Development of an untethered, mobile, low-cost head-mounted eye tracker

Elizabeth S. Kim*
Adam Naples
Giuliana Vaccarino Gearty
Quan Wang
Seth Wallace
Carla Wall
Michael Perlmutter
Fred Volkmar
Frederick Shic**
Yale University Child Study Center

Linda Friedlaender
Jennifer Kowitt
Yale Center for British Art

Brian Reichow
University of Connecticut Health Center

Abstract

Head-mounted eye-tracking systems allow us to observe participants' gaze behaviors in largely unconstrained, real-world settings. We have developed novel, untethered, mobile, low-cost, lightweight, easily-assembled head-mounted eye-tracking devices, comprised entirely of off-the-shelf components, including untethered, point-of-view, sports cameras. In total, the parts we have used cost ~$153, and we suggest untested alternative components that reduce the cost of parts to ~$31. Our device can be easily assembled using hobbying skills and techniques. We have developed hardware, software, and methodological techniques to perform point-of-regard estimation, and to temporally align scene and eye videos in the face of variable frame rate, which plagues low-cost, lightweight, untethered cameras. We describe an innovative technique for synchronizing eye and scene videos using synchronized flashing lights. Our hardware, software, and calibration designs will be made publicly available, and we describe them in detail here, to facilitate replication of our system. We also describe novel smooth-pursuit-based calibration methodology, which affords rich sampling of calibration data while compensating for lack of information regarding the extent of visibility on participants' scene recordings. Validation experiments indicate accuracy within 0.752 degrees of visual angle on average.

CR Categories: H.1.2 [Information Systems] User/Machine Systems--Human information processing, Human factors; J.4 [Social and Behavioral Sciences] Psychology

Keywords: Head-mounted displays and systems, low-cost systems, tools for eye tracking analysis

1 Introduction

Eye tracking can provide information about an individual's cognitive state and about the nature of atypical attentional processes in individuals with neuropsychiatric conditions. In contrast to table-mounted or remote systems, head-mounted eye-tracking systems allow participants to move, extending gaze tracking from constrained stimuli in controlled environments, and out into the real world. Unfortunately, commercially developed head-mounted eye trackers tend to be expensive, costing $10,000-$40,000 per unit. The cost is likely generated by extensive testing, research, and development. For social research in particular, study designs may necessitate eye tracking of multiple participants, making the high price of commercial eye trackers especially prohibitive.

As an alternative to commercial devices, several groups have developed custom-built head-mounted eye-trackers. Unfortunately, to our knowledge, previous low-cost, portable eye-tracking devices have required custom manufacture of specialized parts, limiting accessibility [Järvenpää and Äyräs 2010]; have been tethered to a laptop or a backpack, encumbering or restricting users [Babcock and Pelz 2004; Franchak et al. 2010; Hayhoe and Ballard 2005; Li et al. 2005; Lukander et al. 2013; Noris et al. 2011], or have not been reported to have high tracking accuracy [Rantanen et al. 2011]. To date, there has existed no device that is simultaneously accessible, untethered, and highly accurate—three criteria for affordable, real-world eye tracking.

This paper describes our efforts to create an untethered, head-mounted eye tracking system which is affordable, usable in real-world settings, and functionally equivalent (that is, as accurate as) a commercial system. We discuss hardware design, video-based gaze estimation software design, and the calibration and operation protocols we have developed to create our system. We also describe empirical validation of acceptable accuracy in point-of-regard estimation.

Our system costs approximately $153 in easily obtainable parts and can be assembled and used with minimal technical training. Our selection and modification of components, and our software development are all tuned to the specific formats and shortcomings of our particular equipment. Including our freely available software, our system manifests a highly accessible, highly accurate eye-tracking device. While ours is less robust than a commercial system, we have begun using it successfully in studies of visual attention in community settings, including classrooms and museums, with adults both with typical development and with autism spectrum disorders.

2 Device design

2.1 Cameras and storage media

We use two cameras in our design; one (*eye camera*) pointed at the wearer's eye, to detect the pupil; and the other (*scene camera*) capturing the wearer's point of view. We selected Veho's Muvi Atom camera ($60-$90, 30g or 1.05oz), which records video at a variable frame rate around 30 frames per second (fps), in VGA (640 x 480 pixel) resolution, in RGB 8-bit color, with automatic (and no manual) exposure adjustment, at an internally adjustable focal length. To compensate for observed shortcomings of the cameras, we developed extensive optical, electronic, mechanical, and software solutions. Although the parameters of our solutions

* elizabeth.kim@yale.edu
** frederick.shic@yale.edu

ETRA 2014, March 26 – 28, 2014, Safety Harbor, Florida, USA.

