Labelled pupils in the wild: A dataset for studying pupil detection in unconstrained environments

Marc Tonsen Xucong Zhang Yusuke Sugano Andreas Bulling
Perceptual User Interfaces Group
Max Planck Institute for Informatics, Saarbrücken, Germany
{tonsen,xczhang,sugano,bulling}@mpi-inf.mpg.de

Abstract

We present labelled pupils in the wild (LPW), a novel dataset of 66 high-quality, high-speed eye region videos for the development and evaluation of pupil detection algorithms. The videos in our dataset were recorded from 22 participants in everyday locations at about 95 FPS using a state-of-the-art dark-pupil head-mounted eye tracker. They cover people of different ethnicities and a diverse set of everyday indoor and outdoor illumination environments, as well as natural gaze direction distributions. The dataset also includes participants wearing glasses, contact lenses, and make-up. We benchmark five state-of-the-art pupil detection algorithms on our dataset with respect to robustness and accuracy. We further study the influence of image resolution and vision aids as well as recording location (indoor, outdoor) on pupil detection performance. Our evaluations provide valuable insights into the general pupil detection problem and allow us to identify key challenges for robust pupil detection on head-mounted eye trackers.

CR Categories: H.5.2 [Information Interfaces and Presentation]: User Interfaces—Evaluation/methodology

Keywords: Pupil detection; Head-mounted eye tracking; High-speed; High-quality

1 Introduction

Pupil detection is a core component of shape-based gaze estimation systems and is therefore well established as a research topic in eye tracking [Hansen and Ji 2010]. Robust and accurate pupil detection is challenging, particularly in eye images recorded using head-mounted eye trackers. These trackers are used in mobile everyday settings and eye images are therefore subject to significant influences from changes in ambient light, corneal reflections, pupil occlusions, and shadows (see Figure 1). Accurate pupil positions are particularly important for recent methods that directly map the detected 2D pupil positions to 3D gaze estimates (see [Mansouryar et al. 2016] for an example). Current benchmark datasets have two main limitations that impede further advances in computational methods for pupil detection on head-mounted eye trackers.

First, most existing datasets were recorded using remote cameras and consist of only monocular RGB images (see [Jesorsky et al. 2001] for an example). Images recorded under these conditions are significantly different from the close-up infrared eye region images recorded on head-mounted eye trackers. Second, the few datasets for head-mounted pupil detection that are publicly available are either limited in size, were recorded in controlled laboratory settings and therefore do not cover realistic day-to-day usage scenarios – which, for example, also include transitions of users between indoor and outdoor environments – or contain only low-quality eye images (see Table 1 for a comparison). The dataset presented in [Świrski et al. 2012] includes 600 high-quality close-up eye images and manual ground truth annotations of the pupil centre. While this dataset is a good starting point to evaluate pupil detection algorithms, it is limited in that it only contains eye images of two participants and was collected in the laboratory under controlled lighting conditions. A more recent dataset was introduced in [W. Fuhl 2015]. This dataset is significantly larger and images were recorded with a head-mounted eye tracker in uncontrolled environments, namely while driving and shopping, but not in fully outdoor environments.

