Usability Evaluation of Eye Tracking on an Unmodified Common Tablet

Corey Holland
Department of Computer Science
Texas State University
San Marcos, TX 78666 USA
ch1570@txstate.edu

Jose Cruz
Department of Computer Science
Texas State University
San Marcos, TX 78666 USA
jc2018@txstate.edu

Atenas Garza
Department of Computer Science
Texas State University
San Marcos, TX 78666 USA
af1284@txstate.edu

Oleg Komogortsev
Department of Computer Science
Texas State University
San Marcos, TX 78666 USA
ok11@txstate.edu

Elena Kurtova
Department of Computer Science
Texas State University
San Marcos, TX 78666 USA
ek1106@txstate.edu

CHI 2013 Extended Abstracts, April 27–May 2, 2013, Paris, France.
ACM 978-1-4503-1952-2/13/04.

Abstract
This paper describes the design, implementation, and usability evaluation of a neural network based eye tracking system on an unmodified common tablet and discusses the challenges and implications of neural networks as an eye tracking component on a mobile platform. We objectively and subjectively evaluate the usability and performance tradeoffs of calibration, one of the fundamental components of eye tracking. The described system obtained an average spatial accuracy of 3.95° and an average temporal resolution of 0.65 Hz during trials. Results indicate that an increased neural network training set may be utilized to increase spatial accuracy, at the cost of greater physical effort and fatigue.

Author Keywords
Eye tracking; Mobile device; Neural network; Usability

ACM Classification Keywords
B.4.2 [Input/Output and Data Communications] Input/Output Devices; I.4.8 [Image Processing and Computer Vision]: Scene Analysis – Tracking

General Terms
Design, Experimentation, Human Factors, Performance.

Figure 1: Eye tracker evolution.

Introduction

Eye tracking is an increasingly common input medium in the ever-growing field of human-computer interaction. For decades, eye trackers have provided a fast, stable, and non-intrusive means of communication for disabled users incapable of using the more standard keyboard and mouse [4, 6]. As advances in recent years (shown in Figure 1) continue to improve the cost, accuracy, and portability of eye tracking systems, eye trackers are becoming an ever more integral part of the daily routine of many [3].

A major step in this direction is the recent application of video-oculography based eye tracking techniques on the common web camera [1, 11]. This has allowed an unparalleled advance in the cost and portability of eye tracking systems. Mobile devices such as cell phones and tablet PCs are potential candidates for the application of such techniques, where eye tracking can serve as an additional input channel for able users and as a primary input for disabled users; however, a practical eye tracking solution on such devices is extremely challenging and has not yet been achieved.

The usability of an eye tracking system is primarily dependent on two aspects: spatial accuracy (the difference between the reported and actual point of gaze) and temporal resolution (the rate at which the system captures usable gaze locations) [12]. Spatial accuracy has been the primary focus of continued research [10], as temporal resolution is typically a non-issue for commercial eye tracking systems which are optimized at the hardware level for just such tasks.

Mobile devices are at a distinct disadvantage in both cases, as the typically sub-standard camera and processing power are limiting factors for the application of an eye tracking system. For instance, the iPad 2 is equipped with a web camera capable of capturing 30 frames per second, limiting the maximum temporal resolution of an eye tracking system to 30 Hz. In comparison, commercial eye tracking systems are generally able to achieve a consistent temporal resolution of 60 Hz – 1000 Hz.

While face detection on an unmodified tablet [2] and eye detection on an unmodified cell phone [7] has been achieved previously, the realm of eye tracking is as of yet completely unexplored. In this paper, we present an eye tracking solution on an unmodified mobile platform and the usability evaluation of the described system during its calibration procedure.

Design & Implementation

The described system functions primarily through software, utilizing techniques originally explored in [11], and employing a standard web camera, as shown in Figure 2. A video frame is retrieved from the camera, converted to an appropriate format, and passed to the image recognition module. The image recognition module consists of face detection, eye detection, and iris detection sub-modules, each of which attempts to identify its respective features within a sub-region of the image. The neural network module retrieves the final sub-region identified by the image recognition module, extracts key features of the eye from the image, and passes them into a neural network, mapping the feature set to a position on the screen.

