# Extending the Visual Field of a Head-Mounted Eye Tracker for Pervasive Eye-Based Interaction

Jayson Turner*
Lancaster University

Andreas Bulling†
University of Cambridge

Hans Gellersen‡
Lancaster University

## Abstract

Pervasive eye-based interaction refers to the vision of eye-based interaction becoming ubiquitously usable in everyday life, e.g. across multiple displays in the environment. While current head-mounted eye trackers work well for interaction with displays at similar distances, the scene camera often fails to cover both remote and close proximity displays, e.g. a public display on a wall and a hand-held portable device. In this paper we describe an approach that allows for robust detection and gaze mapping across multiple such displays. Our approach uses an additional scene camera to extend the viewing and gaze mapping area of the eye tracker and automatically switches between both cameras depending on the display in view. Results from a pilot study show that our system achieves a similar gaze estimation accuracy to a single-camera system while at the same time increasing usability.

**CR Categories:** B.m [Hardware]: Miscellaneous—;

**Keywords:** Visual Field, Eye-based Interaction, Multiple Cameras, Mobile Eye Tracking, Pervasive Eye-based Interaction

## 1 Introduction

Our current research focuses on eye-based interaction between two displays in the environment, one public wall mounted display and one personal hand-held display. We use a head-worn video-oculography system. Whilst there are few issues detecting a public display directly in front of a user, a hand-held display is typically positioned closer, below the visual range of an eye trackers scene view. Despite being able to move freely, users must consciously lower their head beyond natural limits to bring the display in to full view of the scene camera. An ideal eye tracking system would allow users to interact with both displays without correcting their head posture to accommodate for the eye tracker.

The use of mobile eye tracking for pervasive interaction requires that systems have the ability to detect displays of different sizes, at varying locations in the environment. An off-the-shelf eye tracker such as the iView X HED[1] from SensoMotoric Instruments (SMI) uses a 1/3" sensor and a 3.6mm lens, providing a viewing angle of ˜56°vertically and ˜74°horizontally The viewing angle of the scene camera can typically be increased using lenses with less than 2mm focal length to achieve a wider view of around ˜97°vertically and ˜129°horizontally, though this can result in a loss of accuracy due to distortions and lowered resolution.

*e-mail: j.turner@lancaster.ac.uk
†e-mail: andreas.bulling@acm.org
‡e-mail: hwg@comp.lancs.ac.uk

[1]http://www.smivision.com

An increasing number of researchers are investigating eye-based interaction with multiple displays in the environment. For example, work by [Turner et al. 2011] and [Mardanbegi and Hansen 2011] has focused on the detection of screens within the environment and techniques to allow interaction. In particular, Turner's work described a multimodal approach that combines touch and gaze to move objects between public and personal displays. The work also prompted discussion of the challenges of robust gaze mapping to screens of different sizes at different proximities.

[Wagner et al. 2006] and [Schneider et al. 2009] showed that a motorised camera can be used to extend the visual field of an eye tracker. They combined two scene cameras to create a single, low resolution wide-angle view of the scene with a smaller but high resolution overlay. The latter was provided by a motorised camera that represents a user's foveal point. Their approach allowed them to record a wider area of the scene but required complex equipment.

This paper presents a system that uses an additional camera to extend the visual field of a head-mounted eye tracker while maintaining a high resolution. We focus specifically on using this for the detection of screens that are not necessarily detectable using a standard system. The issue being that their viewing range combined with the natural physical head-movements of a user is not wide enough. We discuss results from a preliminary study to determine the advantages and disadvantages of the modified system compared with a standard system.

## 2 Issues with Single-Camera Systems

State-of-the-art eye tracking systems use two cameras for gaze tracking. The first camera is used to track one eye from close proximity, while the second, the scene camera, records part of the visual scene. This approach works well for interaction with several displays placed at similar distances to the user, e.g. in a multi-display office setting. Interaction with displays at different distances, however, is often impossible with such systems.

For a head-mounted eye tracker to provide accurate mapping to screens in the environment, the following conditions need to be met: (1) crucially, a view of the eye and infrared reflection in the cornea is needed and cannot be occluded by the eyelid. (2) the screen needs to be in full view of the scene camera.

