uk
[2] Microsoft Corporation, Redmond, USA
sostel@microsoft.com
[3] Microsoft Research, Cambridge, UK
{asellen,ablake}@microsoft.com

Abstract. Gaze has been proposed as an ideal modality for supporting remote target selection. We explored the potential of integrating gaze with hand gestures for remote interaction on a large display in terms of user experience and preference. We conducted a lab study to compare interaction in a photo-sorting task using gesture only, or the combination of gaze plus gesture. Results from the study show that a combination of gaze and gesture input can lead to significantly faster selection, reduced hand fatigue and increased ease of use compared to using only hand input. People largely preferred the combination of gaze for target selection and hand gestures for manipulation. However, gaze can cause particular kinds of errors and can induce a cost due to switching modalities.

Keywords: Hand gestural interface · Gaze interaction · Mid-air gestures · Remote interaction · Large display · Smart living room

1 Introduction

With advances in sensing technologies, people can now interact in much richer ways with computer systems without the need for physical contact or manipulation of devices. Free hand gestures, such as those enabled by the Kinect depth-sensing camera, have already been shown to be an effective method of input for games and interactive television applications on displays at a distance. This raises the question of whether other input modalities which support interaction at a distance, such as eye gaze, can further enhance the way we interact remotely. One reason for this is that many applications performed on a large screen over a distance often involve manipulating contents of the entire screen space, which can be widely dispersed [1]. Because people instinctively look at objects of interest, gaze has been shown to be an efficient modality for targeting remote objects [2–6]. In addition, gaze often precedes a manual action [7] suggesting that we could exploit the combination of these two modalities for more efficient methods of interaction. For example, it suggests that gaze might be used for selection of a target object followed by hand gestures to operate on that object.

J. Abascal et al. (Eds.): INTERACT 2015, Part III, LNCS 9298, pp. 570–577, 2015.
DOI: 10.1007/978-3-319-22698-9_39

Recent work has highlighted the potential of combining gaze and hand gestures for fast and accurate *point and drag* interactions [8], and enabling more attentive and immersive 3D UI interactions [9]. While this work focused on examining speed and accuracy, many other aspects of integrating gaze with hand gesture input for remote large screen interactions are still not well understood. To further explore how people perform and perceive the combination, we ask what benefits gaze might bring to gestural interaction from the user's perspective. We also investigate any potential costs that are incurred by adding gaze as a second modality.

In order to address these questions, we constructed a photo sorting task involving the fast assignment of multiple objects spread out across a large screen to different "piles" or destinations. This task, which involves selection of remote target objects followed by a limited set of repetitive actions on those objects, has been shown to be representative of common tasks that users would want to carry out on large screens [1].

We compared users' experience of two interfaces for this task: one which represents the status quo for device-free remote interaction such as gaming, namely using gestures only *(hand-only),* and another which works across modalities (*gaze-hand).* In the cross modality condition, we assigned gaze and hand gestures in what we surmised would be the optimal way for such a task. Based on the prior work mentioned above, we reasoned that gaze would be best for target selection (of relatively large targets), while hand gestures would be good for expressive but not necessarily precise manipulation. In this case, gestures are used to "fling" photos to different piles, an action which involves coarse-grained (as opposed to fine-grained) control.

2 User Study: Photo Sorting

In this study, our goal was to examine the effect of integrating gaze into a remote, gestural interaction task in terms of both performance and subjective experience. We hypothesized that: (1) the combination of gaze pointing and hand gestures will give rise to faster task completion than sorting with gesture only; and (2) using gaze will introduce an extra mental load due to the need to work across modalities during the task. We made no a priori predictions about subjective experience.

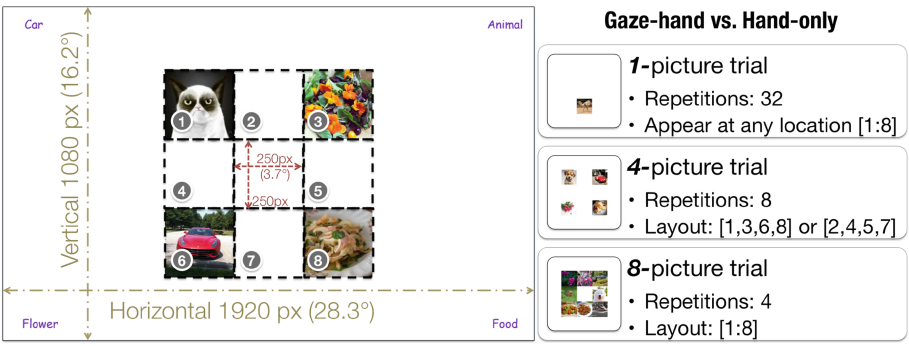


Fig. 1. User interface (left) and study design (right)

