Document Version 1.2.0
19 June 2014
xLabs Pty Ltd, Australia
http://www.xlabsgaze.com/
http://www.xlabs.com.au/

Eye / Gaze tracking system
Technology Whitepaper

1. Introduction

The xLabs Eye/Gaze tracking system is able to continuously calculate where you are looking on a computer screen, while allowing natural freedom of movement. It does not require any special hardware - just any ordinary webcam.

This gaze data is calculated locally and in realtime (i.e. it is available as soon as it happens). To the best of our knowledge, there is no other software-only system with these capabilities.

The system is free to download and use for non-commercial purposes. The user interface is integrated into web browsers (such as Chrome, Safari and Firefox) allowing eye/gaze to be tracked as part of any website or web application. We encourage users and developers to build useful tools and programs on top of the system, using our Javascript interface.

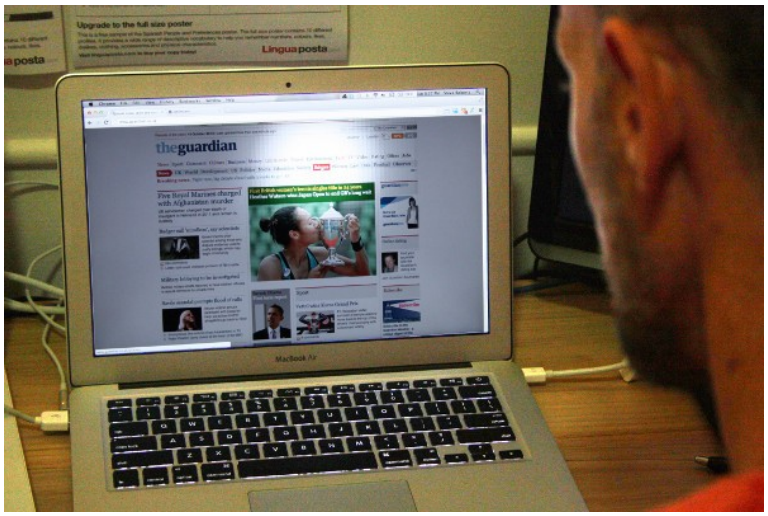

Figure 1: Spotlight on screen, tracking user's gaze

1.1 Objectives

The objective of this study is to quantify the accuracy of the xLabs Eye/Gaze tracking system, both in realistic conditions and on sufficient scale to validate generalization across the gamut of human appearance and variation. These results are necessary to show that the system can be successfully used, at scale, for various 3rd party purposes. Development of applications using the system is contingent on an understanding of any limitations that affect results.

The study should capture all expected difficulties concerning general, mass scale use of eye/gaze tracking using only an ordinary video camera as input. These include:

- variations in human appearance, in particular due to age and ethnicity
- variations in camera hardware used for image capture
- variations in image capture environments, and in particular illumination / lighting

The study should also capture a larger sample of data than used for earlier ad-hoc testing.

A final objective is to validate that the calibration model provided in the software is capable of generating accurate gaze results, in realistic conditions, without expert advice or intervention from the developers. The participants in the study will not have specific skills in imaging or other relevant fields, or strong motivation to behave nicely towards the system. Therefore this study represents and realistic and challenging scenario.

2. Methods

2.1 Study design

The study was intended to be conducted on as large a sample of people as could practicably be obtained. People of all ages and ethnicities were requested to participate.

Accuracy was measured by comparison of screen coordinates we asked participants to look at, with screen coordinates “predicted” by the xLabs client software, given images of participants’ faces.

The xLabs client software was calibrated via a set of images captured during participants’ clicks on known screen coordinates. No methods were employed to ensure that participants looked at the coordinates requested during clicks; however it is assumed that most participants did as asked most of the time. The system is expected to remove outliers (e.g. due to non-compliant users) automatically.

Participants were acquired via their workplaces, via random approach in Melbourne city centre, and via personal contacts of the xLabs team. Images were captured in public, outdoors; inside public buildings; and in real workplace and home environments. Lighting was not controlled, although very dark environments were excluded.

Approximately half the subjects were captured via a third party associate who had only a basic understanding of the difficulties posed by lighting variation and no understanding of the computer vision methods used in the software. Therefore these results should be replicable in real-world conditions without expert knowledge of the technical methods used.

2.2 Software

To validate the calibration system used in the xLabs Eye/Gaze tracking system, the same interface was used for this study. The xLabs client software is capable of recording camera images to disk (but not transmitting them over a network). Therefore, the only software required was the xLabs client software.

Video was recorded as a sequence of JPEG images at a rate of 10-15 frames per second. Images were captured at two resolutions, either 640x480 or 1280x720 pixels (depending on camera and computer used).

