Positional head-eye tracking outside the lab: an open-source solution

Peter Hausamann
peter.hausamann@tum.de
Technical University of Munich
Munich, Germany

Christian Sinnott
csinnott@nevada.unr.edu
University of Nevada, Reno
Reno, Nevada

Paul R. MacNeilage
pmacneilage@unr.edu
University of Nevada, Reno
Reno, Nevada

ABSTRACT

Simultaneous head and eye tracking has traditionally been confined to a laboratory setting and real-world motion tracking limited to measuring linear acceleration and angular velocity. Recently available mobile devices such as the Pupil Core eye tracker and the Intel RealSense T265 motion tracker promise to deliver accurate measurements outside the lab. Here, researchers propose a hard- and software framework that combines both devices into a robust, usable, low-cost head and eye tracking system. The developed software is open source and the required hardware modifications can be 3D printed. The researchers demonstrate the system's ability to measure head and eye movements in two tasks: an eyes-fixed head rotation task eliciting the vestibulo-ocular reflex inside the laboratory, and a natural locomotion task where a subject walks around a building outside of the laboratory. The resultant head and eye movements are discussed, as well as future implementations of this system.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing systems and tools.**

KEYWORDS

Eye tracking, head tracking, gaze estimation, mobile, simultaneous localization and mapping, locomotion, open source

ACM Reference Format:
Peter Hausamann, Christian Sinnott, and Paul R. MacNeilage. 2020. Positional head-eye tracking outside the lab: an open-source solution. In *Symposium on Eye Tracking Research and Applications (ETRA '20 Short Papers), June 2–5, 2020, Stuttgart, Germany.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3379156.3391365

1 INTRODUCTION

Mobile observers tend to fixate environmental features that are stationary in a world-fixed reference frame. However, this behavior is usually impossible to characterize because mobile eye trackers measure eye movements in a head-fixed reference frame (eye-in-head) [Kinsman et al. 2012; Kothari et al. 2019]. Information about head movement relative to the environment (head-in-world) would allow characterizing eye movements in a world-fixed reference frame (eye-in-world), but positional head tracking has historically been difficult to achieve outside the lab. To comprehensively characterize natural interaction with the visual environment (eye-in-world), a convenient mobile solution is needed that allows for simultaneous positional tracking of head-in-world and eye-in-head movements outside the lab.

Measurement of head movement outside the laboratory setting has traditionally relied on microelectromechanical system (MEMS)-based inertial measurement units (IMUs). IMUs are advantageous due to their portability and affordability, but they do not measure 6 degree of freedom (DOF) head-in-world position. They are limited to measurement of angular velocity and sum total gravitoinertial acceleration. Angular position may be estimated by integrating angular velocity, but these estimates are subject to drift [Kinsman et al. 2012; Li et al. 2009]. Estimation of linear position is even more problematic due to the ambiguity of gravitoinertial acceleration. Distinct estimates of gravitational and inertial acceleration may be obtained, e.g. via Kalman filtering [Li and Wang 2012], but double-integration of the resulting inertial acceleration to obtain linear position yields highly unreliable estimates. Nevertheless, several recent studies have investigated the use of IMU data to compensate for head movements during eye tracking [Kothari et al. 2019; Larsson et al. 2014].

More reliable positional estimates of head movement outside the lab may be obtained by fusing IMU data with other data streams. One recent study used a motion capture suit composed of multiple IMUs to track full-body kinematics along with eye-in-head movement [Matthis et al. 2018]. However, use of multiple IMUs does not fully solve the problem of drift. Perhaps the most promising solution involves joint analysis of IMU data along with video recorded from a rigidly mounted camera, known as visual-inertial simultaneous localization and mapping (VI-SLAM). This is a computer vision-based method that relies on frame-by-frame video analysis to create a 3D representation of the environment and simultaneously localize the agent's position relative to that environment [Chen et al. 2018].