2.1 Task and Design

The photo-sorting task required repetitive selection followed by quick gesture commands on the selected items (see Fig. 1). The participants' task was to classify displayed photos into four categories labeled in diagonal corners (animal, car, flower and food), as accurately and as quickly as possible.

We used a 2 × 3 within-subjects design for factors *technique {gaze-hand, hand-only}* and *number of pictures presented {1, 4, 8}* at a time (Fig. 1). We varied the number of pictures presented in any trial reasoning that this might emphasize any differences between the two techniques: selecting amongst 8 possible targets instead of focusing on one at a time might show bigger advantages for the gaze-hand condition, for example.

In total, each session required a subject sorting 192 pictures, 96 for gaze-hand and 96 for hand-only. The order of presentation of the *technique* was counterbalanced and the order of the *number of pictures presented* was randomized among three blocks. Each block consisted of sorting 32 pictures displayed one, four or eight at a time.

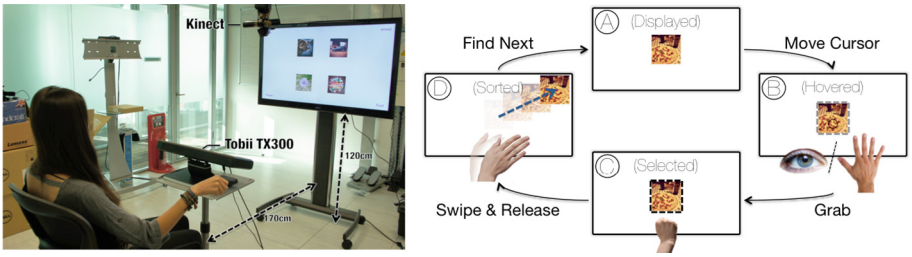


Fig. 2. Our setup: a user sat in front of a large screen using a combination of gaze and hand gestures for photo sorting. After acquiring a photo by either gaze or hand (B), users grabbed (C) and swiped it to diagonal directions (D).

2.2 Setup and Procedure

Participants and Apparatus. We recruited 15 participants (5 female), with a mean age of 26.7 years (SD = 4.5), who had little experience with eye tracking or motion and hand tracking applications. Participants sat on a chair with armrests at a distance of 2.4 m in front of a 55-inch (121 cm × 68.5 cm) display (Fig. 2). To capture gaze input, we used a Tobii TX300 at a sampling rate of 60 Hz. We positioned a Kinect sensor at a height of 1.8 m above the ground and 1.5 m in front of the user. For hand tracking, we used a fast random forest-based hand state classification method [10].

Procedure. After practicing the first technique encountered for approximately 5 min, participants were randomly presented with a set of 32 pictures, either 1, 4 or 8 pictures at a time (Fig. 2(A)). In the hand-only condition, the user's hand position was mapped to a screen cursor; in the gaze-hand condition, the user's gaze was mapped to a screen cursor for selection instead. To confirm to the participant which image was about to be

selected, the cursor-overlaid image was highlighted with a gray border after 150 ms (this is the "hover state", Fig. 2(B)). Thereafter, in both conditions, the user confirmed the selection by a hand grab gesture, indicated by the border color switching from gray to black ("select state", Fig. 2(C)). (Note that we deliberately decided against using gaze dwell time to confirm selection. Short dwell durations can induce the "Midas Touch" problem [2, 5, 11] where users can easily mis-select items when they only intend to look at an image, while long dwell durations can yield lower task performance and user satisfaction, as they are slow and disruptive.) The user then sorted the photo (in both conditions) by swiping their hand diagonally (Fig. 2(D)). We used four diagonal hand swipe directions to assign categories. Once a swipe was detected, the photo flew to one of the screen corners with swoosh sound effect.