The xLabs client software interface includes a calibration screen (see figure 2) which displays a grid of ‘+’ marks. The software randomly picks one mark and highlights it with a red circle. When the user presses and holds the mouse button over the selected mark, the circle turns yellow and then green. It is assumed that the user is looking at the circle during this time. The software collects the clicked coordinates and images of the user captured while the mouse button is being pressed.

In an offline processing stage, after the data was collected, the calibration system in the xLabs client software was then provided half the mouse press coordinates and images captured for

each subject. A calibration was triggered for each subject. The resulting calibration was then used to predict gaze coordinates in all recorded images for that subject during the times the mouse button was pressed.

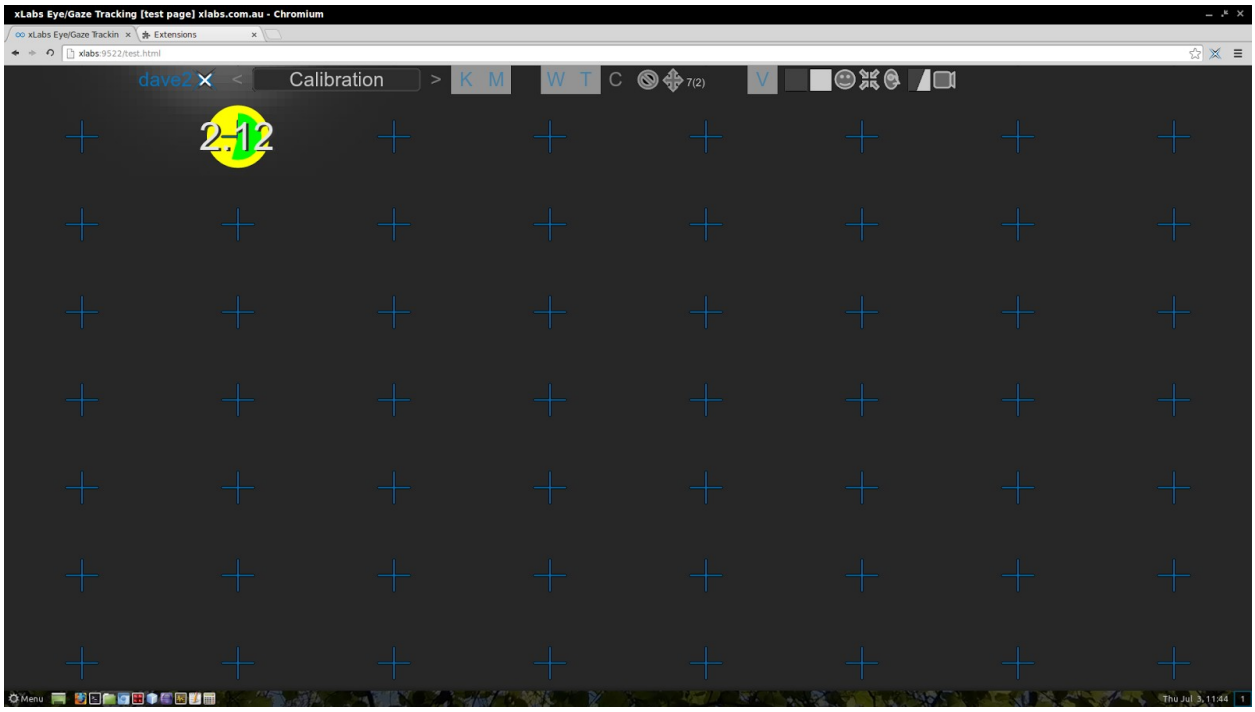

Figure 2: Interface presented to participants in the study. Participants were asked to move the mouse to a randomly selected '+' mark (highlighted with a red circle) and then press and hold the left mouse button for the duration of a countdown. During the mouse press, participants were asked to continuously looking at the coloured timer until it went completely green.

2.3 Instructions to participants

Participants were shown the interface and given the following instructions:

- Sit at a comfortable distance from the computer for normal use
- Requested to press and hold the mouse button on the red circle; note that it changes colour and eventually goes green.
- Try to look at the circle at all times while pressing the mouse button.
- Can move naturally throughout the session.
- Sit however is comfortable, don't have to stay still.
- Blinking is not a problem; just relax.
- If you make a mistake, don't worry, just keep going.
- Asked to repeat the clicking process until 60+ clicks are logged.

Participants were made aware that images were being captured and the purpose of the study was explained to them. Participants were not rewarded for volunteering to be part of the study.

2.4 Hardware

We used an ad-hoc collection of available computers and cameras to collect data for this study, using a variety of operating systems. Laptop inbuilt cameras were used in most cases; however, for completeness an external Logitech USB webcam was used for some data.