ACM 978-1-4503-2751-0/14/03

Figure 1: *Our eye-tracking system costs approximately $153 in off-the-shelf parts. The design centers around a pair of small, lightweight, point-of-view sports cameras, Veho's Muvi Atom model, which record eye and scene videos.*

are specific to Veho Muvi Atom cameras, we expect that the majority of the challenges we confronted are shared by similar low-cost cameras. Therefore, we describe our solutions in detail.

Veho Muvi Atom cameras record to high-capacity secure digital memory cards. We used Kingston Class10 (10MB/sec minimum write speed for unfragmented space) 8GB cards. Muvi Atom cameras can be operated in "webcam mode" by USB connection to a Windows PC, allowing us to use visual feedback as we adjust the eye camera. Unfortunately, these cameras do not allow us to monitor output while recording.

Muvi Atom cameras can operate for approximately 30 minutes on a fully charged, internal, rechargeable battery. A full charge takes 1-2 hours to complete, via USB connection. We experienced a high failure rate among our fleet of Veho Muvi Atom cameras (13/23 became inoperable over 18 months) due to eventual failure to recharge.

Eye Camera: We removed the built-in infrared (IR)-blocking filter and added visible-light-blocking, IR-passing filters (type 87) to eye camera lenses to limit their images to IR. We use intensity-based, dark pupil detection to perform point-of-regard (POR) estimation. Dark-pupil detection depends on high contrast between the iris and pupil, which is vastly enhanced by illuminating the eye using an IR LED lamp. Our lamp was powered by direct connection to a CR2032 lithium ion coin cell battery.

We found that prescription eyeglasses frequently specularly reflected the natural light, including IR reflections that obscured the pupil image from eye camera capture. To control IR illumination we shielded the eye camera from ambient light using a short, wide plastic straw painted matte black (see Figure 1).

Scene Camera: To increase the field of view (FOV) of the cameras, we adhesively mounted a wide 180° FOV lens to the scene camera lens.

2.2 Synchronization lights

To compensate for variable frame rates, which are common among low-power, small, low-cost, untethered cameras, we developed a recurring visual clapboard by which to synchronize eye and scene image streams. We designed synchronized blinking light circuits (*sync lights*), which blink with constant frequency (on for ~1s, with a ~3s period) simultaneously in view of the eye and scene cameras. The sync light LEDs are infrared, so as not to distract the user, and are powered by a shared, simple, monostable 555 timing circuit [Lowe]. The sync light timing circuit is powered using a single CR2032 lithium ion battery. We use 950nm infrared, 1.4V low-power LEDs.

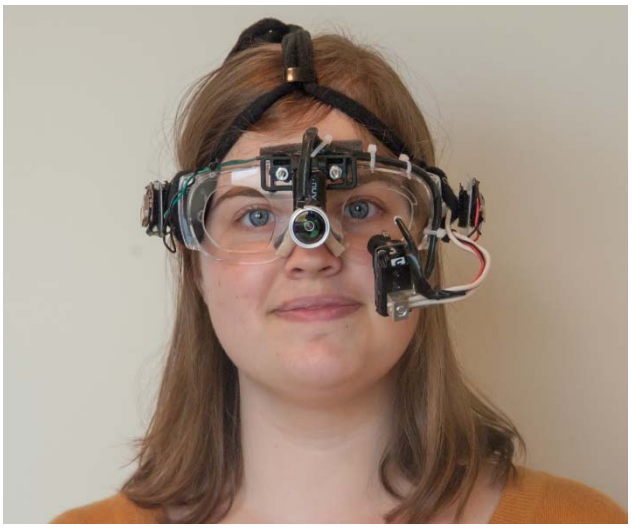

Figure 2. *Our device in use. The scene camera faces out from between the wearer's eyes; the eye camera faces her left eye.*

3 Frames

Our system is mounted on a pair of standard, laboratory safety glasses (the *frame)*. We removed the lenses from the frames using a Dremel rotary tool. We fixed the cameras' proprietary, plastic mounts to the forehead of the safety glasses (scene camera), and to a lightweight, low-gauge wire boom extending from the safety glasses' leg, and snap the cameras into their mounts (eye camera), such that the eye camera sits in front of and below the wearer's left eye. The scene camera sits above the bridge of the frame, facing the wearer's field of vision.