When interacting with a hand-held device it can be difficult to bring the device in to full view of the scene camera, this is due to the device's close proximity to the user and the viewing angle of the camera lens. Figure 1 (left) shows that although the user is looking at a hand-held device below them, mapping gaze to the device is not possible as their eyelid is occluding their eye and the device is not visible in the scene view. Figure 1 (right) shows that the user must lower their head to a point where their eyes can no longer see the device but will allow for the system to detect it. Figure 1 (centre) shows the ideal head position needed to perceive the device with the eyes and reduce occlusions from the eyelids however, the hand-held device still cannot be seen by the scene camera. To allow for robust mapping, the system requires a method of observing the hand-held device without the user needing to compensate by lowering their head to the level shown (right) in Figure 1.

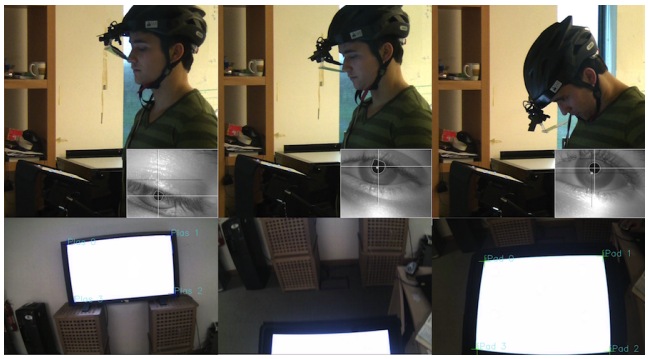

**Figure 1:** *Left: a user looking at the tablet below, the eye is occluded and the public display is in the scene view. Centre: user performing a slight head movement to bring the pupil in to view for tracking, the tablet is not fully visible. Right: the required head position needed to bring the mobile device in to full view.*

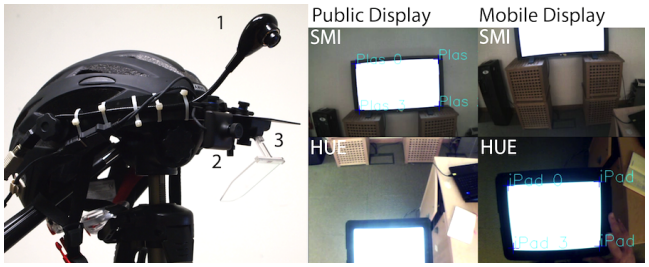


**Figure 2:** *Left: SMI tracker with additional scene camera mounted on a gooseneck (1) HUE Webcam. (2) SMI Scene Camera. (3) Eye Camera. Centre: scene views when the users attention is directed to the public display. Right: scene views when attention is directed to the mobile device.*

## 3 Dual Scene Camera Eye Tracking

Based on the issues outlined in Section 2, we developed a system for robust tracking across multiple displays at varying distances to the user. Our system uses a dedicated scene camera for each display. This aims to provide two main advantages over a standard single-camera system: (1) A wider view of the scene and (2) reduced head movement to bring a hand-held device into view.

### 3.1 Hardware

The basis for our system was SMI's iView X HED eye tracker. This eye tracker uses two cameras, one to track the pupil and corneal reflection by way of a hot mirror and the other to observe the scene. The scene camera provides a vertical viewing angle of ˜74°from a 1/3" sensor fitted with the default 3.6mm lens. It has a resolution of 752x480 pixels at 30Hz though it can only transmit video at 10Hz to third party applications. To extend the camera's field of view we added a HUE HD Webcam (HUE) to the SMI system that runs at 30Hz with a resolution of 640x480 pixels (see Figure 2). This camera has an identical sensor size and lens fitted providing ˜74°of vertical visual field. The HUE camera was mounted on a gooseneck, this gave us the flexibility to realign it for different users. As shown in Figure 2, aligning the two cameras allowed us to extend the gaze mapping and viewing capabilities of the SMI tracker at natural head positions, this allowed for the detection of both displays.

It is important to note that the area covered by the webcam must still be within the tracking range of the eye tracker. For example, aligning the secondary camera too low relative to the primary one would not allow for tracking at that angle due to occlusions caused by the eyelids when the user looks down.

### 3.2 Screen Detection

Our system relies on the detection of screens within each camera view to determine which has the attention of the user.