Computer / Camera	**Operating System**
Toshiba Satellite U400 (13” screen)	Ubuntu Linux
Lenovo Thinkpad SL500 (15” screen)	Windows 7
Lenovo Thinkpad T530 (15” screen)	Linux Mint
Metabox W350ST (15” screen)	Linux Mint
Macbook Pro 15” 2012	OSX
Macbook Pro 15” 2014	OSX
Macbook Air 13” 2012	OSX
Logitech USB camera	n/a

3. Results

Approximately 1 million images were captured during mouse-press events by 300 participants. We present results in two sections. First, accuracy of gaze prediction. Second, a breakdown of the circumstances and participants of the data collected.

3.1 Gaze Accuracy

Participants were required to move the mouse pointer to a '+' mark displayed on screen, then click and hold the mouse for 3-4 seconds at least 60 times. We calibrated the system by using half of the image and click measurements obtained for each participant, and then used the calibrations to predict that participant's gaze points given only the recorded images. This process was entirely autonomous with no manual intervention, data curation or filtering.

We measured gaze accuracy by comparing the measured position of the mouse pointer during each click, to the predicted screen gaze coordinates produced by the calibrated system. Note that we assume the user was looking exactly at the mouse pointer; in reality, the mouse pointer was somewhere within approximately 10-15 mm of the animated target the user was asked to watch. This suggests that system accuracy is actually slightly better than the results presented below, although how much cannot be quantified.

Percentage of all predictions	**Within radius of target, centimetres (~approx. pixels on typical screen)**
50	1.5cm (~90px)
75	2.5cm (~150px)
80	2.9cm (~170px)
90	3.9cm (~225px)
95	5.1cm (~300px)

Errors are not highly correlated over time, which allows median filtering to remove most erroneous predictions. For example, errors beyond the 80th percentile could be largely eliminated using a median filter of the last 5 measurements. Depending on application need and computer performance, the system operates at between 12 and 25 frames / second, so this median filtered output would still allow response times of 200 ms or less. The results shown in this document have no filtering, smoothing or other postprocessing applied.

Prediction errors had very similar distributions in both x and y dimensions suggesting that the system has similar accuracy in both. We compared the distributions of prediction errors for low and high resolution images separately, but did not observe any substantial differences. This suggests that low camera resolutions of 640x480 pixels are adequate (indeed, probably preferable given reduced noise sensitivity).

Ongoing algorithm improvements and more careful data capture will produce higher accuracy.

