Interacting with Communication Appliances: An evaluation of two computer vision-based selection techniques

Jacob Eisenstein [1]
Massachusetts Institute of Technology
jacobe@gmail.com

Wendy E. Mackay [2]
INRIA Futurs
mackay@lri.fr

ABSTRACT

Communication appliances, intended for home settings, require intuitive forms of interaction. Computer vision offers a potential solution, but is not yet sufficiently accurate.As interaction designers, we need to know more than the absolute accuracy of such techniques: we must also be able to compare how they will work in our design settings, especially if we allow users to collaborate in the interpretation of their actions. We conducted a 2x4 within-subjects experiment to compare two interaction techniques based on computer vision: *motion sensing*, with EyeToy®-like feedback, and *object tracking*. Both techniques were 100% accurate with 2 or 5 choices. With 21 choices, *object-tracking* had significantly fewer errors and took less time for an accurate selection. Participants' subjective preferences were divided equally between the two techniques. This study compares these techniques as they would be used in real-world applications, with integrated user feedback, allowing interface designers to choose the one that best suits the specific user requirements for their particular application.

Author Keywords

Communication appliance, domestic technology, EyeToy®, home setting, object tracking, MirrorSpace, motion sensing

ACM Classification Keywords

D.2.2 [Software Engineering]: Design Tools & Techniques, User interfaces, H.1.2 [Models & Principles]: User/Machine Systems Human factors, H.5.3 [Group and Organization Interfaces]: Collaborative computing

INTRODUCTION

We are exploring a particular category of interactive systems that we call *communication appliances*, a form of ambient computing that merges communication and computation capabilities into the physical environment in order to seamlessly integrate remote communication into daily activities [7].

CHI 2006, April 22-27, 2006, Montréal, Québec, Canada.

We define *communication appliances* as simple-to-use, single-function devices that let people communicate, passively or actively, via some medium, with one or more remotely-located friends or family. The goal is to retain simplicity from the user's perspective while providing an object that is robust, enjoyable and aesthetically pleasing. In order to explore the design space, we created a series of *technology probes* [4], each of which offered a single form of interaction for sharing a specific type of data (photos, video, handwriting or sound) via a dedicated 'always on' communication channel among close family members. For example, *MirrorSpace* [9] looks like a mirror on the wall (Figure 1, left), but also acts as an exclusive link between two households, superimposing live images from each. To preserve privacy, images are fuzzy, becoming progressively more distinct as a person approaches the mirror. MirrorSpace provides participants with shared background awareness of each other [3] and a smooth transition between peripheral and focused communication.

Figure 1. Our MirrorSpace and Sony's EyeToy®

Our field studies showed that moving in space works well for controlling privacy, but we need additional controls to manage directed communication. Computer vision offers a possible solution [2], especially for applications that already use a camera. Tracking specified objects or body movements removes the need for additional input devices and provides opportunities for more natural or playful forms of interaction. Unfortunately, computer vision is still not as accurate or robust as users may expect. Computer-vision researchers, understandably, concentrate on technical improvements to their hardware and software. As interaction designers, we need to consider both technical and experiential aspects and make appropriate design tradeoffs based on user require-

[1] Visiting researcher sponsored by INRIA Futurs, In | Situ | group.
[2] In | Situ | group – Pôle Commun de Recherche en Informatique du Plateau de Saclay – CNRS, Ecole Polytechnique, INRIA, Université Paris-Sud.

ments. Our challenge is to design the interface so that the user can collaborate effectively with the system: we create a visual affordance [8] that shows the user the system's best guess and its confidence in that guess. Ideally, the user's instinctive response to that feedback will improve the system's accuracy. This adaptive behaviour keeps the user in control, creating a more robust and satisfying experience, while improving overall system performance.

Artists have been exploring interactions between users and their images for decades. Krueger, an early pioneer, created VideoPlace and other installations in the 1970s. For example, a user sees a live projected silhouette of his body and an animated 'critter' [5] which crawls around the silhouette, reaching for the highest body part. Over the years, computer vision techniques have been refined for diverse art installations [10] and the technology is now sufficiently robust to consider such applications for the home.

Figure 1 (right) shows a commercially successful example: Sony's EyeToy®. Here, the user sees her own image with superimposed objects. When she waves, the object reacts. The feedback is designed so that, if the system has trouble interpreting her gestures, her instinctive reactions provide the information necessary for the system to improve itsperformance. If detection is slow, she waves her hand more quickly, which helps the system detect her movement. She instinctively adapts her interaction to the recognition properties of the system, resulting in a robust, fun interaction.