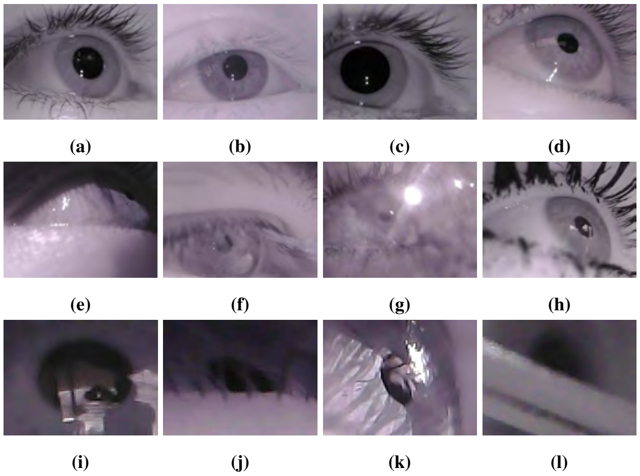


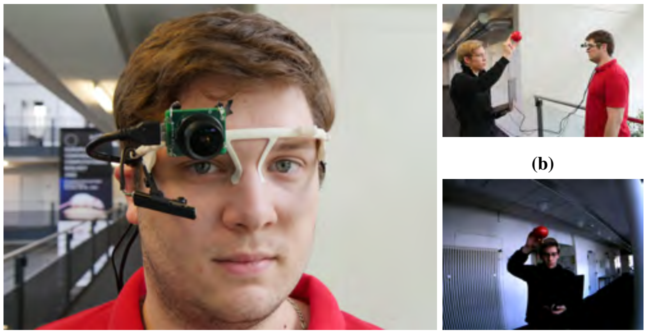
Figure 1: *Diverse set of images from our dataset. First row: (a) (b) (c) and (d) show different eye appearances. Second row, most difficult cases: (e) strong shade, (f) eyelid occlusion, (g) reflection on glasses, (h) heavy makeup. Third row, challenging images cropped around the pupil: (i) reflection on the pupil, (j) self-occluded, (k) strong sunlight and shade, (l) occlusion by glasses.*

In this paper we present *labelled pupils in the wild* (LPW), a novel pupil detection dataset that aims to address these limitations. More specifically, we present a dataset of 66 high-quality eye region videos that were recorded from 22 participants using a state-of-the-art dark-pupil head-mounted eye tracker. Each video in the dataset consists of about 2,000 frames with a resolution of 640x480 pixels and was recorded at about 95 FPS, resulting in a total of 130,856 video frames. The dataset is an order of magnitude larger than existing ones and covers a wide variety of realistic indoor and outdoor illumination conditions, including participants wearing glasses and

ETRA '16,, March 14 - 17, 2016, Charlston, SC, USA
ISBN: 978-1-4503-4125-7/16/03. . .$15.00
DOI: http://dx.doi.org/10.1145/2857491.2857520

	participants	sessions	images	camera angles	lighting conditions	ethnicities	resolution	FPS
[Świrski et al. 2012]	2	4	600	4	1	n.a.	640x480	static images
[W. Fuhl 2015]	17	17	38,401	mostly frontal	≤ 17	n.a.	384x288	25
Ours	22	66	130,856	continuous	continuous	5	640x480	95

Table 1: *Comparison of current publicly available datasets for pupil detection on head-mounted eye trackers.*



(a) (b) (c)

Figure 2: *Data collection setup. (a) The high frame rate eye and scene cameras. (b) Participants move their eyes by looking at the red ball. (c) The image captured by the scene camera.*

eye make-up, and covering different ethnicities with variable skin tones, eye colors, and face shapes. All videos were manually ground-truth annotated with accurate pupil ellipse and centre positions. We further evaluate several state-of-the-art pupil detection algorithms on this challenging new dataset. Our evaluations provide valuable insights into the pupil detection problem setting and allow us to identify key challenges for pupil detection on head-mounted eye trackers. The full dataset and ground truth annotations are publicly available at http://mpii.de/LPW.

2 Labelled pupils in the wild (LPW) dataset

We designed a data collection procedure to 1) have a large variability in appearance of participants, such as gender, ethnicity and use of vision aids and 2) record participants under different conditions, such as lighting or eye camera position. Therefore we took each participant to a new set of locations and recorded their eye movements while fixating a gaze target. Outdoor videos typically have bright(er) natural lighting, while most indoor videos include both natural light from windows and artificial illumination.

Participants and apparatus

We recruited 22 participants (9 female) through university mailing lists and personal communication. Details about our participants can be found in Table 2. The eye tracker used for the recording was a high-speed Pupil Pro head-mounted eye tracker that records eye videos at 120 Hz [Kassner et al. 2014]. We replaced the original scene camera with a PointGrey Chameleon3 USB3.0 camera recording at up to 149 Hz. The hardware setup is shown in Figures 2a and 2b. It allowed us to record all videos at 95 FPS, which is a speed at which even fast eye movements continue through several frames. The video encoding used was Motion JPEG.

Procedure

As shown in Figure 2b, the participants were instructed to look at a moving red ball as a fixation target during the data collection, which is shown in Figure 2c with an image captured by the scene camera. In order to cover as many different conditions as possible, we randomly picked recording locations in and around several buildings. Each location was not chosen more than once during the entire recording of all participants. 34.3% of the recordings were done outdoors, in 84.7% natural light was present and in 33.6% artificial light was present. Besides locations, we have also tweaked the angle of the eye cameras such that the dataset contains a wide range of camera angles from frontal views to highly off-axis angles. This is done by either asking the participant to take the tracker off and put it back on, or manually moving the camera. With each of the 22 participants we recorded three videos around 20 seconds in length, yielding 130,856 images overall. Participants could keep their glasses and contact lenses on during the recording.