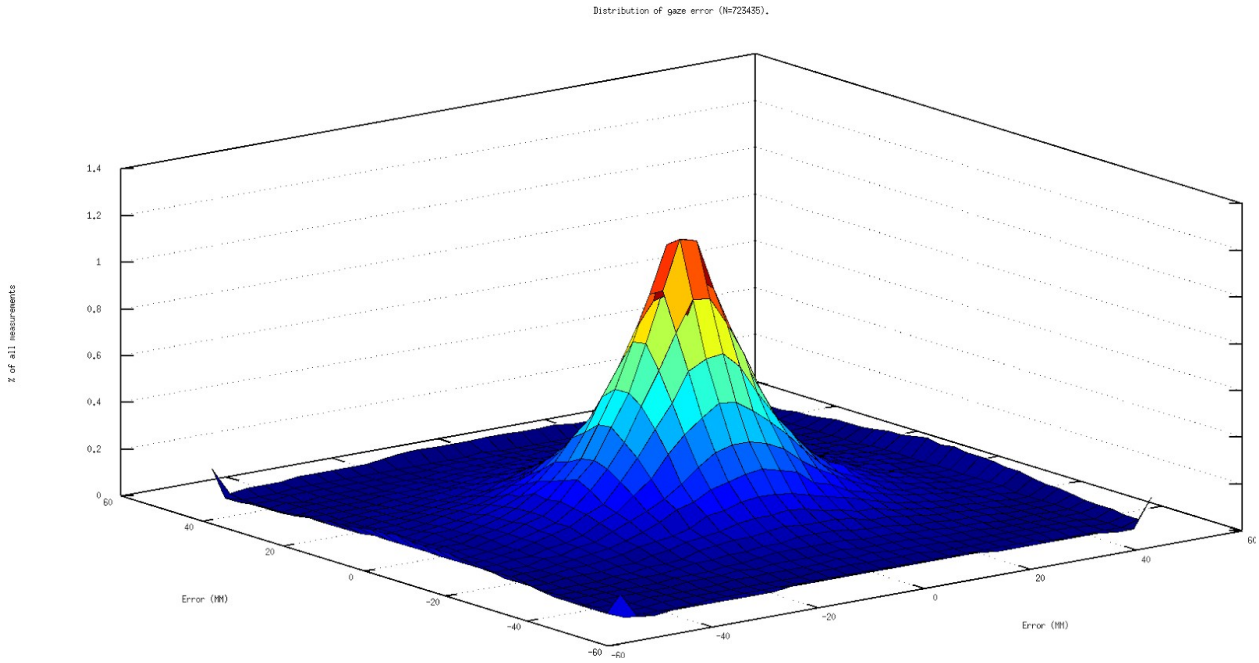

Figure 3: 2-D distribution of gaze prediction errors, in MM. This plot summarizes all images captured during user clicking (approx 1M). Predicted gaze coordinates are tightly clustered within approx. 20mm of the target positions users were asked to look at. ***All errors are visible****: errors beyond the range of the plot are clamped to the boundary, producing the small increases in the corners. It is therefore noticeable that very few gross errors occur beyond these limits.*

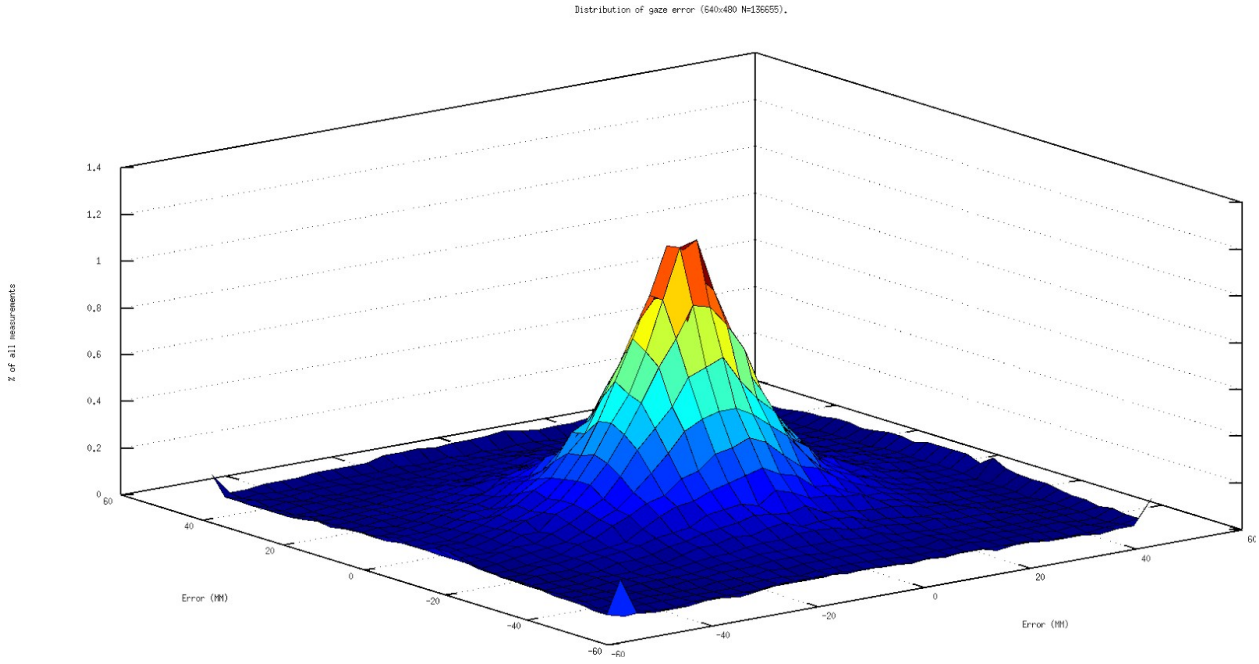

Figure 4: 2-D distribution of gaze prediction errors, in MM, for all images captured at low-resolution (640x480 pixels). There is no clear deterioration in accuracy due to the lower capture resolution.

