Integrating High Fidelity Eye, Head and World Tracking in a Wearable Device

Vasha DuTell
vasha@berkeley.edu
UC Berkeley School of Optometry and Vision Science Program and Redwood Center for Theoretical Neuroscience
Berkeley, California, USA

Agostino Gibaldi
agostino.gibaldi@berkeley.edu
UC Berkeley School of Optometry and Vision Science Program
Berkeley, California, USA

Giulia Focarelli
4068495@studenti.unige.it
University of Genoa DIBRIS
Genoa, Italy

Bruno Olshausen
baolshausen@berkeley.edu
UC Berkeley School of Optometry and Vision Science Program and Redwood Center for Theoretical Neuroscience
Berkeley, California, USA

Martin S. Banks
martybanks@berkeley.edu
UC Berkeley School of Optometry and Vision Science Program
Berkeley, California, USA

ABSTRACT

A challenge in mobile eye tracking is balancing the quality of data collected with the ability for a subject to move freely and naturally through their environment. This challenge is exacerbated when an experiment necessitates multiple data streams recorded simultaneously and in high fidelity. Given these constraints, previous devices have had limited spatial and temporal resolution, as well as compression artifacts. To address this, we have designed a wearable device capable of recording a subject's body, head, and eye positions, simultaneously with RGB and depth data from the subject's visual environment, measured in high spatial and temporal resolution. The sensors include a binocular eye tracker, an RGB-D scene camera, a high-frame-rate scene camera, and two visual odometry sensors, which we synchronize and record from, with a total incoming data rate of over 700 MB/s. All sensors are operated by a mini-PC optimized for fast data collection, and powered by a small battery pack. The headset weighs only 1.4 kg, the remainder just 3.9kg, and can be comfortably worn by the subject in a small backpack, allowing full mobility.

CCS CONCEPTS

• **Hardware → Sensor applications and deployments**; • **Human-centered computing → Ubiquitous and mobile computing systems and tools**; • **Applied computing → Computational biology**; **Engineering**;

KEYWORDS

mobile eye tracking, wearable eye-tracker, data collection hardware, natural scenes, human-computer interaction, open-source

ETRA '21 Adjunct, May 25–27, 2021, Virtual Event, Germany

ACM ISBN 978-1-4503-8357-8/21/05.
https://doi.org/10.1145/3450341.3458488

ACM Reference Format:
Vasha DuTell, Agostino Gibaldi, Giulia Focarelli, Bruno Olshausen, and Martin S. Banks. 2021. Integrating High Fidelity Eye, Head and World Tracking in a Wearable Device. In *2021 Symposium on Eye Tracking Research and Applications (ETRA '21 Adjunct), May 25–27, 2021, Virtual Event, Germany.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3450341.3458488

1 INTRODUCTION

The visual system evolved and developed in the natural environment, so obtaining a full understanding of its function requires studying how vision is engaged in everyday tasks in that environment. Thus, there is a great need to expand vision science beyond the controlled laboratory setting and into the natural world. Data collected in such natural conditions provides crucial information about mechanisms underlying stereopsis [Gibaldi and Banks 2021], eye movements [Gibaldi and Banks 2019], and their coordination with head movements [Hausamann et al. 2020; Kothari et al. 2020], eye optics [Gibaldi et al. 2021], and other motor behaviors [Bonnen et al. 2019; Matthis et al. 2018]. Serious technical challenges accompany expansion into the natural environment; one must record a subject's responses with high fidelity while also documenting the visible stimuli in the environment, all while allowing the subject to move and interact in a natural manner.

In previous work documenting visual properties of the natural environment such as depth statistics [Gibaldi and Banks 2019] and the power spectrum [DuTell et al. 2020], it has been critical to obtain high-resolution data free from compression artifacts. In such work, specialized hardware devices have been used that allow one to obtain high spatial and temporal resolution data. Such hardware is not typically designed for use outside of a laboratory, let alone mobile applications. Furthermore, to create a full account of the sensory-motor relationships present, this high-fidelity scene data must be collected alongside eye tracking, depth, and motion information in a time-synchronized mobile device.