With two CR2032 coin cell batteries, used to power the eye lamp and sync lights, our devices weighed 224g (7.9oz), with the bulk of the weight in the modified cameras (~110g, 3.9oz). The devices had a tendency to rotate forward, pivoting and putting weight on the wearer's nose. To improve stability and comfort, we used cloth straps over the top and around the back of wearer's head, and lined the nose of the frames with felt padding.

4 Point-of-regard estimation software

4.1 Video stream extraction, synchronization, and variable frame rate

Veho Muvi Atom cameras encode video in Motion JPEG (M-JPEG) format, within audio-video interleave (AVI) wrappers, at variable frame rates (~30fps). We used two methods to extract image streams from the eye and scene videos. In the first method, we used VirtualDub, a popular video processing software tool to export image streams, VirtualDub correctly estimated the video frame rate to be approximately 30fps using popular, but due to Veho's apparently non-conventional M-JPEG encoding, could not exactly determine varying frame rates. No other popular video playback and processing tools (including QuickTime, VLC, MediaPlayerClassic, and Mpeg-Streamclip) successfully detected exact varying frame rates.

We initiated eye-scene temporal alignment using the rough audio-visual alignment produced by off-the-shelf video playback software. During all data collection phases, shortly after beginning both eye and scene camera recordings, we produced a distinct

sound (usually a codeword) for the purpose of this rough alignment. At this rough initial alignment phase in image processing, we played back each video and listen for the coarse-grained fiducial sound in both the eye and stream videos, note the times at which these sounds occur, and declare these to be (roughly) the start of alignment. Because audio-visual synchronization within an eye or scene recording has error up to 100ms, we can be sure that the audio-based initial alignment has an error within 200ms, or approximately 7 frames. After initial, rough, audio-based temporal alignment, we then use sync light, fine-grained fiducials to improve the precision of this rough alignment.

We detected fiducial frames by manually selecting a rectangular region of the eye and scene image, within which the sync light was constantly positioned throughout each image stream. We summed the intensity over all pixels and all color channels within the sync light region. We estimated the sync light period to be the mode of the differences between all fiducials' frame numbers, and manually added or removed fiducials which were missing or added by thresholding. Error in fiducial detection tends to be higher in scene streams, because whereas the eye image remains under nearly-fixed illumination conditions (unless the participant turns his head to face a bright source of IR light such as a window), scene camera illumination can change under any visible light changes in the participant's environment. For instance, if a participant turns his head from a dark curtain to face a bright computer screen, the camera takes several frames to automatically adjust exposure to the brighter scene; frames written during this adjustment will have much higher intensity than those immediately preceding. In addition, whereas the background to the eye camera's sync light is fixed to be the wearer's face, behind the scene camera's sync light is the world at large. Therefore, total intensity in the scene stream's sync light's detection region is much higher than in that of the eye stream's, leading to greater threshold-based fiducial detection error in the scene stream.

Under the assumption of temporal alignment of all frames subsequent to fiducial frames, we frequently found misalignment at the end of each sync-light cycle. After manually inspecting and correcting statistically outlying differences between detected sync light onsets, we then truncated the longer stream, to maintain alignment, and to realign eye and scene image streams at the next fiducial. Variability in frame rate tends to be low: in a typical dataset, we truncated 227 eye frames from a total of 69674 frames (0.33%) and 752 of 69571 scene frames (0.11%). We confirmed fiducial alignments by checking for rough alignment between eye and scene audio streams roughly corresponding to sync light fiducials.

After several recordings failed to produce any AVI files, we developed a data scraping tool which performs low-level data extraction from SD cards, allowing us to recover and reconstruct video recordings that have missing end-of-file AVI encoding information.

4.2 Pupil detection, pupil position estimation

We estimated pupil location using dark pupil detection, with an intensity threshold. We estimated the pupil position to be the centroid (based on area) of contiguous pixels in the detected pupil region (see Figure 3). This stage in point-of-regard estimation is exquisitely sensitive to image quality in the eye recordings. For instance, if the eye camera's IR lamp insufficiently illuminates the eye video, then intensity contrast between iris and pupil pixels tends to suffer.

