Remote Gaze and Gesture Tracking on the Microsoft Kinect: Investigating the Role of Feedback

Marcus Carter[1], Joshua Newn[1], Eduardo Velloso[2] & Frank Vetere[1]

Microsoft Research Centre for Social NUI[1]
Computing and Information Systems
The University of Melbourne

Embedded Interactive Systems Group[2]
School of Computing and Communications
Lancaster University

marcusc@unimelb.edu.au, joshua.newn@unimelb.edu.au, e.velloso@lancaster.ac.uk, fvetere@unimelb.edu.au

ABSTRACT

In this paper we present the results of a user experience and preference study into the combination of gaze and gesture in a lounge-style remote-interaction, using a novel system that tracks gaze and gesture using only the Kinect device at a distance of 2m from the user. Our results indicate exciting opportunities for gaze-tracking interfaces that use existing technologies, but suggest that findings from studies of highly-accurate gaze systems may not apply in these real-world simulations where the gaze-tracking is inherently less accurate. We contribute a series of design recommendations for gaze and gesture interfaces in this context, and based on these limitations.

Author Keywords

Gaze-Tracking, gesture tracking, Kinect.

ACM Classification Keywords

H.5.2 User Interfaces: Evaluation/methodology.

INTRODUCTION

Touchless, gestural interfaces are becoming increasingly common, and are being adopted in a wide variety of commercial contexts where remote interaction offers an improved user experience. The Microsoft Kinect is the best known device, having recently been coupled with the Xbox One game console, with 8 million devices sold. Aside from gameplay, the technology has been demonstrated to improve user experience in contexts such as hospitals (O'Hara et al., 2014) and public spaces (Walter, Bailly & Muller, 2013).

The use of and interest in eye-tracking technologies is similarly expanding, with numerous low-cost, reliable eye-trackers entering the market. Whereas eye-trackers sufficiently accurate for health-applications frequently cost in excess of $20,000, affordable trackers such as the Tobii eyeX (~$170) and EyeTribe (~$150)—increasingly marketed as gaming devices—are prompting a spike in interest in this interaction technology. However, these technologies are all limited in that they need to be within

OzCHI '15, December 07 - 10 2015, Melbourne, VIC, Australia

ACM 978-1-4503-3673-4/15/12…$15.00
DOI: http://dx.doi.org/10.1145/2838739.2838778.

Figure 1: A user playing *eyeSheep*, a game we developed to explore the combination of gaze and gesture. This animated figure is best viewed in Adobe Reader.

1m of the user in order to function, precluding console gaming or large public screen use, and require dedicated hardware. This also limits the combination of gestural interaction with gaze-traction, which recent studies have shown to be an ideal combination (see Velloso et al., 2015)

We have developed a system that integrates a 2D-image based eye-tracker (developed by xLabs) into the Microsoft Kinect, allowing for the combination of gaze and gestural interaction at distance of 2m from the screen, such as in a domestic lounge. This system opens up the opportunities for gaze interaction in new contexts, drawing on existing and inexpensive hardware.

However, gaze tracking with RGB cameras is fundamentally less accurate than infra-red tracking. Further, tracking in a living room setting exacerbates these inaccuracies due to the user's distance from the screen. Consequently, design guidelines developed in existing gaze research (which use highly accurate gaze systems, predominantly in desktop-use systems) do not necessarily hold up in large screen, at-a-distance interactions where the gaze is less accurate.

We contend that it is necessary to re-evaluate user interface design assumptions in this new gaze-interaction context. Consequently, in this paper we present the results of a user experience and preference study into the combination of gaze and gesture in a lounge-style interaction with less-accurate gaze tracking, where we interrogated two different versions of gaze-visualisation. Participants played two versions of *eyeSheep* (figure 1), a simple game that we developed that exemplified the interaction technique of using gaze to select an on-screen object, and gesture to manipulate it.

We found that visualising a less-accurate gaze-point significantly detracts from user experience, even in complex tasks where a visualised gaze-point could be expected to assist users in 'correcting' the tracking. Rather than providing feedback about the system's estimation of the user's gaze, we recommend only visualising the system's interpretation of the user's intent, which was preferred and improved user's perception of how accurate the system was.

Our results exemplify the immediate opportunity for gaze and gesture remote interaction, support continued research into designing interfaces and interaction techniques informed by non-perfect gaze estimation, and suggest that more research into the different usability and design of these systems is warranted. Along with contributing numerous implications for design based on our participant's 'natural' attempts to interact with the system, we argue that gesture and gaze are not independent modalities and their combination can be more effectively harnessed in gaze and gesture systems.

LITERATURE REVIEW

In this section we overview the relevant literature on gaze, gaze visualisation, gesture and their combination, situating our study in earlier work.

Gaze

Eye tracking has been an active topic in HCI since the early 80's (Bolt, 1981). Applications for eye tracking fall into one of two categories: diagnostic or interactive (Duchowski, 2007). Diagnostic applications analyse users' natural gaze behaviour to obtain insights on their cognitive processes (Yarbus, 1967) or to evaluate interface designs (Bojko, 2013). In interactive applications, systems use the eye tracker as an input device, either for direct control (e.g. pointing (Ware and Mikaelian, 1987)) or for implicit interaction (e.g. attention-aware interfaces (Vertegaal, 2003)).