We present a solution to these issues in the form of a wearable device optimized to obtain robust, high-fidelity, multi-modal data, while remaining lightweight and portable enough to enable data collection during everyday behavior in the natural environment

Figure 1: Left to Right: The device worn by a participant, highlighting the Ximea camera (blue), the Realsense D435i (red), the Realsense T265 (green), the Pupil Labs eye tracker (purple) and the operating computer (gray). Sample frames collected from Ximea camera, Realsense RGB stream, Realsense depth stream, and from the Pupil Labs binocular eye tracking cameras.

(Fig. 1). This solution adapts consumer electronics and laboratory hardware to the needs of mobile, head-mounted tracking, and combines this with software that enables accurate, high resolution data acquisition, in a convenient interface.

2 HARDWARE

Devices and Sensors. To record information about the subject and scene, our device utilizes a set of six sensors, listed in Table 1. To capture high fidelity video signal, we used a Ximea PCIE camera collecting in 8-bit CMYK color format, with a global shutter at 200fps. To supplement the color video with corresponding depth information, our system included an Intel RealSense D435i, which records both depth and RGB video streams. This device allowed us to not only match the high fidelity world camera data to a lower-resolution depth signal, but when combined with eye tracking, allowed us to estimate the subject's fixation point in the three-dimensional scene. For eye tracking, we used the Pupil Labs binocular eye tracking system [Kassner et al. 2014]. Finally, to track the subject's head and body motion, we used two Intel RealSense T265 tracking sensors [Hausamann et al. 2020]. One was mounted on the subject's back using a strap in order to measure body position and motion. The second tracker was mounted on the head, attached rigidly to the headband along with the other devices, in order to measure head position and motion.

The total data flow produced by all of the sensors was approximately 700MB/s, so that 25 minutes of data collection filled 1 TB of disk space. Although not utilized in the described configuration, the Ximea switchbox supports up to four high speed cameras, of which at least two could be supported given the power and write-speed capabilities of our system. However, as a single Ximea camera is responsible for over 90% of the incoming data, a second camera would nearly double the data stream, greatly reducing collection time, which is limited by storage space.

Device Ergonomics. The design of the head mounted portion (Fig. 1) had two main aims: (1) to be as comfortable and lightweight as possible to the participant, and (2) to allow for an adjustable configuration, depending on the subject's head and face shape, and the task at hand. The core structure of the headband was acquired from a Binocular Indirect Ophthalmoscope and adapted to hold the sensors. Custom components were designed in SolidWorks and 3D printed in PLA, making them robust yet lightweight. The three scene cameras (Ximea, RealSense D435i, and T265), were mounted together on the same 3D printed bracket, connected to the headband via three-point 3D printed adjustable ball and socket joints, and stabilized by clamps. This method allowed for easy camera adjustment to modify the pitch of the camera ensemble depending on the task, from 0 deg for far viewing (*e.g.* walking), to 30 deg downward for near viewing (*e.g.* cooking). The Ximea camera's switchbox was strapped on the headband at the back of the head using velcro. This switchbox converted the PCIE connection from the computer to the ribbon-cable connection on the camera, and powered pair of 20mm fans, that cooled the Ximea camera.

All of the power and data cables connecting the sensors and computer were bound together into a single clean band; we looped this band behind the subject's back with excess slack in the loop. This avoided tangling, yet allowed the subject to move freely without causing tension on the headband. We positioned the body tracker on the back both to avoid occlusions in front of the subject, and we found that a backpack-style strap provided more stable positioning than a chest-mounted strap. We connected the two eye tracking cameras to the headband with spherical joints, which allowed easy and stable camera positioning. The total weight of the head mount is 1.4kg; while this was the lightest setup possible given our large number of sensors, future experiments will explore the degree to which the weight of this device on the head may impact natural motion dynamics.