As *communication appliance* designers, we could simply adapt the EyeToy® motion detection technique: it is easily-understood and robust in every-day use. However the MirrorSpace is intended for home settings with much ambient movement. We need to ensure that users' selections are intentional, without disrupting the shared awareness aspects of the system. Our fieldwork suggested another alternative. Most families installed their communication appliances on a table with aesthetically pleasing, funny or sentimental objects [6]. We wanted to test whether using *object tracking*, i.e. a vision system that recognizes and tracks pre-identified objects in real time, would be as effective as *motion sensing* in the home environment. We hypothesized that object tracking could be made robust, using feedback in a similar way to motion sensing.

This paper describes a 2x4 within-subjects experiment that compared two feedback-enhanced computer vision techniques, *motion sensing* and *object-tracking*. In addition to speed, accuracy and quality judgements, we wanted to identify the threshold after which they begin to breakdown: *2, 5, 11* or *21* items. We describe our implementation followed by the experiment and its results. We conclude with recommendations for designers who want to incorporate feedback-enhanced computer vision techniques into *communication appliances* and similar ambient interfaces.

SELECTION TECHNIQUES

Motion Sensing is modeled after the Sony EyeToy®. The user waves over the desired button which "fills up" to show how much motion is being detected. The selection is triggered when a threshold is reached. We implemented this technique using optical flow detection instead of the simpler image-differencing technique used by the EyeToy®. We felt this would improve accuracy and make a fairer comparison with the object tracking technique [11].

Object Tracking requires the user to place a coloured physical ball over the target button displayed on the screen. The ball is tracked using the camshift algorithm [1], with the backprojection computed by a set of histograms at several levels of precision in the YCrCb color space. The system displays an ellipse on the screen to show its best guess for the location of the tracked ball. The tracker assesses its level of confidence in its own estimate, using goodness-of-fit metrics based on the size, shape, and color of the estimated location and boundary of the tracked object, and reflects this via the size and color of the ellipse. The ellipse is yellow when certainty is low and turns green when certainty is high. If the object tracking system takes longer than a few seconds, users perceive a problem and intuitively hold the ball still, which helps the tracker to recover. Note that users have the opposite intuitive response when the motion sensing system is confused: they then move faster and more precisely.

EXPERIMENT

We conducted a 2x4 within-subjects repeated measures experiment to compare two computer-vision-based selection techniques under varied menu sizes. Each participant was exposed to all conditions, counter-balanced for order. We measured speed and accuracy with respect to selection type: *motion sensing* and *object tracking* and menu size: 2, 5, 11 and 21 items. We also interviewed participants after the experiment to learn their subjective impressions of each technique under different conditions.

Subjects: 12 people (10 men and 2 women in their 20s and 30s), participated in the study. None were familiar with either selection technique and none were paid.

Apparatus: Both selection techniques, *motion sensing* and *object tracking*, were implemented on a Macintosh G5 with an iSight camera and a 20-inch widescreen display set at 1680 by 1050 resolution. Both were implemented with the OpenCV Open Source Computer Vision Library [1] and Roussel's Núcleo video toolkit [2].

Procedure: Participants sat in front of a screen with a live image of themselves. Conditions were controlled with an automated script and instructions were overlaid onto the video display. Participants started each trial with their hands centred on their laps. When the target button began to blink, participants would select it as quickly and accurately as possible and then return their hands to the starting position. To prevent fatigue, the system enforced a rest period after each block of 20 trials and participants chose when to start the subsequent block.

[1] http://www.intel.com/research/mrl/research
[2] http://insitu.lri.fr/ roussel/projects/nucleo/