Here, we combine two devices to create a highly mobile system for simultaneous tracking of head and eye position. To track 6 DOF head position we use a ready-made VI-SLAM system, the Intel RealSense T265 tracking camera, which fuses inertial data from an IMU with two global shutter fisheye cameras to generate a 6 DOF pose estimate at 200Hz. The system is relatively compact and affordable, and the on-board VI-SLAM solution has the benefit of greatly reducing computational load. To track eye movements, we used the Pupil Core head-mounted eye tracker which is also relatively inexpensive and lightweight. In addition, the software

ETRA '20 Short Papers, June 2–5, 2020, Stuttgart, Germany

ACM ISBN 978-1-4503-7134-6/20/06…$15.00
https://doi.org/10.1145/3379156.3391365

suite supporting the Pupil Core is entirely open source facilitating integration with other devices.

In this paper, we present software that allows the Pupil Core and Intel RealSense T265 to be used together for simultaneous positional head-eye tracking outside the lab. We demonstrate the functionality of the system by presenting two example datasets. Tasks were chosen to demonstrate capture of basic physiological head-eye behaviors in both a laboratory and a real-world setting.

2 METHODS

2.1 Hardware

2.1.1 Pupil Core. The Pupil Core system is a mobile eye tracking device from Pupil Labs [Kassner et al. 2014]. Our configuration of the Pupil Core eye-tracker features three cameras. Two of these film the eyes at 200 frames per second (FPS) with a resolution of 192x192 pixels. The last camera is an outward-facing world camera and records at 30 FPS with a 1280x720 pixel resolution and a diagonal field of view (FOV) of 100°. After consulting with other researchers we modified the Pupil Core to add a nose cushion (for ergonomics) and a IR-reflective film around the user's eyes (to reduce errant IR reflections during outdoor recording [Bonnen et al. 2019]).

2.1.2 RealSense T265. Head pose was measured using the Intel RealSense T265 tracking camera. The device consists of two global shutter fisheye world cameras (173° diagonal FOV; 30 Hz frame rate; 848x800 pixel resolution), a 3 DOF accelerometer (±4g range; 62.5 Hz sampling rate), and a 3 DOF gyroscope ($\pm 2000\frac{°}{s}$ range; 200 Hz sampling rate). The T265 uses VI-SLAM to fuse data from these streams to estimate 6 DOF position and velocity of the camera relative to the environment at 200 Hz.

2.1.3 Camera mount. Our solution requires recording simultaneously from both the T265 and the native Pupil world camera, and the two must be rigidly attached to each other. To this end, we designed and 3D-printed a mounting bracket that holds the T265 and can be clipped and glued to the world camera. The CAD model for this bracket is freely available for download[1].

2.2 Software

2.2.1 Overview. Data acquisition and export was implemented as a set of Pupil Core software plugins. These plugins are written in Python and can be used by copying the plugin source file to a folder specified by the Pupil Core software. The software suite developed by Pupil Labs for use with Pupil Core consists of three programs: Pupil Capture, a data acquisition program with GUI; Pupil Player, a data visualizer and exporter; and Pupil Service, a debugging and acquisition (from terminal) program. We use Pupil Capture and Player for purposes of this report.

All data recorded by the plugins is stored in the same format as the data streams recorded by Pupil Capture (such as gaze positions) in order to to streamline the integration with the software. The plugins along with example data and analysis code are open source and freely available for download[1]. In the code, head tracking data is referred to as odometry as it includes positions as well as velocities.

[1]https://github.com/vedb/pupil-t265

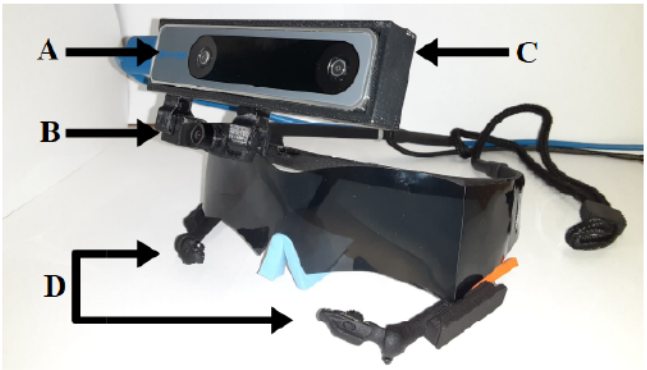

Figure 1: Pupil and RealSense combined head and eye movement tracker. *The RealSense T265 (A) is mounted rigidly above the Pupil Core world camera (B) via a custom 3D-printed bracket (C). The eye cameras (D) are also visible.*

2.2.2 Recording. Eye tracking data is recorded via Pupil Capture. In its original configuration, Pupil Capture allows the user to record from a world-facing camera and up to two eye cameras simultaneously. The user is able to modulate sampling rate, resolution, exposure time, and other image parameters before recording. The software also allows for the user to map gaze in real time (online) or later, after a recording has been made (offline).