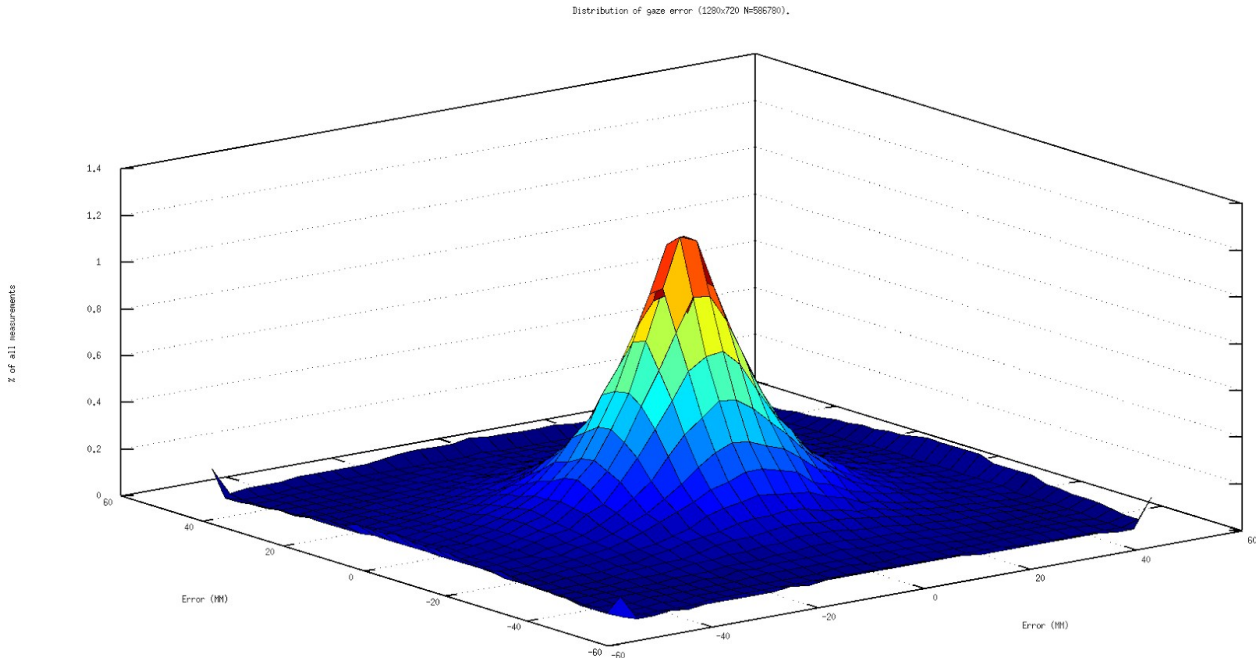

Figure 5: 2-D distribution of gaze prediction errors, in MM, for all high resolution images (1280x720 pixels). There is no obvious improvement, suggesting that the limiting factor is not camera resolution.

3.2 Bias Across Screen

We have plotted mean distance between predicted and measured mouse coordinates for each patch of 30 x 30 pixels in a virtual screen that encompasses all our actual hardware screens. The purpose of this plot is to reveal any reduced accuracy bias in specific screen areas. However, no such bias was observed. The system appears to predict gaze equally well across the screen.

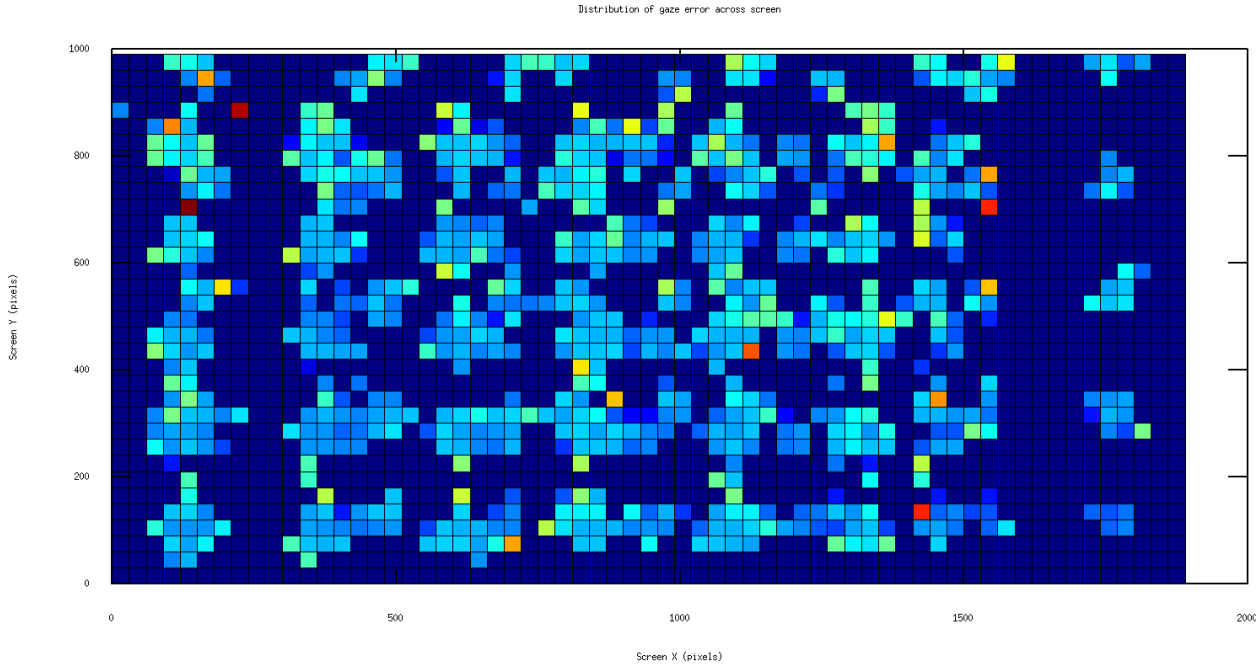

Figure 6: Mean gaze prediction error across the screen. Hotter colours indicate greater mean distance between predicted and mouse coordinates. We do not observe any clear deterioration in performance in any part of the screen. The clusters apparent in the figure are caused by the fixed grid layout of the '+' marks. The smaller clusters in the far right column are due to a wide-format screen being used for only a subset of participants.