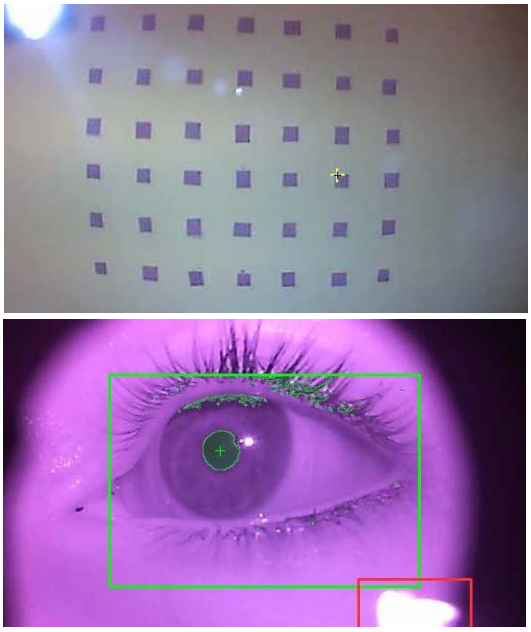

Figure 3. *Temporally aligned scene and eye images from validation experiment, viewed at a distance of 4 feet from stimuli. In the eye image the pupil detection and sync light regions are outlined, and the detected pupil and its centroid are highlighted.*

4.3 Point-of-regard estimation

We used a quadratic polynomial, in horizontal and vertical pixel coordinates, to estimate a map from pupil positions in the eye stream to point-of-regard positions in scene images, trained on known positions of calibration targets and their temporally aligned pupil position estimates.

5 Validation and performance

Using a 7x7 grid of stimuli (see Figure 3), we calibrated the device at distances of 2, 4, and 8 feet from the stimuli, with an average number of 3 samples at each calibration stimulus. We then tested the trained calibration on the same dense grid, at test distances of 3, 7, 9, and 10 feet from the stimuli, again using an average of 3 samples at each test position, producing a total of 517 test samples. The error (Euclidean distance) between the actual, known positions and estimated POR estimates was 0.752 degrees in visual angle on average (SD = 0.733 degrees).

For this dataset, sync light-based alignment required 2 and 18 manual corrections (of false positive and missed onsets of sync lights) out of 197 cycles, in eye and scene streams, respectively.

6 Real-world calibration and testing strategies

Given the short battery lives of our chosen cameras, we developed a rapid protocol for fitting devices to participants. This involved using webcam mode to position the device and check that sync lights were visible in each camera's field of view. In the case of the eye camera, we adjusted the camera's position and IR lamp to minimize pupil-obscuring reflections, to create spatially consistent luminance throughout the eye image, and to maximize contrast between iris and pupil. We periodically checked the cameras' indicator lights to confirm that they were recording throughout the procedure.

During testing in real-world settings, we use smooth pursuit of a target whose coverage of the camera's FOV gradually expands to the camera's limits. When using an unmonitored eye tracker in unconstrained environments, smooth pursuit overcomes challenges presented by calibration on discrete targets. Unconstrained, unmonitored eye tracking does not afford precise knowledge of what objects in the participant's view are also visible to the eye tracker, unlike table-mounted eye tracking over a constrained space, at a fixed distance from a participant's eyes. Table-mounted eye tracking allows us to control and guarantee our awareness of calibration stimuli's locations, whereas our system depends on a scene camera in unconstrained situations to provide calibration stimuli's locations. Eye tracking calibration is susceptible to systematic eccentric error. Therefore, more distal calibration stimuli improve calibration accuracy. However, because the cameras we have chosen do not offer real-time visual monitoring of the camera's field of vision, if we were to use discrete calibration stimuli in the distal corners of the participant's field of vision, we would risk placing the stimuli beyond the scene camera's FOV. Smooth pursuit calibration obviates this risk by allowing us to cover distal regions of the wearer's and scene camera's fields of vision, without necessitating constrained calibration environments or stimuli, which might interfere with applications requiring real world settings or activities.

Smooth pursuit calibration also affords the opportunity to collect data rich in spatial and temporal density. Numerous calibration samples are helpful especially in the face of noisy (e.g., due to low resolution) imaging, unreliable recording (e.g., frames badly encoded by an inexpensive camera processor), and limited precision in temporal alignment of eye and scene data (e.g., due to variable frame rate commonly used by inexpensive cameras).