Ground truth annotation

We used different methods for annotation. In many easy cases such as some indoor recordings, the pupil area has a clear boundary and no strong reflections inside. We annotated these frames by manually selecting 1 or 2 points inside the pupil area, using them as seed points to find the largest connected area with similar intensity values. The pupil centre is defined as the centroid of this area. Some recordings have a clear scene video but strong reflections/noise in the eye video, such as outdoor recordings under strong sunlight. In those cases, we tracked the fixation target (red ball) in the scene videos and manually annotated part of the eye pupil positions in the eye videos. From this calibration data we computed a mapping function from target positions to pupil positions. In addition, the annotators examined each video again to verify the annotation results and to correct mistakes by manually fitting an ellipse to the pupil with 5 points to select the ellipse centre as pupil centre.

3 Results

To evaluate the difficulty and challenges contained in our dataset, we have analysed the performance of five state-of-the art pupil detection algorithms. *Pupil Labs* [Kassner et al. 2014] is the algorithm used in the Pupil Pro eye tracker. *Swirski* [Świrski et al. 2012] and *ExCuSe* [W. Fuhl 2015] are taken as examples of state-of-the-art algorithms. *Isophote* [Valenti and Gevers 2012] and *Gradient* [Timm and Barth 2011] are two simple algorithms designed for the iris shape fitting task on low-resolution remote eye images. In the following sections we examine several performance values and highlight key challenges in our dataset. A set of sample images for the most difficult cases can also be found in Figure 1 e) to h). We ran the evaluations on a Linux system desktop with an Intel E5800 CPU 3.16GHz processor and 8GB memory. The average processing speed of each algorithm was: *Isophote* 225.59 fps, *Pupil Labs* 45.09 fps, *Gradient* 43.52 fps, *Swirski* 5.44 fps, *ExCuSe* 1.90 fps.

Accuracy and robustness

Figure 3 shows the cumulative error distribution of all algorithms on the entire dataset. One can see that *Pupil Labs*, *Swirski* and *ExCuSe* all return very good results in roughly 30% of all cases with less than 5px error; however, their performance falls off quickly. It is worth mentioning that ExCuSe falls off last. The *Gradient*

	P01 (m)		**P02 (m)**		**P03 (f)**		**P04 (m)**		**P05 (f)**		**P06 (n)**		**P07 (m)**		**P08 (m)**		**P09 (m)**		**P10 (f)**		**P11 (m)**	
Nationality	Iranian		German		Iranian		Indian		German		Indian		Indian		Pakistani		German		Indian		Pakistani	
Glasses	No		No		Yes		No		Yes		No		No		No		No		No		No	
Video variability	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out
	2	1	2	1	2	1	2	1	1	2	2	1	1	2	2	1	2	1	2	1	2	1
	Nat	Art	Nat	Art	Nat	Art	Nat	Art	Nat	Art	Nat	Art	Nat	Art	Nat	Art	Nat	Art	Nat	Art	Nat	Art
	3	1	3	1	2	2	2	1	3	1	2	1	3	1	3	1	3	1	2	1	1	2

	P12 (m)		**P13 (m)**		**P14 (f)**		**P15 (f)**		**P16 (m)**		**P17 (m)**		**P18 (m)**		**P19 (f)**		**P20 (f)**		**P21 (f)**		**P22 (f)**	
Nationality	Egyptian		Indian		Indian		German		German		Indian		Indian		Indian		Indian		Indian		German	
Glasses	No		No		No		No		Contacts		No		No		No		No		No		No	
Video variability	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out
	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	3	0
	Nat	Art	Nat	Art	Nat	Art	Nat	Art	Nat	Art	Nat	Art	Nat	Art	Nat	Art	Nat	Art	Nat	Art	Nat	Art
	3	1	3	1	2	1	3	1	3	1	3	1	2	1	3	1	2	1	2	2	2	1

Table 2: *Characteristics of the LPW dataset. The gender of participants has been indicated as female (f) and male (m). The variability of the environment is represented as indoor (In) and outdoor (Out), with natural (Nat) and artificial (Art) light. With each participant we recorded three videos. Note that a single video can contain both natural and artificial light. There is no outdoor video for P22 because it was raining on the recording day. P22 also wore heavy eye make-up.*