Image Recognition

The image recognition module consists of three distinct sub-modules: face detection, eye detection, and iris

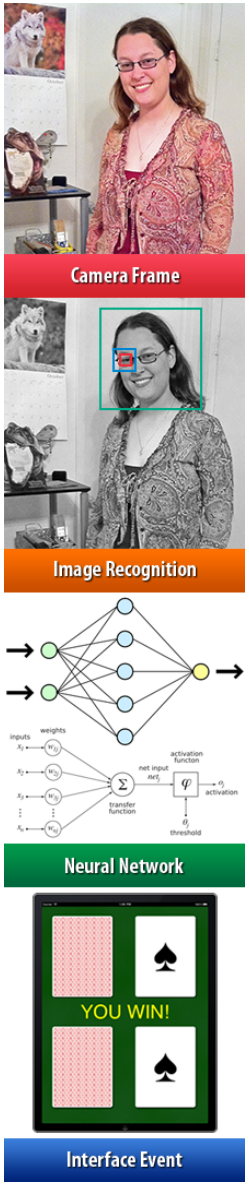

Figure 2: Eye tracking process.

detection. The face and eye detection sub-modules employ Haar classifiers, a method of object detection originally developed for facial recognition. The classifier is trained with a set of features and is able to quickly reject false regions of the image, reducing the overall search time. The iris detection sub-module makes use of visual template matching, sequentially comparing overlapping regions of the image. To reduce computation, the size and resolution of the original image is reduced by ½ and converted to gray-scale.

Each sub-module passes the size and coordinates of the identified image segment to the following sub-module, reducing the search area. To improve the speed of image recognition algorithms, the previously identified image segment is maintained for each sub-module, and is used as a starting point when searching each frame.

By steadily narrowing the search field from the easiest to identify to the most difficult, the ability to accurately identify the eye is improved. Unfortunately, with the limited processing power available to tablet PCs, each stage of image recognition has a substantial impact on the overall performance. Therefore, the temporal resolution of the eye tracking system is largely dependent on the image recognition module. For this reason, the image recognition sub-modules were designed to function independently, allowing "plug-in" functionality for the various stages of image recognition.

The eye detection sub-module is vital to the operation of the eye tracking system; however, it is possible to remove the face and iris detection sub-modules with the application of simple heuristics. In this paper, the face detection sub-module was inactive to ensure faster processing, while the eye and iris detection sub-modules remained active.

Neural Network

The neural network module, composed of a single perceptron, is responsible for mapping the image sub-region identified by the image recognition module to a position on the screen. The common analogy is that artificial neural networks are designed to mimic the ability of biological neural networks to recognize and identify patterns [9]. In practice, neural networks are often used for function approximation, mapping a set of input values to the appropriate set of output values.

The coordinates of an image sub-region are provided by the image recognition module. If the iris detection sub-module is active, this is given by the closest template match within the eye sub-region, otherwise coordinates are provided by the eye detection sub-module to the center of the eye. An image segment of fixed size is formed around the center of the coordinates. With or without the iris detection sub-module, the segment must be large enough that a fraction of the iris is contained within the image.

The brightness of each pixel within the segment is used as an input value to the neural network, giving a number of input neurons equal to the number of pixels in the image segment, and two output neurons for the screen coordinates.

Then, the accuracy of the eye tracking system is largely dependent on the neural network module, assuming a valid image segment is provided by the image recognition module.