Recording of head tracking data is achieved through a plugin for Pupil Capture. The user can start the T265 device by clicking a button, and if the device is successfully started, they will see a live readout of the current tracker sampling rate, confidence, linear and angular position, and linear and angular velocity. When a recording is initiated through Pupil Capture, head tracking data is continuously written to disk along with eye tracking data. It is also possible to show the video stream recorded by the two T265 fisheye cameras in a separate window.

2.2.3 Camera extrinsics estimation. Both the T265 and the Pupil Core tracker define reference frames with respect to their world cameras. However, for analyses directly relating head and eye movements, the exact transformation between the two devices needs to be calculated in order to transform head and eye tracking data into a common reference frame. For this purpose, a second plugin was developed that estimates the position and orientation of the RealSense T265 with respect to the Pupil world camera. This is achieved by capturing a calibration pattern presented on the screen by all cameras from multiple angles and performing a stereo camera calibration for pairs of cameras using the `stereoCalibrate` function from the OpenCV computer vision library [Bradski 2000] (version 3.2, included in the Pupil software suite). The plugin also takes into account the different coordinate system of the T265 pose (x-axis leftwards, y-axis upwards, z-axis backwards) compared to the standard camera coordinates used by Pupil Core (x-axis leftwards, y-axis downwards, z-axis forwards).

As with the recording plugin, the user can inspect the fisheye video stream from the T265 which can be useful to verify that the calibration target is captured by all three cameras. The target, consisting of an asymmetric grid with 44 circle markers, can be

shown in a window or in fullscreen mode. Calibration images (10 in total) can be acquired with a key or button press in the world camera window. The plugin will also draw the extent of the previously recorded targets over the current world camera image so that the user can make sure to cover the entire FOV of the camera.

2.2.4 Data export. Export is implemented as a plugin for Pupil Player. When the plugin is activated, head tracking data and camera extrinsics are saved as `.csv` files each time an export is triggered through Pupil Player. The export format is consistent with other exports that can be made with Pupil Player such as gaze data and includes timestamps and the indices of corresponding world camera frames.

2.3 Data acquisition

To illustrate the solution, we recorded data from two subjects and analyzed this data to demonstrate the workflow. Each subject put on the modified Pupil Core tracker which was attached to a laptop carried in a backpack. A 9-point calibration routine was performed in which the experimenter presented a specialized marker at nine locations configured in a 3x3 grid across the FOV of the world camera. The subject fixated the center of the marker for several seconds at each location, viewing the calibration target grid from a distance of about 1.2 m. This array corresponds to a 72° by 42° displacement of visual angle. Data from this calibration procedure was later used to perform post-hoc gaze mapping in the Pupil Player software.

After the calibration, the first subject was presented with the same kind of marker at the center of the visual field and was instructed to fixate on the marker while nodding and shaking their head several times to elicit a rotational vestibulo-ocular reflex (rVOR) response. Then, while still fixating on the target, they performed several vertical full-body squatting movements to elicit a translational VOR (tVOR) response. This was done at two viewing distances of approximately 1.2 and 2.1 m. The second subject performed the same calibration routine and then walked around the outside of a building for roughly five minutes before returning to the lab.

2.4 Analysis

2.4.1 Reference frame transformations. The T265 measures position and velocity of the device with respect to a world-fixed reference frame whose origin is at the point where the device was started. The Pupil Core measures the 3D position of the gaze point with respect to a reference frame fixed to the world camera. The transformation between the T265 and Pupil Core coordinate systems is computed with the calibration procedure outlined in section 2.2.3. The orientation of the eye with respect to the Pupil Core world camera frame can be estimated by computing the shortest arc rotation between a unit vector perpendicular to the image plane and the gaze point.