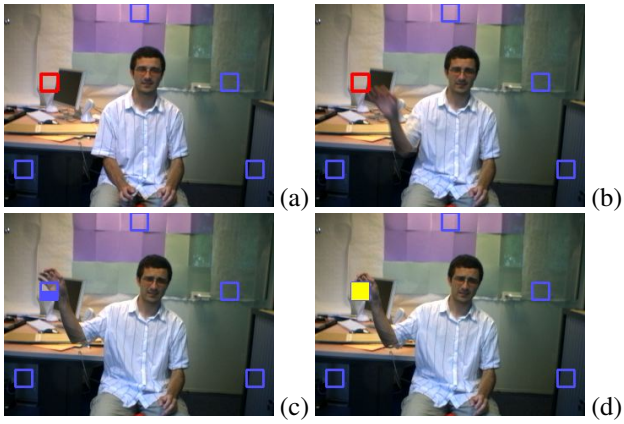

Figure 2. Motion sensing condition, 5 items

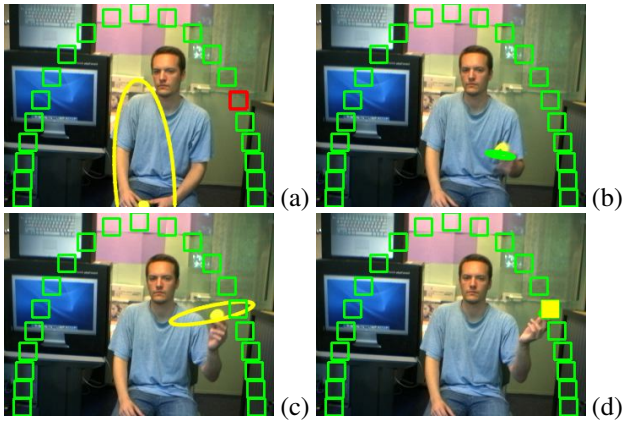

Figure 3. Object tracking condition, 21 items

The training phase allowed participants to experiment with both selection techniques under the various menu conditions. When they felt comfortable with both techniques, participants indicated that they were ready to begin.

The experimental phase consisted of six blocks of 20 trials each. Each block presented one technique (*motion* or *object*) with five presentations of each of the four menu sizes (2, 5, 11 and 21 buttons). Buttons were laid out in a semi ellipse such that no target item would be closer than 15 pixels from the edge of the screen. The angle between the center of the ellipse and the center of each button and the distance from the starting position to the button were held constant for each correct choice. Note that spacing between the buttons was *not* constant and for displays with 11 or 21 buttons, side buttons were closer than top buttons (Figure 3).

We counterbalanced the order of menu sizes and distributed correct answers for each menu size equally across both selection types and across all potential target positions. This ensured that the potential for increased errors was equally distributed across conditions. (Figure 1 shows the five-button display and figure 2 shows the 21-button display.) We alternated the *motion* and *object* blocks; half of the participants began with *motion* and the other half began with *object*). Participants were exposed to each selection technique, with each of four menu sizes, 15 times over the course of the experiment (a within-subjects repeated-measures design), for a total of 720 trials for each selection technique. Quantitative measures included selection and menu type, target button, actual selection, and start and stop times. We also calculated elapsed time and errors.[3] Qualitative measures included perceptions of speed and errors and user satisfaction with each technique.

Figure 2 shows the 5-item *motion* condition: a) target button flashes b) user moves hand towards button c) user waves hand over red button and activation level rises d) activation level reaches threshold and button is selected.

[3]We asked participants to try to attain a target error rate of 4 %. The actual mean error was 4.5%.

Figure 3 shows the 21-item *tracking* condition: a) system is unsure of location; displays a large yellow ellipse b) user moves ball toward target; tracker identifies ball, ellipse turns green. c) system is momentarily confused as object is moved too quickly; ellipse turns yellow. d) target button is selected immediately when tracked object is superimposed.

RESULTS

We used a two-factor analysis of variance to analyse the effect of selection type and menu size on errors and selection time (alpha = 0.05). We counted an error when the button selected differed from the target. Of 1440 possible trials, participants made only 65 errors (4.5%), a highly accurate result given the inaccuracy of these techniques without user feedback.

Errors: A small learning effect obtained: block 1 had the most errors (19 or 29%) followed by block 2 (15 or 23%). All errors in the 11-item condition occurred in the first two blocks and the first two blocks had significantly more errors than the final three blocks ($p < 0.05$). Total errors for individual participants ranged from 2 (3%) to 13 (20%), with a mean of 5.3 (8%). All participants were 100% accurate for the 2-item and 5-item trials. The 11-item condition had a total of 8 (12%) errors, with the remaining 57 (88%) errors in the 21-item condition. Two-thirds of errors (44 or 68%) occurred with the *motion* (EyeToy-like) technique. The *motion-21 item* condition was significantly different from the *object-21 item* condition, which were significantly different from all the remaining conditions (both *motion* and *object* with 11, 5 and 2 items), $p < 0.05$. In summary, the *object* condition was significantly more accurate than the *motion* condition, and errors increased significantly with more items, with most errors in the 21-item condition.