In this work we are interested in using eye tracking for interaction with large displays, such as in public settings or in a living room. However, tracking gaze at long distances is difficult, as even the most expensive eye trackers have a tracking range of less than 1m. To address this problem, most researchers have used wearable eye trackers or placed the tracker closer to the user than to the display (Lander, 2015; Turner, 2013; Turner 2015). To extend the range of eye trackers, one approach is to track the user's head with a separate device (such as a Kinect (Henessey, 2012) or multiple wide-view cameras (Cho, 2012; Cho, 2013)) and face the eye tracker towards to the user's eyes with a servo-motor. Whereas these approaches successfully enable gaze interaction at longer distances, they require additional and expensive hardware. In this work, we explore how to do so using only the wide-view camera already contained in the Microsoft Kinect, enabling simultaneously eye and body tracking.

Video-Based Eye Tracking

The main advantage of video-based eye tracking—as compared to more common approaches such as infra-red pupil-corneal reflection or EOG tracking—is how cheap and easy it is to deploy, since it can work with conventional webcams. However, it requires an unobstructed view of the eyes and the tracking quality can be affected by the lighting conditions, reflections from glasses, droopy eyelids, squinting eyelids, and heavy makeup (Majaranta, 2014). In general, webcam eye tracking, though more convenient to deploy to a general audience, offers a visual angle error in the order of 2-4 degrees, substantially higher than the error in the order of 0.5 degree that dedicated trackers provide (Sesma, 2012). Examples of works that take this approach include a webcam mounted on a hat for experiments with children (Noris, 2008) and a calibration-free content scrolling interface for public displays (Zhang, 2013).

With the goal of providing web analytics using consumer hardware, many companies provide webcam-based eye tracking as a web service, including *GazeHawk* (recently acquired by Facebook), *Sticky*, and *xLabs*—the tracking algorithm we chose for this experiment.

Gaze Visualisation

There are several ways of displaying gaze information to the user. *Scanpaths* are linearly connected sequences of points representing the coordinates of the gaze point, often aggregated by fixation time (e.g. larger points representing longer fixations). *Heatmaps* display an aggregate visualisation of the distribution of users' visual attention. Whereas scanpaths are better suited to display temporal information, heatmaps better display spatial information and data from multiple users. Even though these are the two most predominant approaches for visualising gaze data for visual attention analysis (Duchowski, 2012), they are not well suited for real-time cursor control.

Few studies have examined the impact of cursor visualisation for direct gaze interaction. This scenario presents three significant challenges (Zhang, 2011). First, the jittery movement of the eyes makes the cursor movement appear to be noisy. Second, device and algorithmic accuracy issues can add error to the gaze point estimation. Third, the cursor itself can draw users' attention, creating a positive feedback loop in which their attention tends to drift further away from the desired location (Jacob, 1993).

The most basic gaze visualisation technique for real-time interaction is displaying the most recent raw or lightly filtered data from the tracker. To account for the jittery movement of the eyes, it is common to display several recent data points, rather than a single one. To account for the visual angle error, instead of a point, another possibility is to use a larger area, in the form of a circle, a spotlight, or a fisheye lens that increases target sizes close to the gaze point (Ashmore, 2005). A more subtle visualisation technique is instead of using a continuous visualisation, highlighting the currently selected object.

Seifert compared three visualisations in a fast selection task: a continuous cursor, highlighting the selected target and no feedback (Seifert, 2002). She found no significant differences in performance or perceived workload, but the condition with no feedback yielded shorter reaction times, fewer false activations, and fewer target misses. Participants preferred the condition with no feedback.

Majaranta et al. compared different feedback techniques for dwell-based typing: speech, 1-level selection (red background on the selected key) and 2-level selection (highlight the key under the gaze point, and red background after the dwell time) (Majaranta, 2004). The two visual feedback techniques outperformed the speech technique, but there was no significant difference between them. However, participants found the 1-level feedback less confusing than the 2-level feedback. In a separate study, Majaranta et al. compared other feedback modalities for the same task: visual feedback, audio click + visual, speech + visual, and speech only (Majaranta, 2006). In this study they found that the visual feedback combined with the audio click yielded the best typing speed, and was the technique most participants preferred. These authors recommend confirming selection with a non-speech sound, combining speech with visual feedback, using one-level feedback with short dwell times, making focus and selection distinguishable in 2-level feedback, using animation to support focus with long dwell times, and allowing feedback parameter adjustment.

Combining Gaze and Gesture

Early work combining gaze and gesture dates back to the early 90's (Koons, 1993), and since then several works have investigated the topic. In the 3D domain, Song et al. built a CAD system that uses gaze to select 3D objects and mid-air gestures to manipulate them (Song, 2013), a concept that Velloso et al. empirically showed to outperform other hand pointing techniques (Velloso, 2015). Velloso et al. also found that even when object selection is performed by (high accuracy) gaze, users still reach out in the general direction of the object in order to manipulate it (Velloso, 2015). In a gaming context, Arcade+ is the prototype of an arcade machine that combines eye tracking with overhead finger gesture tracking (Velloso, 2015-2)

Yoo et al. proposed a system that combines hand gestures with head orientation as an approximation for the gaze point (Yoo, 2010), however it has been reported that few users naturally align their heads with their gaze (Mubin, 2009). Kosunen et al. compared hand and gaze selection in a similar task to ours, and found the eyes to be faster and more accurate (Kosunen, 2013). Cha and Meier proposed a multi-display system controlled by a combination of a wearable eye tracker with discrete hand gestures, but did not evaluate it (Cha, 2012).

Summary

In summary, gaze has been successfully used in HCI for many decades, but due to the high cost of trackers, gaze interaction has remained confined to research labs. As eye trackers finally become affordable, there has been increased interest in expanding their potential applications to other areas, such as gaming; and combining gaze with other modalities, such as mid-air gestures. However, infra-red eye trackers have very short ranges, making them unsuitable for living room interaction without additional hardware. Video-based eye tracking then arises as a possible solution, but tracking accuracy and lag issues hinder their potential application for real-time control. In this work, rather than attempting to improve the accuracy of the tracker (as is typically the case), we instead investigate how feedback can overcome these issues in a Kinect-based video eye tracker.