**Detection of the public display** To detect a wall-mounted screen our implementation used several functions from the OpenCV computer vision library. For this algorithm to work correctly, lighting conditions need to be dimmed. We first convert incoming scene frames to the Hue, Saturation and Value (HSV) colour model. This allows us to threshold the image using a high Value, to produce binary images to identify the brightest parts. We then use *cvFindContours* to detect contours within the image and *cvApproxPoly* to reduce the number of points describing them. The next step compares each of these contours by first discarding those below an area of 1000 pixels. The dot product of each of the joining vectors is then calculated to compare the angles at which they meet. If each of these equate to ˜90°the contour is regarded as a screen. As we use two different screens in our setup, the ratios of their sides can be used to differentiate between them.

**Detection of the portable device screen** The screen of a hand-held touch device can not be detected in the same way as for the public display as its corners and sides may be obscured by fingertips and hands during operation. For this method it is assumed that the tablet will be the only bright object in view. This allows us to search for and select the largest blob in view using the cvBlob[2] library. The bounding box and convex hull of the blob are found. We then find the closest convex hull point within a set distance to the bounding box corners. If a suitable hull point cannot be found close enough, that corner is used instead as an estimate.

### 3.3 Calibration

As we want to map gaze to two displays at different distances, it is necessary to perform two calibrations that our software can switch between. The user is first calibrated for the SMI system's scene camera using the 4 corner points and the centre point of the wall-mounted display. Once completed, the user then lowers their head to bring the tablet into view of the HUE camera to undergo another 5 point calibration using points on the tablet. During this calibration our software uses gaze values reported by the SMI, some of which may not be within the original scene camera's bounds. For each calibration point, its location within the HUE cameras bounds and gaze value are recorded. These values are then used to calculate a homography. This is then applied to incoming gaze values to estimate the gaze location within the HUE camera view.

### 3.4 Operation

During operation our system attempts to detect screens in each camera view. By default, the gaze reported by our system is relative to the upper SMI scene camera. As the system is calibrated at two difference distances, it can switch between these calibrations to ensure gaze mapping remains accurate in relation to which screen the user is interacting with. The system uses a simple condition to infer which display has the attention of the user. If the tablet is detected

[2]http://code.google.com/p/cvblob/

within the HUE camera view, the homography calculated from the second calibration described above is used. If the wall-mounted display is detected in the SMI scene camera, the gaze data reported by the SMI system is used. Both screens will not be detected simultaneously due to the alignment of the two cameras.

Once the screen in focus has been detected and the gaze values have been obtained, they are transformed from scene coordinates to screen coordinates. This is done by taking the four corner points of the screen to compute a homography that is then applied to the gaze data.

# 4 Prototype Evaluation

We conducted a short study with eight participants (2 female, 6 male, aged between 23 and 30 years) to compare our dual-camera system with a standard single-camera eye tracker. Participants performed a simple eye-based interaction task: moving similar sized objects between two screens in the environment using a technique that combines touch and gaze-based input. The environmental screens were a mobile device and a public display. To move an object, the participant had to fixate on it, perform a touch hold action on the mobile device, direct their gaze to the mobile device's screen and then release the touch. This dropped the object at the location of their gaze. We measured the accuracy and time in which participants could pick up, move and drop objects using the above technique. We evaluated this alongside qualitative information obtained from questionnaires to determine if our modified system showed any performance drop in this particular situation.

## 4.1 Apparatus

The study was performed under dimmed lighting conditions to ensure there was no interference from fluorescent lighting. We used the following equipment:

- SMI iView X HED head-mounted eye tracker
- SMI tracker with additional HUE HD Webcam
- Public Display: 50" plasma, 1280x768 @ 60 Hz
- Mobile Device: tablet computer, 1024x768 @ 60Hz

In addition, the study used our screen detection software as described in section 3.2, this software delivered gaze data to our object moving application running on the PC. The objects to be moved were light blue coloured circles measuring 150px in diameter on both displays. The targets were the same shape and size but coloured white with a black dotted outline.

## 4.2 Procedure

After arriving in the lab, participants were first introduced to the study and asked to fill the first part of the questionnaire on demographics. Participants were then asked to stand about 150cm from the display to ensure it was in full view of the scene camera. To exclude influences of distance on eye tracking accuracy, the tablet was mounted on a tripod in front of each participant.

Each participant was asked to perform 10 repetitions of two tasks with each of the systems. The first task required participants to move an object from the wall-mounted display to the tablet. The second task required participants to do the opposite, i.e. move an object from the mobile device to the public display. After an initial eye tracker calibration, participants were asked to look at a circle at the centre of the public display to initiate the first repetition. Each consecutive repetition was started in the same way. For all repetitions, we recorded the pick up and drop locations of each object and target as well as the full gaze path on both screens. Each of these events, including touch events, were timestamped.