3.3 Participants

A total of 300 participants were obtained for the study. The majority of participants were obtained by asking companies for permission to approach their employees in office or workplace environments. However, approximately 50% of participants were obtained by randomly requesting members of the public to work with our laptops. In the latter category we were unable to properly control lighting and other environmental characteristics, resulting in very challenging circumstances for controlled video capture (indoor and outdoor). In addition to difficult lighting, participants outdoor typically rested the laptops on their legs, resulting in varying and extreme poses throughout the process, in comparison to use of a laptop on a desk.

We present the following breakdown of participants' characteristics, where known, to evidence generalization across variations in human appearance due to age and ethnicity.

Demographic information (such as age) was not available for all participants, and was estimated from appearance in these cases. Country of origin was requested for some participants but not all; however, we can state that at least one participant self-reported originating in each of the countries marked below.

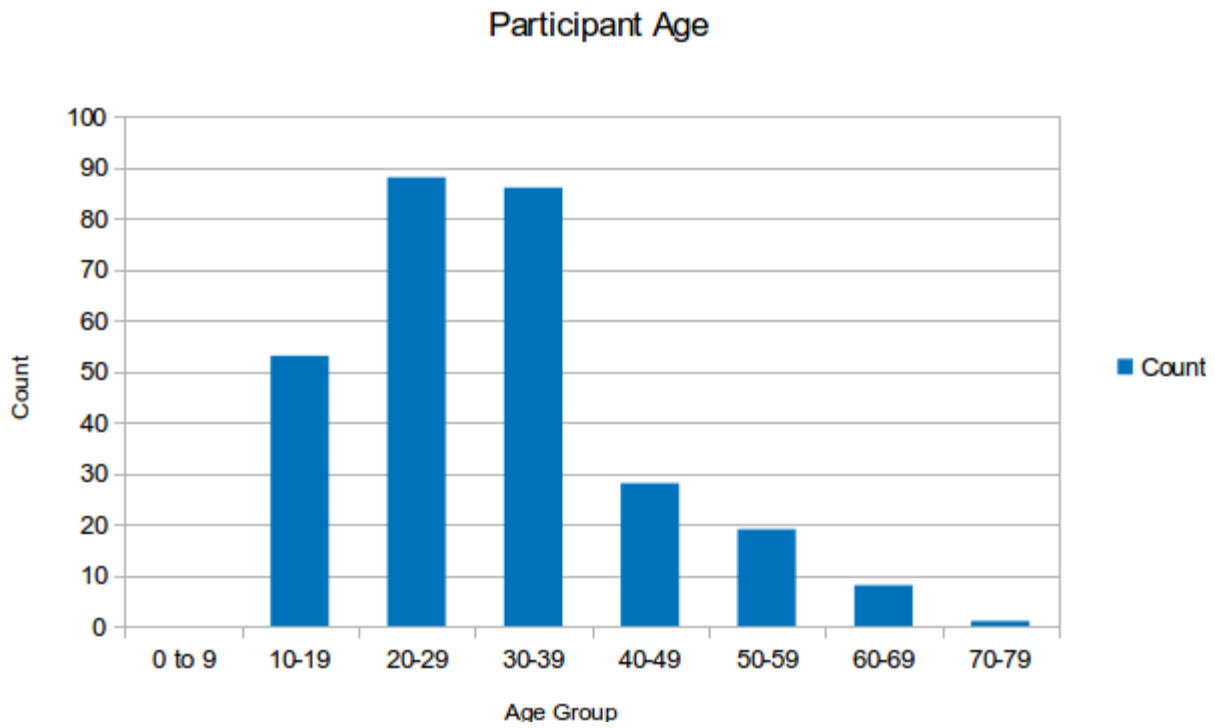

Figure 7: Histogram of participant age. Age affects participant appearance and therefore could have an impact on accuracy. We attempted to sample a wide range of age groups.