METHOD

In this section, we will detail in turn the underlying gaze and gesture system that we have built, the game *eyeSheep* that demonstrates the opportunity for this system, and the user experience and preference study design.

Experimental Set Up

Our system combines a 2D-image based eye tracking software developed by xLabs (www.xlabsgaze.com), the gestural tracking capabilities of the Microsoft Kinect, and a large (40") LCD television.

The xLabs gaze tracking software continuously calculates where the user is looking on the screen in real-time, and will work with any 2D webcam. It is designed for offline analysis of visual attention, so the gaze-tracking can 'lag' if the users' gaze moves rapidly across the large screen. Our current system feeds a cropped version (480x270) of the HD camera from the Kinect device into xLabs. We currently use the default calibration procedure provided by xLabs, which is a 5 point head-rotation calibration. Following calibration, the software creates a 2D map of the face with a high number of facial feature points. This enables the user to have some degree of flexibility with head movement that allows them to leave and return to the same position without having to calibrate again.

We found that the system was able to accurately detect eye-gaze to within a $4cm^2$ area at a distance of up to 2m from the Kinect, with the limiting factor being the resolution of the Kinect HD camera. At this distance, the system ranged from consistently and accurately tracking user-gaze, to being misaligned by up to 6cm. As we were not using a high-accuracy tracker, we were unable to quantify the specific levels of inaccuracy which ranged for each user. Tracking gaze at this 2m distance is suitable for tracking gestural information, where the user is recommended to stand at a minimum of ~1.4m from the Kinect camera. This is allows us to replicate a lounge-style interaction (such as console gaming or TV use), or the comfortable distance that a user might stand from a 40" display in a public space without dedicated hardware placed nearer the user.

Combining Gaze and Gesture in a Game: eyeSheep

In order to evaluate user experience and preference regarding the interaction techniques possible in this system, we developed a simple game that utilized gaze for object selection and gesture for object manipulation. Games are frequently utilized in HCI research to motivate participant use in usability trials, particularly with novel interfaces (Carter et al 2014). In our study, we found that the thematic simplicity of the game (sheep = good, wolves = bad) meant that our participants quickly learnt the goal of the task and were motivated by their timed performance.

eyeSheep challenges the user with sorting sheep into a pen within a one minute time limit. Making the task

Figure 2: Instructions given to *eyeSheep* players.

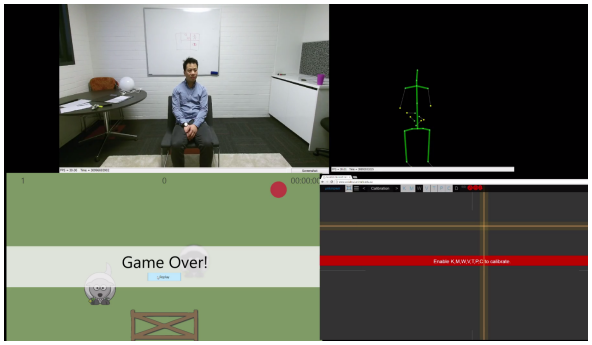

Figure 3: Using OBS we recorded (A) video, (B) skeletal data, (C) the game, and (D) the gaze coordinates.

challenging, there are wolves in sheep's clothing hiding among the flock, hoping to be sorted into the pen (see Figure 1). If the user places a wolf in the pen, the game ends. Each level increases the number of sheep and wolves, making it more difficult to complete the task within the time limit.

The user 'grabs' with their hand by closing their hand into a fist. When this event occurs, the sheep (or wolf) selected is based on the system's estimation of the user's gaze. The user can then move the sheep (or wolf) around the screen by moving their hand (1:1 movement replication), with the goal of placing the sheep inside the gate which becomes highlighted when the sheep is placed within it. Their gaze during this time is irrelevant to their interaction. The user 'lets go' of the sheep (or wolf) by opening their hand, and if the sheep is 'let go' in the pen, it disappears.

Sheep selection is based on a nearest-object calculation (Sibert, 2000). The nearest object is measured from the center of the gaze cursor to the center of the object. This creates a large but dynamic interaction area for an object.

User Experience and Preference Lab Study

We wanted to understand the user experience associated with the combination of gaze and gesture at a lounge-esque distance, and the role of gaze visualisation where the gaze-tracking is imperfect which has received little attention as research has focused on using inauthentic technology set ups to simulate high-accuracy gaze tracking. Consequently, we assembled a mock lounge environment where the user sat 2m from the screen and Kinect device.

We designed two versions of *eyeSheep*. In both versions, the system highlighted which object would be selected if the user 'grabbed' at that particular moment. In our system, this worked by blurring every object on the screen except for the selected object. We felt that this was a thematically coherent way to provide feedback in a system that harnessed user gaze.

Version one (with feedback, or WF) placed a 20x20 red dot on the screen that represented where the system was estimating the user's gaze to be. Version two (no feedback, or NF) provided no visualisation of the user's gaze-point, only the feedback of what object was selected. It was unclear to us which system would be preferred by users, and in what cases.

We considered that the WF version, which provides feedback about how the system is interpreting user gaze, would be useful for users in adapting their behaviour to the limitations of the system. However, as the system is imperfect, this red dot is often in the user's peripheral vision, which could be distracting or exacerbate issues with fatigue. Thus, we wanted to understand if – where gaze-tracking is imperfect - the gaze-visualisation would frustrate users, or provide a useful (or necessary) feedback.