# 5 Results

## 5.1 Qualitative Feedback

From the questionnaire we found that all participants had prior knowledge of and experience with eye tracking systems. Table 1 shows the means and standard deviation taken from participant questionnaires. These results are calculated from a 7-point Likert scale. To analyse the perceived usability of each system statistically we used a Wilcoxon signed-rank test on participant responses in both conditions.

Overall, participants found themselves in significantly greater control with the dual-camera system than with the single-camera system, $Z = -2.03, p < .05$. They also found that the dual-camera system required significantly less mental demand ($Z = -1.83, p < .05$) and significantly less frustration ($Z = -1.73, p < .05$). Most importantly, participants found that they had to put significantly less conscious effort into getting the dual-camera system to work, $Z = -1.75, p < .05$.

In the free text comments two participants stated that the single-camera system caused them physical discomfort, one participant said *"It requires too much effort, I finished the activities with a pain in the neck"*. Another stated *"The system caused neck pain very quickly"*. This was due them needing to lower their head beyond a comfortable limit to allow for the tablet to be detected. Participants generally stated that they found it easier to perform the tasks with the dual-camera system saying it was *"preferred and easier to use"*. In terms of accuracy, speed and control participants wrote in favour of the dual camera system saying *"I felt more in control"* and that it *"appeared to be more accurate"*. In contrast, one participant highlighted that they found the single-camera system easier simply because they had learned to move their head much lower than needed to allow the system to register the tablet display. Although this made the system work, they said they couldn't correctly perceive the tablet within their vision.

## 5.2 System Performance

Table 3 shows the individual and overall mean error and standard deviation in accuracy when picking up and dropping objects. The mean error was calculated using the distance between participants gaze and the centre of an object or target. We removed runs where the user was unable to move the object to the desired screen. The single-camera system had a total of seven failed runs, five when dropping the object on the tablet and one dropping on the display. The two-camera system had three failed runs when dropping objects on the tablet. Table 3 shows that P2, P7, and P8 were distinctly more erroneous with the single-camera system than with the dual-camera system. P4 was more erroneous with the dual-camera system when dropping objects.

Table 2 shows the average times required for each participant to complete each run of each task. In general, the mean time taken for each task across both systems can be considered to be similar though P4, P6 and P7 demonstrated significantly higher task times.

| | Single Cam Mean (SD) | Dual Cam Mean (SD) |
|---|---|---|
| 1. You were able to accurately pick up the object | 3.12 (1.64) | 4.75 (2.05) |
| 2. You were able to drop the object accurately within the target zone | 3.12 (1.55) | 4.62 (1.68) |
| 3. You felt as if you were in full control of the system | 2.62 (1.68) | 4.5 (1.60) |
| 4. The system responded to your commands and actions quickly | 4 (2) | 4 (2.13) |
| 5. How much mental demand did this technique require? | 6 (0.75) | 4 (2) |
| 6. How much frustration did this system cause? | 5.75 (0.70) | 3.37 (1.92) |
| 7. How much conscious effort did you put in to head movement? | 6.37 (0.91) | 4.37 (1.59) |

**Table 1:** *Qualitative questionnaire results relating to tablet interaction. Means and standard deviation from a 7 point likert scale. 1=Very Low, 2=Moderately Low, 3=Slightly Low, 4=Undecided, 5=Slightly High, 6=Moderately High, 7=Very High.*

| | Single Camera | | Dual Camera | |
|---|---|---|---|---|
| | Task 1 | Task 2 | Task 1 | Task 2 |
| P1 | 4.174 | 3.472 | 4.395 | 3.999 |
| P2 | 8.144 | 4.508 | 8.811 | 5.370 |
| P3 | 5.017 | 3.936 | 5.924 | 3.864 |
| P4 | 5.806 | 3.347 | 14.193 | 2.735 |
| P5 | 3.367 | 3.263 | 7.405 | 3.626 |
| P6 | 7.306 | 2.325 | 9.927 | 13.042 |
| P7 | 14.520 | 5.315 | 5.187 | 2.979 |
| P8 | 9.498 | 4.357 | 6.440 | 2.365 |
| Mean (seconds) | 7.229 | 3.815 | 7.785 | 4.747 |
| SD (seconds) | 3.591 | 0.918 | 3.170 | 3.477 |

**Table 2:** *Values in pixels. Mean times for each task and system as well as the mean and standard deviation across all participants.*

## 6 Discussion

The results of our short study demonstrate that across eight users our dual-camera system performed better in terms of accuracy but similarly in terms of speed when compared to the single-camera system.