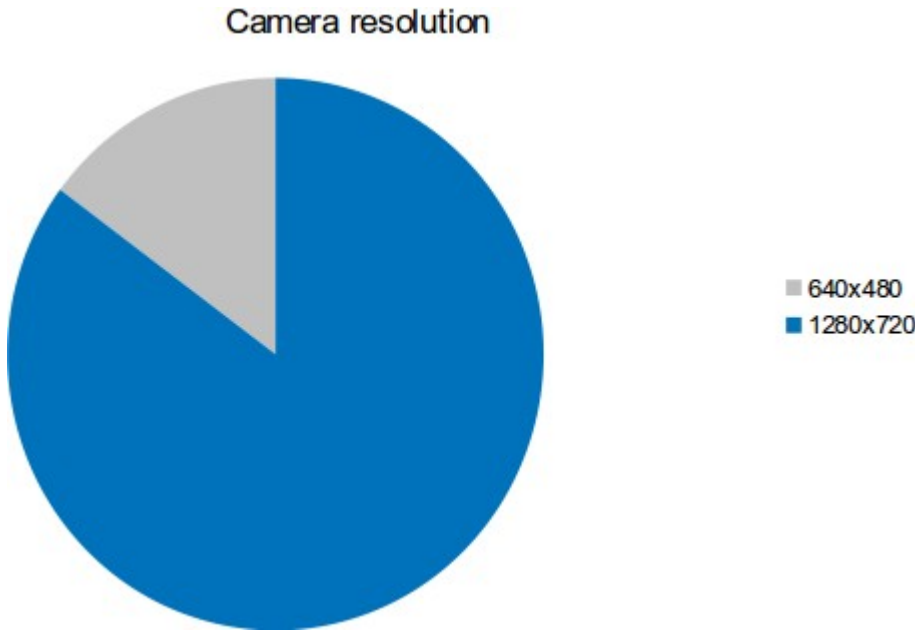

Figure 8: Pie chart showing proportions of subjects imaged at low and high resolutions. Resolution was later determined not to substantially affect results.

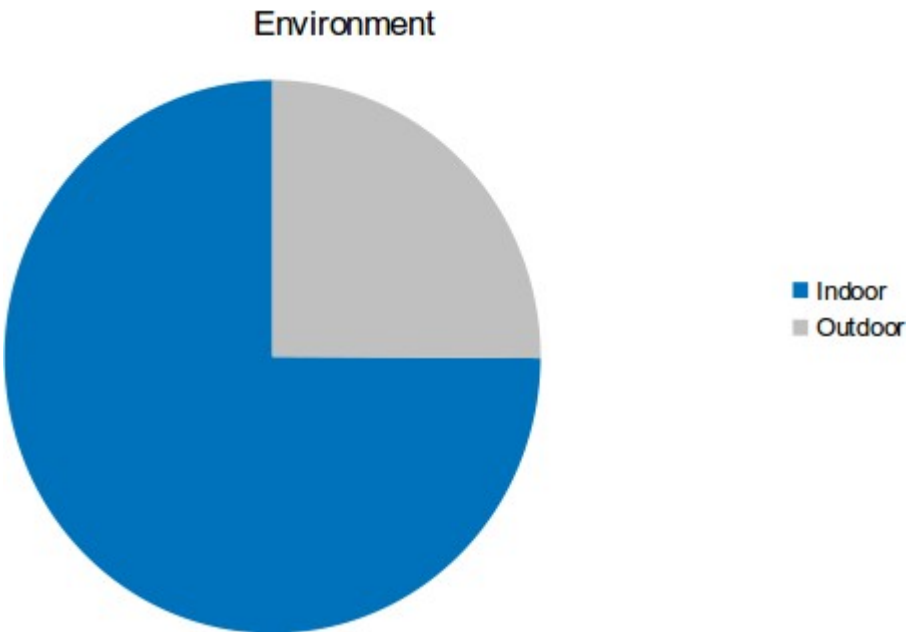

Figure 9: We found it impractical to bring all participants to an indoor environment with controlled lighting. Therefore, some participants were imaged in an outdoor environment using daylight. This did not prevent successful gaze tracking.