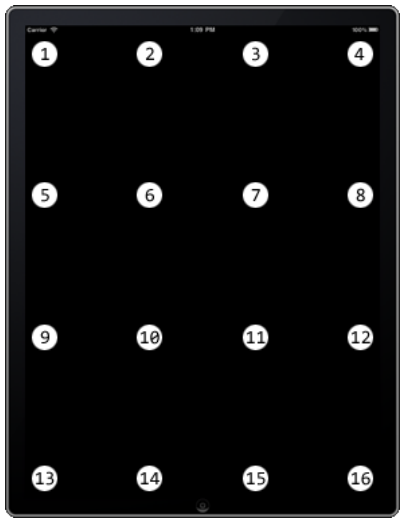

Figure 4: 4×4 calibration grid.

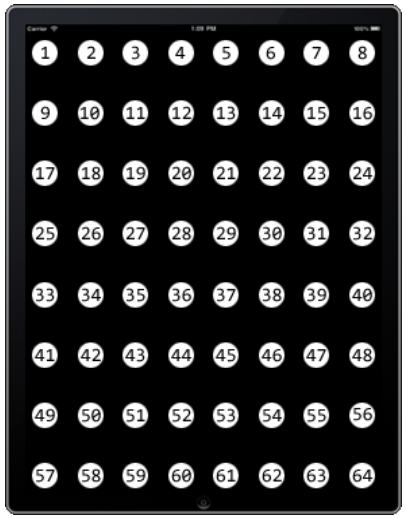

Figure 3: 8×8 calibration grid.

Methodology

To evaluate the proposed eye tracking system, a proof-of-concept application was developed, using the Open Source Computer Vision library (OpenCV) for image recognition and the Fast Artificial Neural Network library (FANN) for the neural network [5, 8].

Apparatus

The experiment was conducted on the Apple iPad 2, with a 1 GHz dual-core processor, 512 MB memory, and a screen resolution of 1024×768. The built-in front camera has a resolution of 0.3 megapixels and is capable of video capture at 30 frames per second. A portable easel was used to secure the iPad in a fixed, upright position to establish a reliable performance baseline.

Participants

A total of 21 subjects volunteered for the experiment and provided informed consent. Participant ages ranged from 19 to 34 years old (M = 26.14, SD = 3.79), of which there were 12 male and 9 female. Among these, 11 had normal vision and 10 required corrective lenses. Of the participating volunteers, 18 provided usable data and 3 were unable to complete the experiment. For those unable to complete the experiment, failure occurred in the eye detection sub-module due to occlusion by the eyelid or otherwise non-typical eye shape.

Procedure

The tablet was secured on an easel with the front-facing camera at approximately eye-level. Participants were seated 20.0 inches (±1.9 inches) from the camera and were allowed a free range of head motion. Each participant was given 5 minutes to complete calibration and 5 minutes to complete verification, for a maximum task completion time of 10 minutes.

During calibration, a series of dot stimuli were presented on the screen in an evenly spaced n×n grid covering the entire screen. Each stimulus point remained visible until 2 valid image samples were collected for the point, the stimulus point was then hidden, and the next stimulus point in the sequence was presented. Training of the neural network occurred incrementally for each stimulus sample collected during both calibration and verification. Accuracy calculations were performed during verification, while the sampling rate was measured during both calibration and verification.

We hypothesize that an increased neural network training set will enhance the spatial accuracy of the system at the cost of overall system usability. To examine the effect of the neural network training set on system performance and usability, two calibration schemes of varied size were compared. The first calibration set employed a 4×4 grid, shown in Figure 3, for a neural network training set of 32 samples, and the second calibration set employed an 8×8 grid, shown in Figure 4, for a neural network training set of 128 samples. The calibration schemes were alternated evenly, such that 9 participants performed calibration on each.

After each trial, participants were asked to complete a usability survey as suggested by the ISO 9241-9 standard. The survey contained 6 questions, each rated on a 7-point Likert scale, with 7 as the most favorable response, 4 as the mid-point, and 1 as the least favorable response. Survey categories included: ***1)*** *General comfort;* ***2)*** *Shoulder fatigue;* ***3)*** *Neck fatigue;* ***4)*** *Eye fatigue;* ***5)*** *Physical effort;* ***6)*** *Mental effort*.