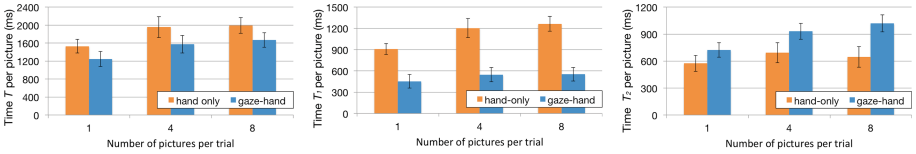
Participants filled out a questionnaire after each technique was completed. During the study, participants were allowed to take a short break after each experiment block. At the end of the entire session, we also conducted interviews to collect feedback. Each session lasted for approximately 1 h.

Dependent Variables. We logged three variables (Fig. 2): **T** the overall selection time, defined as the time between states A to C for single picture trials and from D to C for multiple picture trials; The time **T** can be further divided into: $\mathbf{T_1}$ the time from initial movement of the cursor, starting from picture being presented to confirmation of the hover state, and $\mathbf{T_2}$ the time between hover confirmation and selection using the grab gesture (B to C). Swipe time (from C to D) was assumed, and was indeed found to be, similar across conditions. In addition, sorting errors were logged.

3 Results

3.1 Overall Selection Time (T) per Picture

The time to select a picture was faster with gaze-hand than hand-only (shown by a significant main effect of technique on T, $F_{1,14} = 32.4$, $p < 0.0001$), being on average 0.33 s faster per picture (Fig. 3(A)). There was also a significant main effect of the number of pictures on T ($F_{1.3,18.25} = 28.8$, $p < 0.0001$), with the overall selection time increasing as the number of pictures increased. There was no interaction between technique and number of pictures on T ($F_{2,28} = 0.346$, $p = 0.71$).



A) Overall selection time T per picture B) The cursor moving time T_1 per picture C) The grabbing time T_2 per picture

Fig. 3. Different average performance time for each technique and number of pictures presented at a time (mean ± 95 % CI).

3.2 Cursor Moving Time (T_1) and Grabbing Time (T_2) per Picture

Despite the overall difference in T or selection time, it is more instructive to break this down into two components.

Time to move the cursor to a hover state (T_1) was significantly faster (by on average 0.61 s) with gaze-hand than with hand-only ($F_{1,14} = 308.9$, $p < 0.0001$). There was also a significant main effect of the number of pictures on T_1 ($F_{1.39,19.39} = 14.18$, $p < 0.001$) but as Fig. 3(B) shows, there was also a significant interaction between technique and number of pictures ($F_{2,28} = 5.99$, $p < 0.01$) which suggests this effect is due an increase in T_1 as the number of pictures increased in the hand-only technique.

When we look at T_2, the time from hover to select, we see a very different story. Here the hand-only technique was faster than gaze-hand ($F_{1,14} = 47.33$, $p < 0.0001$), being on average 0.25 s faster per picture (Fig. 3(C)). Number of pictures again gives rise to a main effect ($F_{2,28} = 21.87$, $p < 0.0001$) but again it is made more difficult to interpret due to a significant technique by number of pictures interaction ($F_{2,28} = 6.06$, $p < 0.01$). This indicates that this time it is the gaze-hand condition that accounts for this increase and not the hand-only condition.

3.3 Sorting Accuracy and Error Analysis

We analyzed the sorting accuracy per block ($N = 32$) (see Table 1). The success rate is generally high for all trials. Error analysis revealed that users made significantly more mistakes in the 8-picture trials in the gaze-hand condition than hand-only ($t(14) = 2.902$, $p < 0.012$).

Table 1. Average sorting accuracy (mean ± std) of the 15 participants made for each block.

	Gaze-hand	Hand-only	Paired samples t-test
One	92.1 % ± 6.0 %	93.3 % ± 3.9 %	$t(14) = 1.000$, $p = .334$
Four	91.7 % ± 5.7 %	93.3 % ± 4.1 %	$t(14) = 1.035$, $p = .318$
Eight	90.0 % ± 7.4 %	95.4 % ± 3.1 %	$t(14) = 2.902$, $p = .012$

There were many kinds of errors. Most were incorrectly sorted images due to swiping in the wrong direction (categorization errors). Some were due to the way the system was configured, or the peculiarities of the interaction. For example, this included triggering a swipe response when returning the hand to the resting position or not completely releasing a grabbed picture.