Each participant played both versions of *eyeSheep*, with half of them playing the WF version first (in anticipation of a learning effect). Participants were not told before they began what the difference between the two versions would be, just that 'you'll play two versions of the game that work slightly differently'. After calibrating (which took ~3 min), participants played 5 levels of the first version of the game (which took between 3 and minutes), and then filled in a visual analogue scale measuring 10 dimensions of their experience (detailed in the results section) which was more suitable for our measures than a Likert scale (Shaik and Link, 2003). Before each playthrough, we displayed the game instructions on the screen (see Figure 2).

Play was followed by semi-structured interview questions about their experience, and participants were also asked to explain any strong (positive or negative) responses to the visual analogue scale. Participants re-calibrated the eye-tracker before playing the second version, filling in a second visual analogue scale which included a scale for indicating their preference between the two versions. Two researchers carried out the data collection for this study.

We used Open Broadcaster Software (OBS) to record the Kinect vision data (see Figure 3), gestural tracking, gaze tracking, game-play and audio in a single video file that was transcribed with reference to the user's body-language, physical behaviours and the performance of the gaze and gesture tracking. We also recorded the 3D location of the gestures in reference to the gaze data and gestural tracking, and the speed at which our participants completed each level.

In total, we had 24 participants in the study. 7 participants were omitted from the data analysis for various reasons, including the gaze system not working, the user being unable to calibrate, and interruptions in data collection (such as the computer crashing). In the following sections, we present the results from the 17 successful participants.

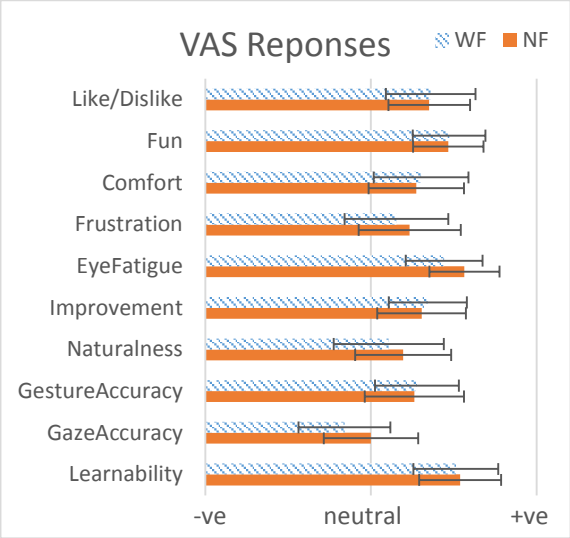

Figure 4 – Reponses to the VAS scale.

RESULTS

Preference

Out of our 17 participants included for analysis, only 4 preferred the version with the active red-dot visualisation (3 having played the no feedback version first), while the remaining 13 participants strongly preferred the version without active gaze visualisation. This significant preference for no gaze feedback is in distinct contrast to earlier research on gaze interaction.

Visual Analogue Scale Responses

After playing each version of *eyeSheep*, participants responded to Visual Analogue Scales (VAS) measuring 10 different dimensions of their experience, grounded in prior work. These were; learnability, perception of gaze accuracy & gesture accuracy, naturalness, improvement over time, eye fatigue, frustration, comfort, fun and like/dislike. These responses were measured so we could identify explanations for different preferences with a higher granularity, and served as useful prompts in the semi-structured interviews following play.

We performed a paired samples T-test in SPSS 22 on these responses split on each version played first, and identified no statistically significant difference ($.825 < p < .090$) between the version with feedback (WF) and the version with no feedback (NF) of *eyeSheep* in any of the 10 dimensions. That is, while users indicated a strong preference towards the NF visualisation, this was not reflected in the 10 dimensions measured in the VAS scale. Figure 4 depicts these responses, including standard deviations. Though not significant, we note the largest difference was a perception that the gaze tracking was more accurate in the NF version.

We also performed a paired samples t-test based on the order of play as a possible effect (learning effect). Marginally significant differences based on order ($p=.039$ and .052 respectively) were only identified for learnability and comfort. The increase in learnability is unsurprising, as users did not have any additional learning demands in their second play. A significant increase in comfort, however, does indicate that future research could allow participants a longer-play time, as comfort with the system did increase after under 5 minutes of use.

Qualitative Reponses

As we identified a strong preference in favour of the version without the red-dot gaze visualisation, unexplained by the 10 dimensions that we measured using VAS, we now turn to the qualitative responses to understand this difference and identify other insights relevant to the design and implementation of gaze and gesture at a distance.

Strategies for Overcoming Low-Accuracy Gaze

Our participants demonstrated a number of emergent adaptive behaviours to compensate for the low-accuracy of the gaze-tracking at the 2m distance we believe serve as useful resources for design.

Intense Stare

Some participants (P2, P5, P8, P24) attempted to overcome inaccuracy by staring intently at the desired sheep they wanted selected. P8, assuming that "*this technology is based on the whites of eyes*", stared as wide-eyed as possible. P5, playing first the NF version, noted:

> Well for example sometimes the system couldn't detect the one that I wanted to choose. But then in my peripheral sight I could see what was chosen, for example which sheep was not blurry, so I tried to focus on that sheep [intensely].

This method of 'staring intensely' at the desired object, waiting for the system to correct itself was problematic, as mostly the system did not correct itself, and appears to have been be associated with higher levels of fatigue. Occasionally this did work where lag was the issue, because the gaze would catch up to the user's gaze location. P10 noted that it would be better if the gaze "*could respond faster*" because the game was timed.