Operating Computer. To collect data from all these sensors simultaneously, we designed and built a PC using consumer parts. For a small form factor we used a Mini ITX motherboard (Asus OG Strix Z390-I) with 32GB of RAM, a dual M.2 support, a PCIE port, and integrated WiFi. We chose the Intel i7-8700 processor, which has sufficient computational power, yet maximized battery life due to its limited power consumption (65W). To maintain sufficient disk write speed and avoid RAM overflow, we used two M.2 SSD (Samsung 970 EVO), capable of writing at 1.2 GB/s. We mounted a touchscreen inside the PC case, for quick viewing and basic controls of the computer while mobile. For power, we used a pair of compact batteries designed to power a health-care device (CPAP machine), capable of 12V/8Ah each. The batteries were connected in parallel and power both the computer's DC power supply and the PCIE camera's external power supply. We modified a standard mini ITX computer case with a custom 3D-printed enclosure, covering the ports at the back of the computer case, exposing only the ports

Table 1: Device sensors and settings utilized by the system. While these settings gave the best overall results for our experimental setup, resolution and frame-rate settings for the RGB-D and eye tracking cameras are easily modified in the GUI. The Ximea camera's spatial and temporal resolution are easily changed in a YAML file, and the field of view adjusted with a lens change.

Sensor Type	Resolution	Field of View	Model	Location	Data Format
Eye Tracker	192 × 192 @ 200Hz	37° × 37°	Pupil Labs	L/R Eye	MPEG-4
RGB Scene Camera	2064 × 1544 @ 200fps	61° × 46°	Ximea MX031CG SY-X2G2-FL	Head	Raw Binary
RGB-D Scene Camera	1920 × 1080 @ 30fps (color) 848 × 480 @ 30fps (depth)	64° × 41° 86° × 57°	RealSense D435i	Head	MPEG-4 NumPy/PNG
IMU 1 IMU 2	200Hz	-	RealSense T265	Head Body	PupilLabs (.pldata)

for DC power, an external monitor, and Ethernet, leaving the band of sensor cables permanently connected. For computer cooling, a single CPU heatsink/fan was sufficient.

A video overview of the device hardware is available at: https://www.youtube.com/playlist?list=PLEloutX3oXFbi2CoA3_koqFSwKpdxLliF

3 SOFTWARE

Software Structure. We wrote all the device software in Python, as plugins for Pupil Labs' Pupil Capture software [Kassner et al. 2014], allowing for control of all the combined devices within a single graphical interface. We used the RGB sensor on the Intel D435i as the world camera, and modified the Pupil Capture software to save depth information as either raw *NumPy* values or lossless PNG images, rather than the default lossy MPEG-4 encoding. Our software includes a plugin to align the Realsense depth and RGB streams online, though we found this reduces the effective framerate achievable. We also wrote a plugin to view and record from the Ximea camera, as well as load and apply camera settings from a YAML file. For the odometry sensors, we used the tracker code from [Hausamann et al. 2021], modified slightly to support recording from both tracking devices and the Intel RGB/depth device simultaneously.

During collection, we used the Pupil Labs Capture software, modified by our plugins, to observe and control the computer, switch between sensor views, run eye calibration, and to start and stop collection using a GUI interface. At times when the subject was non-mobile, we were able to most easily control the computer and observe the video stream using an external monitor and Bluetooth keyboard and mouse with the computer sitting atop the table next to the subject. During mobile tasks, we found Remote Desktop over WiFi worked well. For eye tracking, we utilized the default Pupil Capture eye camera recording software, recording greyscale video of each eye at 200Hz. In addition, we wrote a Pupil Capture plugin to visualize a 9-point marker placement within the field of view of the world camera, together with a custom 3D calibration routine that is beyond the scope of this paper.

High Speed Acquisition. The biggest design challenge was the acquisition and writing of the high-speed RGB data from the Ximea camera, particularly in accommodating the high rate of data input, 637MB/s for this sensor alone. To interface with and control the camera, configure settings, and collect data, we used Xiapi, Ximea's Python API. We utilized Python's threading and queue packages to create a data collection worker threads that continuously checked for and collected images and their associated timestamps from the camera's buffer, then placed them in FIFO queues. These queues were simultaneously checked by data saving worker threads, which wrote queued frames and timestamps to disk. We saved frames in the raw binary format from the camera (400 images per file), for offline conversion to a standard image format. This was the only method we found that was sufficient to handle the high data rates from this camera; other configurations for acquisition resulted in either frames dropped or overwritten in the camera's internal buffer due to buffer overflow, or a buildup of frames in the computer's RAM due to insufficient transfer of frames from RAM to disk. We found that a similar queuing strategy for saving depth frames further stabilized effective framerates.