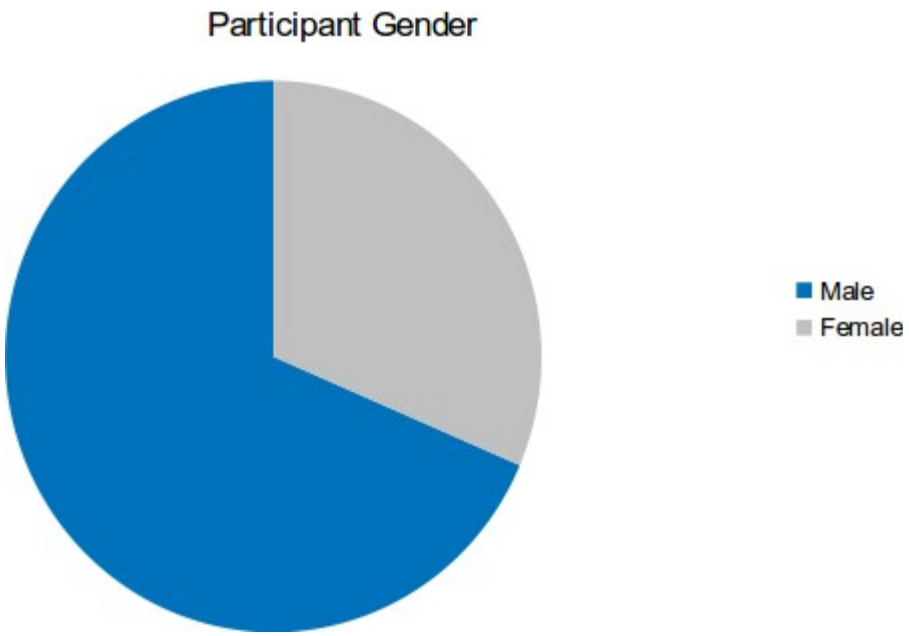

Figure 10: Although an equal number of participants from both genders was sought, collection was eventually heavily biased towards males.

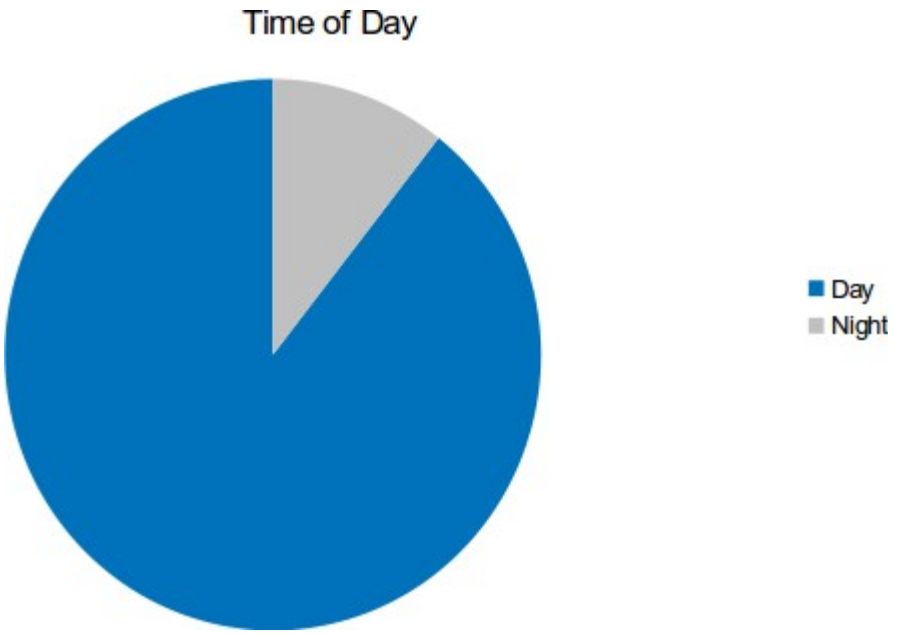

Figure 11: Some participants provided data captured at night in an indoor, artificially lit environment.

Figure 12: Participant volunteered country of origin, where known. Map based on outline image from WorldAtlas.com.

3.4 Exclusions

7 participants were excluded because their faces were not fully within the view of the camera.

4. Conclusions

The intent of this study was to quantify the accuracy of the xLabs Eye/Gaze tracking system, in realistic conditions, and on sufficient scale to validate generalization across the gamut of human appearance and variation.

The results show that eye-gaze prediction can be achieved accurately and robustly in uncontrolled circumstances, day and night, indoor and outdoor. We have tested the system on participants having a wide range of ages and ethnicities, both factors that could affect face-feature localization and thereby accuracy. We did not observe any difficulties with specific ages or ethnicities. Our data captures a wide range of skin and eye colouration, face-shape and other variables.

Given the difficult environments and limited sample data of each participant, it is reassuring that this study produced successful results. Gaze was predicted to within 2.9cm of the target coordinate on screen for 80% of all images of all subjects in all conditions. Some part of this error is due to the subject not reliably looking at the coordinates requested, or failing to look at the very centre of the target. The target itself had a size of approximately 1.4cm, and participants were likely looking anywhere in this region.

In more controlled circumstances, with better lighting and more limited poses (e.g. subjects sitting at a desk, with the laptop placed on the desk) it is likely that accuracy would be increased further.

Camera resolution did not substantially affect results. This shows that even low resolution cameras are adequate for accurate gaze-tracking. We did not detect any bias towards better or worse performance in specific areas of the screen. Within the range tested, the make and model of computer and camera had no effect on results.