Finally, we define a coordinate system for analysis where the x-axis points forward, the y-axis to the left and the z-axis upwards. All measurements in our analysis are transformed from their original reference frame to this coordinate system. While this coordinate system is easier to interpret from an anatomical perspective, it is still relative to the Pupil Core world camera rather than an anatomical reference such as Reid's plane (see discussion). Directions and orientations are also expressed in a spherical coordinate system where the azimuth angle is measured from the positive x-axis.

2.4.2 Data pre-processing. The 3D gaze position was transformed to the spherical coordinate system defined in the previous section and smoothed with a boxcar filter with a window length of 500 ms. Afterwards, gaze velocity was computed by differentiating the gaze position with respect to time. Pupil Core provides a confidence score of the gaze estimate between 0 and 1. All gaze data with a confidence below 0.8 was excluded from the analysis. Similarly, the T265 reports a tracking confidence from 0 to 3, with a score of 3 corresponding to "high" tracking confidence. All tracking data below the highest score was also excluded. Additionally, gaze data whose velocity had a magnitude above $1000\frac{^\circ}{s}$ was excluded from analysis [Holmqvist and Andersson 2017].

2.4.3 Vestibulo-ocular reflex. The angular and linear movements of the subject while fixating the calibration target were used to demonstrate the reflexive eye movements induced by the rVOR and tVOR. Both rVOR and tVOR are combined head and eye movements performed constantly in normally functioning vertebrates for image stabilization [Einhäuser et al. 2007], and represent an ambiguous case for gaze classifiers when head movement data is unavailable. For the rVOR, we compared eye yaw and pitch velocity with negative head yaw and pitch velocity, respectively. For the tVOR, we compared eye pitch velocity with the angular velocity of the head relative to the marker with $\omega = \arctan(v_z/d)$ where v_z is the vertical head velocity and d is the distance of the marker from the head. VOR velocity gain was calculated by computing the median ratio of eye velocity to head velocity along the corresponding axis for each segment.

2.4.4 Eye-in-world velocity. Motion at the retina is driven predominantly by how the eye is moving relative to the world-fixed visual environment. Therefore, characterization of eye-in-world velocity can provide insight into typical visual motion stimuli during natural behavior [MacNeilage et al. 2019; Matthis et al. 2019]. As described above, reconstruction of eye-in-world position and velocity is only possible by combining measures of head-in-world and eye-in-head movement.

To illustrate one application of positional head-eye tracking, we analyze the direction of linear velocity (heading) in eye coordinates (eye-in-world) during outdoor walking. Linear head velocity measured in the world frame by the T265 was transformed into spherical head coordinates and then further transformed into spherical eye coordinates using eye-in-head position (see section2.4.1). We plot 2D histograms of heading as well as the eye-in-head position between ±45° azimuth and elevation.

3 RESULTS

3.1 Vestibulo-ocular reflex

During the rVOR task, eye velocity and negative head velocity were closely aligned and exhibited a gain of 1.01 and 0.83 for yaw and pitch, respectively (figures 2A and B). Similar alignment was observed for eye velocity and angular velocity of the head relative

to the fixation target during the tVOR task, although eye movement seemed to over-compensate for head movement (VOR velocity gains 1.32 and 1.52 for near and far fixation, respectively, figures 2C and D). This can be explained by the fact that the subject simultaneously performed small involuntary head pitch rotations that also need to be compensated.

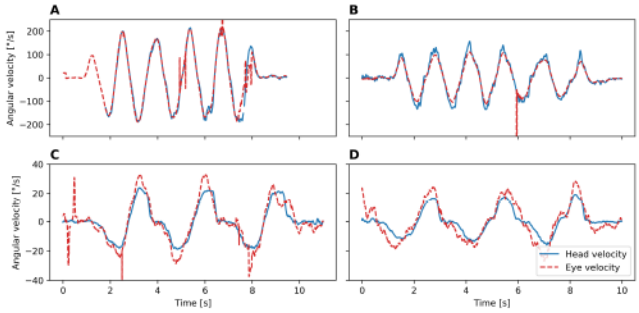

Figure 2: Comparison of eye and head movements during fixation task. ***Eye yaw velocity and negative head yaw velocity (A). Eye pitch velocity and negative head pitch velocity (B). Eye pitch velocity and angular velocity of the head relative to the fixation target (1.219 m distance, C). Eye pitch velocity and angular velocity of the head relative to the fixation target (2.143 m distance, D).***