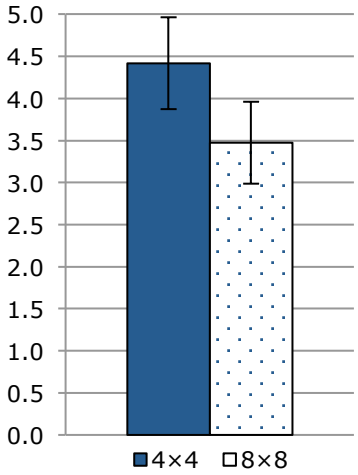

Figure 6: Average spatial accuracy (deg).

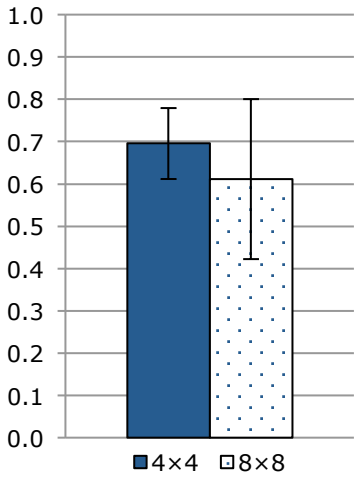

Figure 5: Average temporal resolution (Hz).

Results

Spatial Accuracy

The spatial accuracy is calculated as the error in degrees of the visual angle between the stimulus position and the position calculated by the neural network. Shown in Figure 5, the average spatial accuracy of the eye tracking system was 4.42° (±0.55) or 203.85 pixels (±17.02) for the 4×4 calibration set, and 3.47° (±0.49) or 157.91 pixels (±21.20) for the 8×8 calibration set. These results were statistically significant and indicate that on average the increased neural network training set provided a 27% improvement in spatial accuracy ($t(8) = 4.72$, $p < 0.001$).

Temporal Resolution

The temporal resolution is calculated as the usable frames processed by the neural network per second. Shown in Figure 6, The average temporal resolution of the eye tracking system was 0.70 Hz (±0.08) for the 4×4 calibration set, and 0.61 (±0.19) for the 8×8 calibration set. Consequently, the average task completion time was 1.55 minutes (±0.18) for the 4×4 calibration set, and 7.70 minutes (±2.72) for the 8×8 calibration set. These results were not statistically significant, indicating that the temporal resolution was unaffected by the size of the calibration set ($t(8) = 1.46$, $p = 0.092$).

Usability Evaluation

According to the subjective evaluation, shown in Figure 7, there was little or no perceived difference between the two calibration sets. There was no statistically significant difference between the scores in any category ($p > 0.05$); however, there was a noticeable tendency for the subjective scores of the 4×4 calibration set to average slightly more favorable responses than the 8×8 calibration set.

Discussion & Future Research

In comparison to commercial eye tracking systems, the specifications achieved by the proposed system are underwhelming, and would obviously not be suited for high-quality data collection. Despite this, the proposed system is capable of distinguishing on-screen objects presented at a low resolution (screen quadrants, for example), which may find uses in attentional studies and engagement analysis. Further, the price and portability of the system provide an attractive alternative for the typical consumer.

System	Accuracy	Resolution	Portable	Cost
EyeLink 1000	0.5°	1000 Hz	No	$10000+
Tobii TX300	0.5°	300 Hz	No	$10000+
Sony HMZ-T1	1.0°	90 Hz	Yes	$10000+
iPad 2	2.8°	0.87 Hz	Yes	$500

As was stated previously, the spatial accuracy of the eye tracking system is largely dependent on the neural network module, while the temporal resolution is largely dependent on the image recognition module. The current work has shown that an increased neural network training set is a reliable method to enhance the spatial accuracy of the neural network; however, a large training set becomes impractical with such a low temporal resolution, as the time required to complete calibration increases drastically with each point added.