Speed: Participants were significantly faster with the *object* technique (mean = 1.63") than *motion* technique (mean=1.94"). Although average selection time increased monotonically with the number of buttons, this effect was not significant. The *object* trials were significantly faster than the *motion* trials (mean = 1517.1 vs. mean = 1755.3 milliseconds), regardless of number of items, and the 21-item trials were sig-

	motion	object	total
2-item	0 (0%)	0 (0%)	0 (0%)
5-item	0 (0%)	0 (0%)	0 (0%)
11-item	7 (11%)	1 (1%)	8 (12%)
21-item	37 (57%)	20 (31%)	57 (88%)
total	44 (68%)	21 (32%)	65 (100%)

Table 1. Errors by selection technique and # items

	motion	object	no pref
Which is faster?	1	10	1
Which is more accurate?	3	9	0
Which did you prefer?	5	6	1

Table 2. Qualitative results.

nificantly slower than the rest ($p < 0.05$). When we examined the slowest trials (> 3000 milliseconds), we found that only 1 of 26 (3.8%) were errors, compared to the overall error rate of 4.5%. Note that participants placed their hands in the same location before each trial to ensure comparability across conditions and timing measures. However, this constraint is not necessary in home settings. With the *motion* technique, if the user makes a series of choices, the system requires the same level of user movement to detect each new choice. In contrast, with the *object* technique, once the system recognises the object, subsequent choices should be faster. Thus, in real-world settings, the *object* technique is likely to be even more rapid than the *motion* technique.

Qualitative impressions: Most participants reported that the *object* condition was both faster and more accurate (see Table 2), but were almost evenly divided as to which they preferred. Of the participants who preferred the *motion* condition, two said they liked the fact that gradual feedback enabled them to see if they were about to make a mistake.

DISCUSSION

Designers of technologies for the home must select solutions that meet specific user requirements. Although some computer vision techniques may be relevant, they are difficult to assess, since most are evaluated only in terms of their absolute accuracy. This paper compares two computer vision techniques, taking the user-feedback loop into account, which improves accuracy and user enjoyment.

We studied *motion sensing* similar to the commercially successful EyeToy®, which acted as a control for *object tracking* which we find more appropriate for *communication appliances* such as MirrorSpace. Although both techniques were highly accurate, the *object* technique was faster and more accurate than the *motion* technique, and was better able to handle large numbers of alternative choices. Both techniques were 100% accurate with a few menu items, but the *motion* technique broke down sooner, as the number of items increased. Participants were roughly evenly divided in their subjective preferences, with half preferring the *object tracking* technique.

Both techniques have strengths and weaknesses. *Object tracking* is robust to background movement, but was overly sensitive to lighting changes in our implementation. Requiring a separate object emphasizes intentional interaction and is less susceptible to accidental use. However, objects can be lost, which is why we chose an inexpensive, easy-to-replace colored ball. *Motion sensing* does not require a separate object, but is overly sensitive to background motion, which can confuse the system, or worse, result in unintended selections. We plan to add *object tracking* to a home version of MirrorSpace and will evaluate other computer vision techniques, e.g. direction of motion, in this context.

We argue that assessments of computer vision techniques should include cases with feedback to the user, since when users collaborate with the system, interpretation is more accurate and robust, and the resulting user experience is improved. This study also suggests that *object tracking* is as good or better than the commercially successful *motion sensing* approach and is worth exploring in home settings.

ACKNOWLEDGMENTS

Thanks to Nicolas Roussel, for help with the Nucleo toolkit and Michel Beaudouin-Lafon for advice on the paper.

REFERENCES

1. G. Bradski. Computer vision face tracking for use in a perceptual user interface. *Intel Tech. J.*, Q2, 1998.
2. W. Freeman, P. Beardsley, H. Kage, K. Tanaka, K. Kyuma, and C. Weissman. Computer vision for computer interaction. *SIGGRAPH Comput. Graph.*, 33(4):65–68, 2000.
3. C. Heath and P. Luff. Collaboration and control: Crisis management and multimedia technology in london underground line control rooms. *J. CSCW*, 1(1):24–48, 1992.
4. H. Hutchinson, W. Mackay, B. Westerlund, B. Bederson, A. Druin, C. Plaisant, and M. Beaudouin Lafon. Technology probes: Inspiring design for and with families. *CHI'03 Human Factors in Computing Systems*, pages 17–24, 2003.
5. M. Krueger. *Artificial Reality 2*. Addison-Wesley, 1991.
6. W. Mackay. The interactive thread: Exploring research methods for multi-disciplinary design. *DIS'04*, 2004.
7. B. V. Niman, D. Thanh, J. Herstad, and H. Hüttenrauch. Transparent communication appliances. In *HCI International '99*, pages 18–22, NJ, 1999. Erlbaum.
8. D. Norman. *The Invisible Computer*. MIT Press, 1999.
9. N. Roussel, H. Evans, and H. H. Mirrorspace: using proximity as an interface to video-mediated communication. *INRIA Technical report*, 2004.
10. J. Truckenbrod. Touchware gallery, 1998. ACM SIGGRAPH.
11. Z.Zivkovic. Optical-flow-driven gadgets for gaming user interface. In *Proceedings of the 3rd International Conference on Entertainment Computing*, 2004.