However, the lower accuracy in the 8-picture condition of the gaze-hand technique is probably explained by difficulty in coordinating gaze and gesture. For example, in the 4- and 8-picture trials, participants occasionally performed a swipe as their eyes prematurely moved ahead to the next image - the eyes "jumping the gun" before the hand had finished its work, which is similar to the synchronization problem identified in prior work [11]. Further, the logged gaze data revealed that neighboring pictures in the gaze-hand condition, largely in 8-picture trials, distracted some participants. During selection, their gaze shifted between the selected and the adjacent image, resulting in incorrectly sorting the wrong object. Participants' feedback in post-task interviews

suggested that both of these kinds of errors occurred when they tried to speed up their performance, which in turn caused them to slow down to be more careful.

3.4 User Experience and Preferences

The questionnaire data indicate only two significant differences across techniques. First, hand fatigue was rated significantly higher with the hand-only technique than gaze-hand ($Z = –2.751$, $p < 0.01$). Most participants (10/15) commented on this. For example, one said, *"holding the arm high in mid-air is tiring"* and *"[with gaze] I could rest my hand and use my hand only for sorting"*.

Second, participants felt that the hand-only technique was easier to learn than the gaze-hand technique ($Z = −2.309$, $p < 0.05$). Though participants were positive about the combination of gaze for selection and hand for grasping and throwing, they felt the gaze-hand combination was something that required time to get used to. Despite this, when we tested for learning effects over trials, we found no evidence of differences between gaze-hand and hand-only techniques.

Finally, participants rated which technique they preferred and the perceived speed and accuracy in the sorting task (Fig. 4). Thirteen users preferred the gaze-hand technique: although they found both techniques *"intuitive"*, some felt that the gaze-hand technique was *"less demanding"* (4/15) and *"easier to use"* (10/15), primarily due to less physical fatigue and fast and accurate gaze selection.

Further explanation for this can be found in the interview data. For example, one participant said, *"I am pretty confident with the gaze to select pictures, as what I looked at was what I wanted to select"*. In contrast, users felt that positioning the cursor on top of a picture using the hand was trickier, as it required constant checking and it was difficult to control the cursor accurately on a 2D screen while moving their hands freely in 3D space.

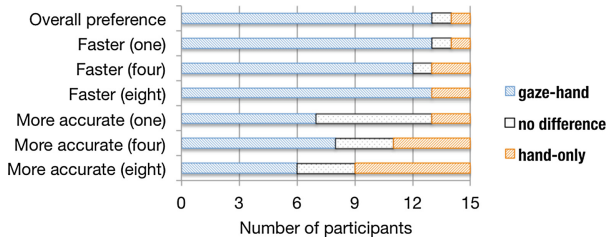


Fig. 4. Participants' technique preference, perceived speed, and accuracy in sorting as rated at the end of the experiment.

4 Discussion

4.1 The Benefits of Combining Gaze and Hand Gestures

Our results show improvements in both speed and user experience when we add gaze to hand gesture input in this remote interaction task. Previous work showed that gaze is

faster than the hand for pointing and positioning a single target [8]. Our study expands on this showing that the speed advantage of gaze persists when used for target selection and when combined with gesture in a single technique. More than this, it persists in the context of multiple targets and indeed becomes more pronounced as the number of targets increases. This was shown in our analysis of T_1, the time to move the cursor, which increased significantly for the hand-only technique with more targets, but which did not increase when gaze directed the movement of the cursor. This explains the overall speed advantage for the task of image selection.

The majority of participants also perceived the gaze-hand technique as faster than hand-only, and users' subjective feedback further confirmed the fact that gaze complements gestural interaction when it comes to indicating objects of interest on a large screen. Here, users explained that hand gestures alone were slow, tiring, and inaccurate compared to incorporating gaze into the technique.

Taken together, this suggests that the combination of gaze and hand gestures is particularly well suited for applications that involve repetitive manipulation of multiple objects. Here, gaze selection can be both faster and reduce hand fatigue, especially when the task requires frequent pointing. Of course, arm fatigue could be even further reduced by attending to the ergonomics of the hand gestures too, such as designing them to require minimal effort, providing arm support and so on.