Improving the temporal resolution, then, will allow an increased neural network training set by reducing the overall time required for calibration and, consequently, user fatigue. Areas of further research are likely to include: optimization of image recognition and processing algorithms, improvements in the heuristics used to estimate eye position, more advanced methods of image

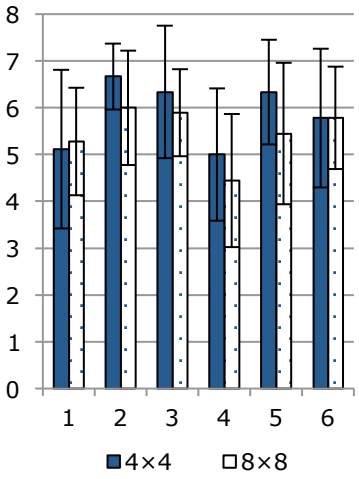

Figure 7: Questionnaire results.

manipulation and machine learning, and the possible removal of both the face and iris detection sub-modules from the image recognition module.

Conclusion & Acknowledgement

This paper has described the design and implementation of a neural network based eye tracking system on an unmodified common tablet, and objectively and subjectively evaluated the effects of the neural network training set on eye tracking performance and user satisfaction. The described system obtained an average spatial accuracy of 3.95°, maximum spatial accuracy of 2.82°, average temporal resolution of 0.65 Hz, and maximum temporal resolution of 0.87 Hz during trials. Source for the Neural Network Eye Tracker (NNET) is available at: http://cs.txstate.edu/~ok11/nnet.html.

Based on the results, it is clear that an increased neural network training set may be utilized to increase spatial accuracy. Unfortunately, calibration time will increase linearly with the size of the training set, increasing user effort and fatigue; however, temporal resolution remains relatively stable, as the increased number of calculations does not affect the rate at which calculations are performed.

This work was supported in part by Texas State University, the National Science Foundation CAREER Grant #CNS-1250718 and GRFP Grant #DGE-1144466, and the National Institute of Standards Grants #60NANB10D213 and #60NANB12D234.

References

[1] Agustin, J.S., Skovsgaard, H., Hansen, J.P. and Hansen, D.W. Low-cost gaze interaction: ready to deliver the promises. In *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*, ACM (2009), 4453-4458.

[2] Allan, A. Face Detection. Face Detection. *iOS Sensor Programming*, O'Reilly, 1-320.

[3] Duchowski, A. A breadth-first survey of eye-tracking applications. *Behavior Research Methods 34*, 4 (2002), 455-470.

[4] Hutchinson, T.E., White, K.P., Martin, W.N., Reichert, K.C. and Frey, L.A. Human-computer interaction using eye-gaze input. *IEEE Transactions on Systems, Man, and Cybernetics 19*, 6 (1989), 1527-1534.

[5] Intel. Open Source Computer Vision (OpenCV). (1999).

[6] Majaranta, P. and Räihä, K.-J. Twenty years of eye typing: systems and design issues. In *Proceedings of the 2002 symposium on Eye tracking research & applications*, ACM (2002), 15-22.

[7] Miluzzo, E., Wang, T. and Campbell, A.T. EyePhone: activating mobile phones with your eyes. *Proceedings of the second ACM SIGCOMM workshop on Networking, systems, and applications on mobile handhelds* (2010).

[8] Nissen, S. Fast Artificial Neural Network (FANN). (2002).

[9] Nissen, S. Neural Networks Made Simple. *Software 2.0 2*, (2005), 14-19.

[10] Rayner, K. Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin 124*, 3 (1998), 372-422.

[11] Sewell, W. and Komogortsev, O.V. Real Time Eye Gaze Tracking With an Unmodified Commodity Webcam Employing a Neural Network In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, ACM (2010), 1-6.

[12] Zhang, X. and MacKenzie, I.S. Evaluating eye tracking with ISO 9241 - part 9. In *Proceedings of the 12th international conference on Human-computer interaction: intelligent multimodal interaction environments*, Springer-Verlag (2007), 779-788.