Data Synchronization. Importantly, given the very fast frame rate, the computer's Unix timestamp was synced with the high speed camera's internal timestamp at both the beginning and end of collection to ensure no large temporal drifts had occurred, and to ensure successful synchronization with the other sensors. Timestamp synchronization for the other sensors utilized Unix timestamps directly. When tested, synchronization was found to match between cameras within a single 200Hz frame (+/-5ms), with typically fewer than 1 dropped frames over a 1 minute collection period.

The software plugins are available on Github at: https://github.com/vdutell/hmet_aquisition

4 CONCLUSION

We report the design considerations and build of a mobile eye, head, body, and environmental motion capture device designed to be worn by a human subject while moving freely in the natural environment. To our knowledge, this device is the first to combine this level of high-fidelity, data-intensive, and synchronized multi-sensor signal capture in a mobile eye tracking device. These specific design considerations allow for a high-quality reconstruction of both the natural visual input as seen by the human retina as a subject goes about day-to-day activities, and the subject's body, head, and eye movement during natural visual behavior. Such a device lays the foundation for an improved understanding of the visual and motor systems and their adaptations to the natural environment.

We anticipate the data collected from it will be of use to the wider Vision Science, Neuroscience, and Computer Science communities.

ACKNOWLEDGMENTS

This work was supported by the Center for Innovation in Vision and Optics, NSF Grant IIS-1718991 (BAO), and the National Defense Science and Engineering Graduate Fellowship. We thank Emily Cooper, Hany Farid, Steve Cholewiak, Teresa Cañas-Bajo, and Peter Hausamann for assistance in hardware and software design, and eye tracking.

REFERENCES

Kathryn Bonnen, Jonathan S Matthis, Agostino Gibaldi, Martin S Banks, Dennis Levi, and Mary Hayhoe. 2019. A role for stereopsis in walking over complex terrains. *Journal of Vision* 19, 10 (2019), 178b–178b.

Vasha DuTell, Agostino Gibaldi, Giulia Focarelli, Bruno Olshausen, and Marty Banks. 2020. The Spatiotemporal Power Spectrum of Natural Human Vision. *Journal of Vision* 20, 11 (2020), 1661–1661.

Agostino Gibaldi and Martin S Banks. 2019. Binocular eye movements are adapted to the natural environment. *Journal of Neuroscience* 39, 15 (2019), 2877–2888.

Agostino Gibaldi and Martin S Banks. 2021. Crossed–uncrossed projections from primate retina are adapted to disparities of natural scenes. *Proceedings of the National Academy of Sciences* 118, 7 (2021), e2015651118.

Agostino Gibaldi, Vivek Labhishetty, Larry N Thibos, and Martin S Banks. 2021. The blur horopter: Retinal conjugate surface in binocular viewing. *Journal of Vision* 21, 8 (2021), –.

Peter Hausamann, Christian Sinnott, Martin Daumer, and Paul MacNeilage. 2021. Validation of the Intel RealSense T265 for Tracking Natural Human Head Motion. (2021).

Peter Hausamann, Christian Sinnott, and Paul R MacNeilage. 2020. Positional head-eye tracking outside the lab: an open-source solution. In *ACM Symposium on Eye Tracking Research and Applications*. 1–5.

Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction. In *Adjunct Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Seattle, Washington) *(UbiComp '14 Adjunct)*. ACM, New York, NY, USA, 1151–1160. https://doi.org/10.1145/2638728.2641695

Rakshit Kothari, Zhizhuo Yang, Christopher Kanan, Reynold Bailey, Jeff B Pelz, and Gabriel J Diaz. 2020. Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities. *Scientific reports* 10, 1 (2020), 1–18.

Jonathan Samir Matthis, Jacob L Yates, and Mary M Hayhoe. 2018. Gaze and the control of foot placement when walking in natural terrain. *Current Biology* 28, 8 (2018), 1224–1233.