Head Shake

Other participants (P4, P8, P10, P19) attempted to 'refresh' the system by shaking their head, sometimes rigorously, others more slowly and deliberately. P8 spoke about this following their play:

> P8: the strategy of just shaking my head around to give it a slightly different angle on my eyes, the equivalent of shaking the old mouse with the rolla-ball thing. So that and that was kind of based on the configuration aspect of, like, when I did pitch and yaw [to configure].

Consequently, we believe that this re-calibration method was encouraged by the calibration method (described earlier), which presents a model for how the software works which influences emergent behaviours. The nature of the gaze software we were using meant that this occasionally worked, as it actually would 'reset' the system's understanding of the users' face orientation.

Body Movements

As we were providing minimal instruction, others tried ineffective strategies such as waving their hand (P10, P21) or leaning into the screen (P18, P22). P2 would lean or gesture their head towards the direction that they wanted the gaze point to move (only observed when the gaze point was visualised), in effect hinting at the direction they

wanted the red-dot to move towards. Similarly P20 would place their hand in their lap (only observed when the gaze point was not visualised) to "*start over again"* and "*reset"* the system. P24 began saying "*sheep, sheep"* when the sheep wasn't correctly selected during the NF version.

Gaze Offset: Other than these, the most common way of overcoming inaccuracy was by correcting where the user was looking. For instance, if a wolf to the left of the sheep was selected, the user would look to the right of the sheep to select it. P4 looked above the screen to select an object near the top in order to correct their gaze in this way. P5 suggested that this practice, where they were focusing on the plain green background in order to select a sheep, contributed to their fatigue and dislike of the system. This much more frequently shown in the red-dot version of the system (where the users' had some indication of what direction the gaze-tracking was incorrect).

Combining Gaze and Gesture

Echoing Velloso et al (2015), our participants were very positive about the combination of gaze and gesture, considering it "*natural", "compelling"* and "*intuitive"*. However, it is worth noting that our system only combines gaze and gesture as independent modalities; the sheep selected is based *only* on gaze, and the manipulation of the objects is based *only* on gesture.

Hand-Eye Coordination

For all of our participants (except P17) the 3D position of their 'grab' gesture was in relation to the location of the object on the screen (see also Velloso, 2015). That is, if the sheep was on the user's left, they would reach out to the left when grabbing the sheep. This is despite the location of the 'grab' gesture having no bearing upon what is grabbed (with this only being informed by gaze). Many users who grabbed a wolf accidentally would then correct the location of their 'grab' in relation to the wolf (e.g., moving the grab further left if a wolf to the right of a sheep had been grabbed). P10 actually thought that this was a feature in the system:

> R: You preferred the second version [NF]. Could you tell me about that?
>
> P10: Well I don't know but, I just feel it's easier to do it. Umm. I don't know if I'm right, it seems the system is … responds to my hand? It seems when I move my hand the selection will move as well. I don't know if that's the case?

Similarly, P18's first comment on the system was a question about *how* (not *if*) they should be gesturing; in the general direction of the sheep, or on a Cartesian plane (like the current Xbox One interface, that they had experience using). We asked P20 about how they had done this in both versions, who clarified that they understood the location of the gestures didn't matter, but that doing large movements and grabbing in relation to the object's location on the screen made them feel "*more into the game"* and that it "*seemed more natural"*. P13 did this with their whole body, not just their hand gestures. P21 did not believe that the gaze was working at all during the NF version, instead interpreting the absence of the red-dot feedback as indicating the system was only using the gesture locations.

We believe that these demonstrations of this natural behaviour should be incorporated into the design of gaze and gesture interfaces. We will discuss this further at the conclusion of the paper.

Device-Free Interaction

Another reason that our participants were excited about combination of gaze and gesture was that, by using hands and eyes as controllers, anyone could interact with the technology. For example, P20 stated:

> P20: I think [the combination of gaze and gesture is] great. I do not need a game pad to play the game. And like, I think that is good because … I think later on this can be developed as a game which can be played by a lot of people because you do not need the [physical] connection.

Several other participants noted similar opportunities, with some reflecting on opportunities for the system in walk-up public spaces, suggesting contexts like manipulating train-timetables in a subway station (P9) and hospitals (P12).

Focus/Blur Selection Feedback

Several of our participants commented on the blurring of the sheep/wolves to provide visual feedback to the user regarding what would be selected. P8 felt that this method of visualising feedback made sense as it related to the nature of gaze:

> P8: so the blurriness of the animals indicates where the system thinks that I am looking, and so if you were to do that, and you know blur is associated with gaze so if I am looking in the distance at the sheep and its over there, and there is another sheep that is over there, I can't see, that sheep is more blurry than that sheep, so that kind of works.

Only P5 disliked this form of providing feedback, reporting increased fatigue when staring at the blurred sheep when the gaze was not selecting the correct object. While this was not an experimental condition within our research, we believe that this subtle form of feedback worked very well, as no participant had any difficulty understanding or interpreting the feedback.

Sorting Objects

Another strategy that nearly all players employed in the game was sorting the sheep and wolves. Rather than trying to pick the sheep from among the wolves, several users began grabbing the wolves and 'throwing' them to the side to allow easier selection of the sheep. This was a play strategy we had not anticipated. After asking why they began doing this, both P1 and P4 explicitly noted that (as well as being strategic) this was thematically congruent, and fun:

> P4: In order to make it easier for the gaze detector to differentiate between when I was looking at wolves and when I was looking at sheep.
>
> R: Did that feel natural?
>
> P4: Yeah, especially because they're a negative thing. They're a baddie in the game. Throwing them away isn't just strategic, its, "you don't belong here!" [while emphatically throwing aside, and laughing]

While reflective of the gestural excess (van Ryn, 2013; Harper & Mentis, 2013; Apperley, 2013; Downs et al. 2014) noted by others as a key source of fun in Kinect-based games, we also believe this emergent strategy to improve performance could be a suitable design for gaze and gesture interfaces. Giving users the option to permanently discard incorrectly selected buttons would prevent recurring incorrect selections, improving the accuracy of the system in a natural way.