Several participants accuracy results when dropping objects were considerably worse for the single-camera system when compared to the dual-camera system. When compared to the equivalent results when picking up objects the results are much more accurate. This can be attributed simply to participants becoming impatient with the system and dropping objects out of the target zone. This is reflected in Table 1 question 7 where the level of frustration reported is significantly higher than with the dual-camera system.

Qualitative feedback from users supports our dual-camera system and shows improved usability over the single-camera system, in general participants commented on finding it easier to use, more controllable and comfortable during interaction.

It is clear from the information presented in Section 2 and the details of our results that there is a usability issue with regards to using head-mounted eye tracking to interact with close proximity displays. Our dual-camera system allowed for the use of more natural head movement to detect a tablet device without a toll on performance when compared to a standard single scene camera system.

| | Single Camera | | Dual Camera | |
|---|---|---|---|---|
| | Tablet | Plasma | Tablet | Plasma |
| P1 (Pick) | 34 | 51 | 31 | 30 |
| P2 (Pick) | 55 | 58 | 46 | 60 |
| P3 (Pick) | 40 | 69 | 46 | 33 |
| P4 (Pick) | 51 | 42 | 51 | 55 |
| P5 (Pick) | 45 | 55 | 60 | 50 |
| P6 (Pick) | 43 | 55 | 36 | 41 |
| P7 (Pick) | 53 | 54 | 38 | 39 |
| P8 (Pick) | 46 | 46 | 49 | 36 |
| Mean (pixels) | 45.87 | 53.75 | 44.62 | 43 |
| SD (pixels) | 7.01 | 8.10 | 9.28 | 10.82 |
| P1 (Drop) | 24 | 18 | 34 | 23 |
| P2 (Drop) | 102 | 79 | 51 | 53 |
| P3 (Drop) | 45 | 59 | 46 | 47 |
| P4 (Drop) | 122 | 44 | 204 | 62 |
| P5 (Drop) | 57 | 60 | 63 | 46 |
| P6 (Drop) | 57 | 55 | 62 | 60 |
| P7 (Drop) | 195 | 34 | 52 | 25 |
| P8 (Drop) | 108 | 73 | 67 | 59 |
| Mean (pixels) | 88.75 | 52.75 | 72.37 | 46.87 |
| SD (pixels) | 54.85 | 20.11 | 54.23 | 15.28 |

**Table 3:** *Values in seconds. Accuracy of both systems when dropping and picking objects. Shown are the mean errors for each participant along with overall means and standard deviation.*

## 7 Conclusion

In this paper we have highlighted the issues around interaction with a public and close proximity display. We showed that users had to perform unnecessary and uncomfortable head movements in order to register a hand-held display within the scene view of a head-worn eye tracker. We presented a prototype solution that utilised an additional scene camera for tracking the lower area of a user's visual field. We compared our system with a standard single-camera system in a user study with eight participants. These results show that our system provides improved usability over a standard system with a significant reduction in head movement without a toll on speed and accuracy in interaction.

## References


MARDANBEGI, D., AND HANSEN, D. W. 2011. Mobile gaze-based screen interaction in 3d environments. In *Proc. NGCA*, ACM Press, 2:1–2:4.

SCHNEIDER, E., VILLGRATTNER, T., VOCKEROTH, J., BARTL, K., KOHLBECHER, S., BARDINS, S., ULBRICH, H., AND BRANDT, T. 2009. EyeSeeCam: An Eye Movement-Driven Head Camera for the Examination of Natural Visual Exploration. *Annals of the New York Academy of Sciences 1164*, 1, 461–467.

TURNER, J., BULLING, A., AND GELLERSEN, H. 2011. Combining gaze with manual interaction to extend physical reach. In *Proc. PETMEI*, ACM Press, 33–36.

WAGNER, P., BARTL, K., GÜNTHNER, W., SCHNEIDER, E., BRANDT, T., AND ULBRICH, H. 2006. A pivotable head mounted camera system that is aligned by three-dimensional eye movements. In *Proc. ETRA*, ACM Press, 117–124.