When performing smooth pursuit calibration, it is important to move the calibration stimulus slowly enough to maintain smooth visual pursuit. Targets that move too quickly will induce saccades, making it difficult to align eye and scene image streams. The limit of smooth pursuit eye movements is 30 degrees/sec, though in practice our smooth pursuit movements are closer to ~5 degrees/sec [Stampe 1993].

6.1 Discussion and Conclusion

We have designed a portable eye-tracking device that is unique in its combined affordability, unencumbered usability in real-world settings, and high tracking accuracy. Our low-cost, untethered, devices have made it possible to collect eye tracking data in real-world settings, with high accuracy. Despite short battery life and lack of active video monitoring, we have successfully used the device to track gaze patterns of young adults with ASD. The high cost of commercial devices, and tethering constraints of previously described custom-built devices would have otherwise restricted research applications of this technology.

We have prioritized affordability and usability in our design. We selected low-cost components and created an easy-to-assemble design, to make our device accessible to a broad user base. Our technical innovations have focused on compensating for the limitations introduced by low-cost, untethered cameras, especially in our development of semi-automatic POR estimation processing software pipeline. We have achieved a self-contained, untethered device capable of reliable recording of video and audio data; temporal alignment of video and audio streams; scene camera capture comparable to the wearer's field of vision; semi-automatic algorithmic point-of-regard estimation; and stable positioning of the cameras with respect to the device and to participants' faces. Most importantly, our untethered device is comprised of off-the-shelf components totaling $31[1] – $153, and can be easily assembled using a hobbyist's tools and skills.

Our design makes an important step toward widely accessible, real-world eye tracking, which we expect, in turn, to provide large quantities of ecologically valid data to scientific inquiry. We will make our designs and software freely and publicly available.

[1]We have identified an alternative camera that would bring the device's total cost down to $31. This alternative camera appears to be functionally identical to the one we used, but we have not tested it.

References

BABCOCK, J.S. AND PELZ, J.B. 2004. Building a lightweight eyetracking headgear. *Proceedings of the 2004 symposium on Eye tracking research & applications*, 109–114.

FRANCHAK, J.M., KRETCH, K.S., SOSKA, K.C., BABCOCK, J.S., AND ADOLPH, K.E. 2010. Head-mounted eye-tracking of infants' natural interactions: a new method. *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, 21–27.

HAYHOE, M. AND BALLARD, D. 2005. Eye movements in natural behavior. *Trends in Cognitive Sciences 9*, 4, 188–194.

JÄRVENPÄÄ, T. AND ÄYRÄS, P. 2010. Highly integrated near-to-eye display and gaze tracker. *SPIE Photonics Europe, 2010*, 77230Y–77230Y–10.

LI, D., WINFIELD, D., AND PARKHURST, D.J. 2005. Starburst: A hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches. *Computer Vision and Pattern Recognition (CVPR) Workshops, 2005*, 79–79.

LOWE, D. Electronics Components: 555 Timer Chip in Monostable (One-Shot) Mode - For Dummies. http://www.dummies.com/how-to/content/electronics-components-555-timer-chip-in-monostabl.html.

LUKANDER, K., JAGADEESAN, S., CHI, H., AND MÜLLER, K. 2013. OMG!: a new robust, wearable and affordable open source mobile gaze tracker. *Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services*, 408–411.

NORIS, B., KELLER, J.-B., AND BILLARD, A. 2011. A wearable gaze tracking system for children in unconstrained environments. *Computer Vision and Image Understanding 115*, 4, 476–486.

RANTANEN, V., VANHALA, T., TUISKU, O., ET AL. 2011. A Wearable, Wireless Gaze Tracker with Integrated Selection Command Source for Human-Computer Interaction. *IEEE Transactions on Information Technology in Biomedicine 15*, 5, 795–801.

STAMPE, D.M. 1993. Heuristic filtering and reliable calibration methods for video-based pupil-tracking systems. *Behavior Research Methods, Instruments, & Computers 25*, 2, 137–142.

7 Acknowledgments

This study was supported by NIMH grant R03 MH092618-01A1, CTSA Grant Number UL1 RR024139, Expedition in Computing (award #1139078), and by the Associates of the Child Study Center.