Participants who preferred the Red Dot feedback

Only 4 (of 17) participants preferred the version of *eyeSheep* with gaze visualisation. All four appreciated the continuous gaze feedback, which could be used to adapt the user's gaze to the inaccuracies of the system. P1 (NF, WF) cited "*the fact that there was something on the screen that kind of showed you where you were looking at seemed to help*", and P5 (NF, WF) noted that the red dot helped them "*to better adjust because I know that* [is] *where the system is detecting my gaze*". Without the red dot on the harder levels, P5 (NF, WF) reported feeling like they were "*competing with the technology*" to select the correct sheep, whereas the red dot helped them "*to better adjust because I know that where the system is detecting my gaze*". P9 (NF, WF), who slightly preferred the red-dot visualisation, similarly suggested "*I guess it was kind of interesting to see that element of what information the system was using. So that could contribute to it becoming less frustrating at the time*". This is congruent with our initial assumptions that by displaying the system's estimation of the user's gaze, the user's mastery at the system would improve as a result of being able to adapt their behavior in response to the way it was inaccurate.

P4 (WF, NF) also noted a different dimension of how the continuous feedback improved their experience, explaining their preference as "*I think that's because I could see the red-dot reacting*". Whereas in the NF version they felt like their gaze was "*stuck*" on a sheep, leading them to think the system had stopped working entirely. Whereas in the version without feedback the gaze point could be moving but the selected object stays the same, the red dot was constantly moving. This feedback, in terms of the system actively attempting to get the correct gaze, entrusted P4 with a more positive attitude towards the system. This suggests that systems should provide active feedback in some form, and future work should explore other methods of doing so.

Participants who preferred no feedback

The remaining participants all indicated a preference towards the version of *eyeSheep* that did not provide continuous gaze visualisation in the form of a red dot.

Unnecessary for Simple Tasks

P10 (WF, NF) had a very poor impression of the gaze in the red dot version, considering it very inaccurate "*because I found that I look at the sheep and I can grab it but actually there's a red dot fooling around in my periphery*" whereas in the clear version "*I look at that sheep, and that sheep is being selected, then actually I think that's right, that's correct*". P12 (NF, WF) also preferred the clear visualisation referring to it as being "*like it reads my mind*", noting that they forgot during the version without the red-dot that their gaze was being tracked.

The first level of *eyeSheep* presents the user with a single wolf, and a single sheep. As we utilized nearest-object selection, the gaze could be highly inaccurate but still select the correct object. In the clear version, this is "*seamless"* (P18, NF, WF), and gives the user the impression that the system is working perfectly. When gaze is visualised, even if the performance is the same (e.g., the correct object is still selected), users trust and confidence in the system and themselves is significantly diminished.

P17 (WF, NF) strongly preferred the version without citing that "*If I don't have it, I still believe that maybe the software is working and its, more accurate, I have the feeling its more accurate* [without the red dot]. The key term that P17 kept returning to was that it feels "*more natural"* when there is no red dot. P19 (NF, WF) also suggested that the inaccurate red-dot made it "*a little bit less easy just to gaze and grab*", which had felt "*pretty natural"* without the red dot. P5 (who preferred the WF version) did suggest after they played without a dot that "*in the first half it was more accurate*", that is, when it was a less complex task, "*but then it* [gaze accuracy] *dropped*" when the task became more complex.

Overall, many participants reported that the NF version was more accurate than the WF version (e.g., P13 (WF, NF) thought they were two different types of gaze-tracking, and was surprised to learn that they were the same). This suggests that – in particular for low-complexity tasks, such as the initial levels of *eyeSheep* – gaze feedback is unnecessary, despite the gaze-tracking being relatively low accuracy. As the tasks become more complex, feedback may be necessary. However, other reasons suggest that inaccurate feedback should be avoided.

Inaccurate Gaze Feedback is Distracting

Numerous participants reported finding the red dot "*distracting"* (P8, P12, P17, P20, P24) and "*annoying"* (P8, P24). P8 (NF, WF) referred to it as "*the distracting and annoying dot"*, and P24 (WF, NF) remarked that "*the red dot is really annoying … it was really confusing because it was not matching my eye gaze. Not following my eye properly. The red dot was distracting"*. Similarly, P20 strongly recommended to the interviewer that the red-dot should be removed so that they "*can keep focused on the theme of the game, to catch the sheep"*. While the users who preferred the red-dot visualization found it useful for adapting their gaze, those who did not need it (potentially because the gaze was sufficiently accurate) reported these significant detractions from their experience.

Inaccurate Gaze Feedback contributes to Fatigue

In addition to being distracting, users also suggested that the RD increased the fatigue of using gaze as an interaction technique. P22 (NF, WF) noted that "*the one with the red dot was harder. Like, it was drawing my gaze in a different direction. I couldn't focus on the sheep that I wanted or something"*. We believe that this could be a strong,

underlying reason for why users were distracted and annoyed by the feedback.

Red Dot is Thematically Incoherent

In a similar theme to the positive comments regarding the use of blur/focus to indicate which object is selected, P8 attributed the red dot as "*breaking my immersion"* since since "*there's no metaphor associated with it… there's nothing that moves like that"*. While the task and aesthetic of the game was thematically coherent, a large red-dot which jumped around the screen was not. P8 suggested that the gaze visualization should "*narratively make sense*" with the sheep and 'sheep dog trial' task of the game. This may indicate that, for example, using 'cross-hairs' in a first-person shooter style game to provide feedback might be more suitable than a circular red-dot.