3.2 Eye-in-world velocity

Direction of linear head-in-world velocity during walking for this subject and environment showed an elongated distribution along the elevation axis (figure 3A). The distribution for eye-in-head position was more circular (figure 3B). The offsets of these distributions relative to the origin are likely observed because the respective coordinate systems are not aligned with an anatomical reference frame such as Reid's plane. In contrast, eye-in-world velocity exhibited a Gaussian-like distribution, centered in both azimuth and elevation direction (figure 3C) reflecting a tendency to look in the direction of linear motion. This result shows that although head velocity and gaze position are not aligned with an anatomical reference, the resulting eye-in-world velocity is aligned with an anatomical eye-fixed reference frame.

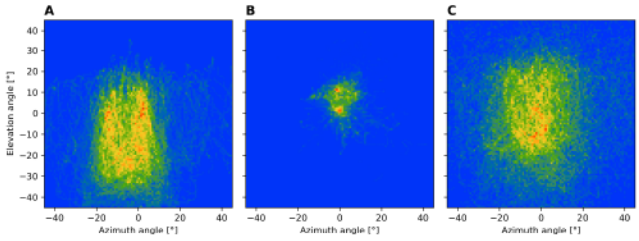

Figure 3: Distributions of head/eye velocities and gaze direction during walking task. ***Direction of linear head-in-world velocity (A). Eye-in-head gaze direction (B). Direction of linear eye-in-world velocity (C).***

4 DISCUSSION

We demonstrate a mobile head- and eye-tracking platform capable of robust positional measurement and present analysis of two paradigms to validate the system, a fixation and a walking task. Simultaneous head and eye measurements during the fixation task revealed the expected compensatory eye movements (figure 2). Joint analysis of head and eye measurements during walking yielded a novel characterization of linear eye-in-world velocity that is a strong determinant of retinal image motion (figure 3).

While our head-eye tracker is relatively light on the user's head, it still requires use of a dedicated computer. We used a high-performance laptop carried in a backpack for data collection that was relatively heavy. Additionally, online data monitoring was impossible with this hardware configuration. This configuration is similar to another off-the-shelf product, Magic Leap. Our solution has the advantage of higher gaze sampling rates and open access to software and hardware collecting these data.

Additionally, weight of future hardware configurations is likely to decrease as personal computing improves. However, online data monitoring must be implemented by modifying how the laptop is carried to make the screen visible. One promising solution uses a lightweight hydration backpack with a window cutout to display the screen of the tablet being used for data collection [Santini et al. 2018].

Another concern is minimizing headset slippage. We attempted to control slippage using a sport strap fixed to the end-pieces of the Pupil Core eye tracker glasses. With this solution, headset slippage did not prove to be an issue in the collection of this pilot data, but collection over longer periods must be tested.

To address slippage as well as conversion of measurements into an anatomically defined reference frame such as Reid's plane, it would be useful to measure the 3D position and orientation of the device relative to the head. This would be possible, for example, by scanning the participant's head with headset just before data collection. One possible method, originally developed to create custom head stabilization devices [Gao et al. 2017] employs a stereo camera attached to a cell phone. To quantify the amount of headset slippage experienced by any single participant one could compare scans acquired before and after a recording session.

Using VI-SLAM to measure head position is promising, but VI-SLAM-based solutions can deteriorate when visual features of the environment move, for example in a snowstorm, when walking in a crowd or when surrounded by foliage that is blowing in the wind. This is because VI-SLAM works under the assumption that environmental features are world-fixed. Roboticists have tried solving this problem by orienting cameras towards the ceiling in indoor environments [Jeong and Lee 2005], but this is not feasible with our system outdoors.

The Intel RealSense T265 was developed primarily to facilitate navigation in autonomous robots such as drones or self-driving vehicles [Chen et al. 2018]. Future human-centric development of a VI-SLAM pipeline leveraging regularities of human movement should lead to better estimation of head position. Higher quality head movement data can then be used to improve gaze classification performance [Kinsman et al. 2012; Kothari et al. 2019].