4.2 The Costs of Combining Gaze and Hand Gestures

Although participants indicated that they preferred the gaze-hand condition and reported little trouble with eye-hand coordination, the performance data (T_2) shows that it actually took longer to switch from gaze-hover to hand-select, than it did to transition from hand-hover to hand-select. Further, this effect was made worse as the number of targets to choose from increased. This may be because multiple targets increased visual distraction, delaying the ability for participants to confirm that the hovered object was indeed the intended target. This was not the case with selection within modality. In other words, there is a cross-modality speed of performance cost here, which may only be amplified as the interface becomes more complex.

Added to this, the analysis of errors provides some indication of accuracy issues induced by the mixed modality interaction. Sometimes it was clear that participants were visually distracted by adjacent pictures, leading to errors in selection. Other times, especially in the 8-picture condition, participants sometimes moved their gaze too quickly in advance of completing a gesture, causing errors. Participants were very aware of these synchronization errors and said that this was further amplified when they wanted to be fast. In other words, while gaze was often a fast way to reach the next target, the cost was that users had to wait for the hand to finish the previous manipulation. Solving this problem through good design starts by recognizing this problem.

Participants' feedback also suggests that working across modality was more difficult to learn than the single modality technique. While we might expect that these cross-modality costs would disappear with practice, we found no obvious learning effects during the course of the study. It may be that longer-term use is needed to investigate these aspects.

However, it is notable that despite the costs we outline above, participants said that overall they preferred the gaze-hand technique and expected it would be *"more productive"* in the long run. So clearly, these judgments are weighing up the benefits against the costs we have highlighted.

5 Conclusion

In this paper, we investigated both the costs and benefits of combining gaze and hand gestures for remote interaction. Our work contributes to a better understanding of how users perceive this input combination, and shows that gaze can complement free-hand gestural interfaces if the task is designed appropriately. Despite the costs, users mainly prefer the combination of gaze and gesture, but designers must take account of the kinds of errors that users can make and expect that such techniques may be initially perceived as more difficult to learn.

References

1. Vogel, D., Balakrishnan, R.: Distant freehand pointing and clicking on very large, high resolution displays. In: Proceedings of UIST 2005, pp. 33–42. ACM (2005)
2. Sibert, L.E., Jacob, R.J.K.: Evaluation of eye gaze interaction. In: Proceedings of CHI 2000, pp. 281–288. ACM (2000)
3. Zhai, S., Morimoto, C., Ihde, S.: Manual and gaze input cascaded (MAGIC) pointing. In: Proceedings of CHI 1999, pp. 246–253. ACM (1999)
4. Zhang, Y., Bulling, A., Gellersen, H.: SideWays: a gaze interface for spontaneous interaction with situated displays. In: Proceedings of CHI 2013, pp. 851–860. ACM (2013)
5. Stellmach, S., Dachselt, R.: Look and touch: gaze-supported target acquisition. In: Proceedings of CHI 2012, pp. 2981–2990. ACM (2012)
6. Turner, J., Alexander, J., Bulling, A., Schmidt, D., Gellersen, H.: Eye pull, eye push: moving objects between large screens and personal devices with gaze and touch. In: Kotzé, P., Marsden, G., Lindgaard, G., Wesson, J., Winckler, M. (eds.) INTERACT 2013, Part II. LNCS, vol. 8118, pp. 170–186. Springer, Heidelberg (2013)
7. Pelz, J., Hayhoe, M., Loeber, R.: The coordination of eye, head, and hand movements in a natural task. Exp. Brain Res. **139**, 266–277 (2001)
8. Kosunen, I., Jylha, A., Ahmed, I., An, C., Chech, L., Gamberini, L., Cavazza, M., Jacucci, G.: Comparing eye and gesture pointing to drag items on large screens. In: Proceedings of ITS 2013, pp. 425–428. ACM (2013)
9. Yoo, B., Han, J-J., Choi, C., Yi, K., Suh, S., Park, D., Kim, C.: 3D user interface combining gaze and hand gestures for large-scale display. In: Proceedings of EA CHI 2010, pp. 3709–3714. ACM (2010)
10. Keskin, C., Kıraç, F., Kara, Y.E., Akarun, L.: Hand pose estimation and hand shape classification using multi-layered randomized decision forests. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 852–863. Springer, Heidelberg (2012)
11. Kumar, M., Paepcke, A., Winograd, T.: EyePoint: practical pointing and selection using gaze and keyboard. In: Proceedings of CHI 2007, pp. 421–430. ACM (2007)