Along the top of the screen of *eyeSheep* is the user's current level, current points score, and time remaining for that level. Though this information has little impact on the user experience, few noticed the timer and all players were surprised if the time ran out. AS P8 stated, "*I'm not looking at the clock. My gaze is controlling the selection, why would I bother looking up there [in the corner] to know the time?"* While this case is specific to this game-based UI, it is worth noting that by using gaze as a key modality for interaction, users become less aware of peripheral information often displayed in game UI's that are integral to their experience.

P8, who preferred the NF version, suggested that the red dot could be replaced by an icon also indicating other information (such as a clock showing how much time is remaining). They speculated that by making the gaze-point visualisation more useful in this way, it would make it significantly less distracting and annoying.

DISCUSSION

In this paper, we have presented the results of our user experience and preference study of two versions of *eyeSheep*, a simple game that combines gaze-based object selection and gesture-based object manipulation at a 2m distance (similar to console gaming or public screen use).

Whereas previous research has typically circumvented the low-accuracy of gaze-tracking in this context dedicated and expensive hardware, we instead used immediately available technologies (the Microsoft Kinect). Due to the inherent limitations of gaze-tracking at these distances, gaze tracking is not as accurate. Consequently, we chose to examine the impact of gaze visualisation on user experience and preference where the gaze tracking is not accurate, along with the overall impressions and natural behaviours our participants exhibited when interacting with this novel interface.

We observed no significant differences between the two versions based on the 10 dimensions of experience we evaluated, but found a significant preference towards the version without an active visualisation of where the system is interpreting the user's eye-gaze. We believe that the primary reason for this strong preference is that gaze-feedback is not necessary when the task is relatively simple, and by not-visualising gaze, users can 'forget' their gaze is being tracked and more seamlessly and naturally use the system, leaving them with an improved impression of the accuracy of the gaze and confidence in their own abilities. Further work similar to this study is necessary to develop best-practice guidelines for designing with less-accurate gaze-tracking.

While the accuracy of the gaze was a recurring issue, this study and proof-of-concept system demonstrate that low-accuracy gaze can be successfully implemented using existing technologies (there are over 8 million Kinect devices in living rooms around the world). While we have highlighted in this paper the issues that users' reported with the system, users were overwhelmingly positive about the fun they had, the opportunities for the system, and were enthusiastic to continue playing to improve their ability. As the sheep in *eyeSheep* are a similar size to the buttons in the Xbox One 'metro-style' interface, gaze provides an immediate opportunity for improving the experience and quality of gestural interaction at a distance.

In this study, we have utilized a simple game in order to better understand this interaction technique, rather than a work-based task used in other studies of gaze and gesture (e.g., Velloso et al. 2015). Reasons for this are twofold. Firstly, we were interested in user experience and preference rather than objective, quantitative task performance. We felt that by engaging users in a competitive, game-based task, they would be more engaged in the successful performance of the task (sorting the sheep into the pen). Secondly, as gaze-based gaming continues to emerge as a standalone area of research and as a commercial drive for gaze-tracking, our study has shown that play-based gaze and gesture interactions are similar to task-based gaze and gesture interactions.

FUTURE WORK

We designed this experimental set up as a proof-of-concept to demonstrate the immediate opportunities of gaze and gesture on the Microsoft Kinect. Future research could explore low-accuracy gaze-tracking usability and design using a high-accuracy gaze tracker, with the accuracy controlled, in order to identify the thresholds that might change what type of feedback is necessary. Future work should continue to explore different methods for providing feedback, and their ability to circumvent issues associated with low-accuracy gaze tracking.

As all but one of our participants gestured in the direction of the object they were selecting (despite it being irrelevant to what object was selected), future work should begin to explore combining gaze and gesture in more integrated ways. For example, by using the 3D location of the gesture to improve the accuracy of the selection. Similarly, we intend to explore the other natural behaviours that we identified in this study (such as head-shaking, and discarding incorrect selections) as opportunities for further improving the usability and experience of gaze-and gesture interaction at a distance.

REFERENCES

Apperley, T. The body of the gamer: game art and gestural excess. Digital Creativity 24, 2 (2013), 145-156.

Ashmore, M., Duchowski, A.T. and Shoemaker, G. Efficient eye pointing with a fisheye lens. In Proc. Graphics Interface, ACM Press (2005), 203–210.

Bojko, A. Eye tracking the user experience: A practical guide to research. Brooklyn, NY: Rosenfeld Media.

Bolt, R.A. Gaze-orchestrated dynamic windows. In ACM SIGGRAPH Computer Graphics (1981), 109–119.

Carter, M., Downs, J., Nansen, B., Harrop, M. & Gibbs, M.. Paradigms of games research in HCI: a review of 10 years of research at CHI. In Proc. CHI Play 2014, ACM Press (2014), 27-36.

Cha, T. and Maier, S. eye gaze assisted human-computer interaction in a hand gesture controlled multi-display environment. In Proc. Gaze-In'12, ACM Press (2012), article no. 13.

Cho, D.-C. et al., Long range eye gaze tracking system for a large screen. IEEE Transactions on Consumer Electronics 58, 4 (2012), 119-1128.

Cho, D.-C. & Kim, W.-Y. Long-Range Gaze Tracking System for Large Movements. IEEE Transactions on Biomedical Engineering 60, 12 (2013), 3432–3440.

Downs, J., Vetere, F., Howard, S., Loughnan, S. and Smith, W. Audience experience in social videogaming: effects of turn expectation and game physicality. In Proc. CHI'14, ACM Press (2014), 3473-3482.