Several recent studies have characterized natural head and eye movements during everyday activities in natural environments [Bonnen et al. 2019; MacNeilage et al. 2019; Matthis et al. 2019, 2018]. Novel, versatile, and accessible methods, like the ones presented here, will help advance these research endeavors by enabling longer recording sessions with data collection across a wider range of activities, environments, and subjects. Ultimately, this knowledge will lead to better understanding of human visual exploration and head-eye coordination, and this knowledge can be applied in the development of interactive technologies that rely on tracking and prediction of head and eye movements, such as augmented and virtual reality.

ACKNOWLEDGMENTS

This research was supported by NSF under grant number OIA-1920896 and NIH under grant number P20 GM103650.

The authors would like to thank Kathryn Bonnen, Jonathan S. Matthis, and Mary Hayhoe, whom we consulted with regarding best practices for outdoor eye-tracking and data collection.

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). Issue 178b.

G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).

Chang Chen, Hua Zhu, Menggang Li, and Shaoze You. 2018. A review of visual-inertial simultaneous localization and mapping from filtering-based and optimization-based perspectives. *Robotics* 7 (7 2018), 1–20. Issue 45.

Wolfgang Einhäuser, Frank Schumann, Stanislavs Bardins, Klaus Bartl, Guido Böning, Erich Schneider, and Peter König. 2007. Human eye-head co-ordination in natural exploration. *Network: Computation in Neural Systems* 18 (9 2007), 267–297. Issue 3.

James Gao, Alex Huth, and Young Park. 2017. *Case Forge.* Retrieved February 3, 2020 from https://caseforge.co/

Kenneth Holmqvist and Richard Andersson. 2017. *Eye-tracking: A comprehensive guide to methods, paradigms and measures.*

WooYeon Jeong and Kyoung Mu Lee. 2005. CV-SLAM: A new ceiling vision-based SLAM technique. *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems* (2005), 1–6.

Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication* (9 2014), 1151–1160.

Thomas Kinsman, Karen Evans, Glenn Sweeny, Tommy Keane, and Jeff Pelz. 2012. Ego-motion compensation improves fixation detection in wearable eye tracking. *Proceedings of the Symposium for Eye Tracking Research and Applications* (3 2012), 221–224.

Rakshit Kothari, Zhizhuo Yanh, Christopher Kanan, Reynold Bailey, Jeff Pelz, and Gabriel Diaz. 2019. Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities. *arXiv* (5 2019).

Linnéa Larsson, Marcus Nyström, Andrea Schwaller, Martin Stridh, and Kenneth Holmqvist. 2014. Compensation of head movements in mobile eye-tracking data using an inertial measurement unit. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication* (9 2014), 1161–1167.

Feng Li, Jeff B. Pelz, and Scott J. Daly. 2009. Measuring hand, head and vehicle motions in commuting environments. *Proceedings of SPIE* 7240 (2009), 7240–1–7240–11.

Wei Li and Jinling Wang. 2012. Effective adaptive Kalman filter for MEMS-IMU/Magnetometers integrated attitude and heading reference systems. *The Journal of Navigation* 66 (2012), 99–113. Issue 1.

Paul R. MacNeilage, Luan Nguyen, and Christian Sinnott. 2019. Characterization of natural head and eye movements driving retinal flow. *Journal of Vision* 19, 10 (2019). Issue 147d.

Jonathan S. Matthis, Karl S. Muller, and Mary M. Hayhoe. 2019. Retinal optic flow and the control of locomotion. *Journal of Vision* 19, 10 (2019). Issue 179.

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

Thiago Santini, Hanna Brinkmann, Luise Reitstätter, Helmut Leder, Raphael Rosenberg, Wolfgang Rosenstiel, and Enkelejda Kasneci. 2018. The art of pervasive eye tracking: Unconstrained eye tracking in the Austrian gallery Belvedere. *PETMEI '18: 7th Workshop on Pervasive Eye Tracking and Mobile Eye-Based Interaction* (2018), 1–8.