Duchowski, A. Eye tracking methodology: Theory and practice, Springer Science & Business Media (2007).

Duchowski, A., Price, M. and Meyer, M. Aggregate Gaze Visualisation with Real-Time Heatmaps. In Proc. ETRA'12, ACM Press (2012), 13-20.

Fisk, L., Carter, M., Yeganeh, B. R., Vetere, F. and Ploderer, B. Implicit and Explicit interactions in video mediated collaboration. In Proc. ozCHI'14, ACM Press (2014), 250-259.

Harper, R. and Mentis, H. The Mocking Gaze: Social Organization of Kinect Use. In Proc. CSCW'13, ACM Press (2013), 167-180.

Hennessey, C. and Fiset, J. Long range eye tracking: Bringing eye tracking into the living room. In Proc. ETRA'12, ACM Press (2012), 249–252.

Jacob, R.J. Eye movement-based human-computer interaction techniques: Toward non-command interfaces. Advances in Human-Computer Interaction, 4 (1993), 151–190.

Koons, D.B., Sparrell, C.J. and Thorisson, K.R. Integrating simultaneous input from speech, gaze, and hand gestures. In M. Maybury (ed.) Intelligent Multi Media Interfaces, Cambridge, MA.: MIT Press (1993), 453-454.

Kosunen, I. et al. Comparing eye and gesture pointing to drag items on large screens. In Proc. ITS'13, ACM Press (2013), 425-428.

Lander, C., Gehrig, S., Kruger, A., Boring, S., and Bulling, A. GazeProjector: Location-independent gaze interaction on and across multiple displays. In DFKI'15 (2015).

Majaranta, P., Aula, A. & Räihä, K.-J. Effects of feedback on eye typing with a short dwell time. In Proc. ETRA'04, ACM Press (2004), 139–146.

Majaranta, P., MacKenzie, S., Aula, A. and Raiha, K-J. Effects of feedback and dwell time on eye typing speed and accuracy. Universal Access in the Information Society 5, 2 (2006), 199–208.

Majaranta, P. & Bulling, A. Eye Tracking and Eye-Based Human-Computer Interaction. In Advances in Physiological Computing, Springer (2014), 39–65.

Mubin, O., Lashina, T. & van Loenen, E. How not to become a buffoon in front of a shop window: A solution allowing natural head movement for interaction with a public display. In INTERACT 2009. Springer (2009), 250–263.

Noris, B., Benmachiche, K. & Billard, A.G. Calibration-Free Eye Gaze Direction Detection with Gaussian Processes. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE (2008).

O'Hara, K. et al. Touchless interaction in Surgery. Communications of the ACM 57, 1 (2014), 70-77.

Schaik, P. and Ling, J. Using online surveys to measure three key constructs of the quality of human-computer interaction in web sites: psychometric properties and implications. Int. J. Human Computer Studies 59, (2003), 545-567.

Seifert, K., Evaluation multimodaler computer-systeme in frühen entwicklungsphasen. PhD thesis, Department of Human-Machine Systems, Technical University Berlin (2002).

Sesma, L., Villanueva, A. & Cabeza, R. Evaluation of pupil center-eye corner vector for gaze estimation using a web cam. In Proc. ETRA'12, ACM Press (2012) 217–220.

Sibert, L.E. and Jacob, R.J. 2000. Evaluation of eye gaze interaction. In Proc. CHI'00, ACM Press (2000), 281–288.

Song, J. et al. GaFinC: Gaze and Finger Control interface for 3D model manipulation in CAD application. Computer-Aided Design, 46 (2014), 239–245.

Turner, J., Alexander, J., Bulling, A. and Gellersen, H. Gaze+ RST: integrating Gaze and multitouch for remote Rotate-Scale-Translate tasks. In CHI'15, ACM Press (2015), 4179-4188.

Turner, J., Alexander, J., Bulling, A., Schmidt, D. and Gellersen, H. Eye Pull, Eye Push: Moving Objects between Large Screens and Personal Devices with Gaze Touch. In Proc. INTERACT'13, Springer (2013), 170-186.

van Ryn, L. Gestural Economy and Cooking Mama: Playing with the Politics of Natural User Interfaces. Journal of Media Arts Culture 10, 2 (2013).

Velloso, E. et al. An empirical investigation of gaze selection in mid-air gestural 3D manipulation. In Proc. INTERACT'15, ACM Press (2015).

Velloso, E. et al. Arcade+: A platform for public deployment and evaluation of multi-modal games. In Proc. CHI Play'15, ACM Press (2015).

Vertegaal, R. Attentive User Interfaces. Communications of the ACM 46, 3 (2003), 31-33.

Walter, R., Bailly, G., & Müller, J. Strikeapose: revealing mid-air gestures on public displays. In Proc. CHI'13, ACM Press (2013), 841-850.

Ware, C. & Mikaelian, H.H., An evaluation of an eye tracker as a device for computer input. In ACM SIGCHI Bulletin (1987), 183–188.

Yarbus, A.L. Eye movements and vision. Cambridge, UK: Springer (1967).

Yoo, B.I. et al. 3d user interface combining gaze and hand gestures for large-scale display. In Proc. CHI EA'10, ACM Press (2010), 3709–3714.

Zhang, Y., Bulling, A. and Gellersen, H. Sideways: A gaze interface for spontaneous interaction with situated displays. In Proc. CHI'13, ACM Press (2013) 851–860.

Zhang, X., Feng, W. & Zha, H., 2011. Effects of Different Visual Feedback Forms on Eye Cursor's Stabilities. In Internationalization, Design and Global Development. Springer, pp. 273–282.