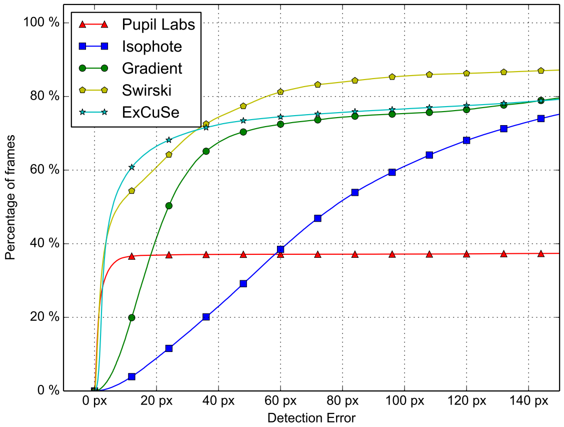


Figure 3: *Cumulative error distribution of each algorithm on the LPW dataset. The y-axis shows the percentage of detections with a pixel error smaller than the corresponding x-value.*

detector follows a similar curve but shifted to the right, indicating a higher error on average. The *Isophote* detector's curve rises the least steeply indicating the highest error on average. *Pupil Labs* stands out by cutting off very early. While giving fairly accurate results in almost 40% of all cases, it completely fails in the other 60%. *ExCuSe*, *Swirski* and the *Gradient* detector return reasonable results with an error of roughly 40px in about 70% of all cases, indicating a higher robustness in comparison to *Pupil Labs*.

Overall, so far there is no satisfying performance on the dataset for gaze estimation. This indicates the difficulty of our dataset, i.e., pupil detection in the wild is still challenging for current methods. According to our observations, the hardest samples are mainly cases of strong shadows, eyelid occlusions, reflections from glasses and heavy make-up (see also Figure 1 (e), (f), (g) and (h)).

Indoor vs outdoor

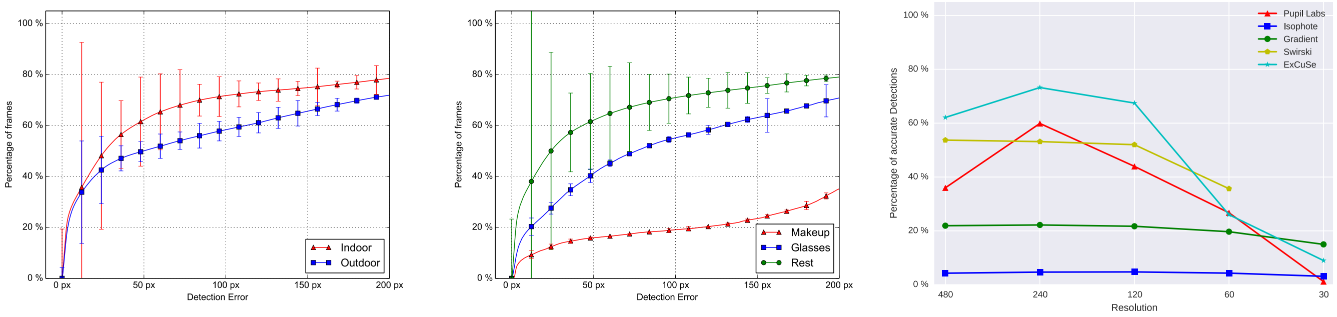
Outdoor images are especially challenging for pupil detection algorithms, since the infrared portion of strong sunlight can create reflections and shadows on the pupil and iris (see also Figure 1 (e), (i) and (k)). Light falling directly into the camera lense can create additional reflections. Figure 4a shows the cumulative error distribution for the mean error of all algorithms for indoor and outdoor scenes. While for indoor scenes roughly 60% of all detections had an error of 50px or lower, for outdoor scenes it is only about 50%.

Glasses and makeup

For users with impaired vision, the possibility to wear glasses along with the eye tracker is very important. However, glasses can cause intense reflections and the pupil will often be partially occluded (see also Figure 1 (g) and (l)). The performance of the examined algorithms is significantly worse for participants wearing glasses compared to those without glasses (see the Figure 4b). According to our evaluation, makeup also greatly disturbs the performance of the examined algorithms, which is also visible in Figure 4b. Although the number of participants wearing glasses or makeup is too small in our dataset to make any statistically significant statements, this behaviour is to be expected, since all algorithms either look for large black blobs or strong edges, which both could also be created by makeup or glasses.

Resolution

The examined algorithms have been designed for different systems working with different image resolutions. Namely, the *Isophote* and *Gradient* detectors have been designed to work on low-resolution images, while the others are generally meant for higher resolutions. In Figure 4c, we show the performance of each algorithm for different resolutions. The error is normalised by image width, and the percentage of detections with an error lower then 0.02 is shown. Parameters depending on the image size have been modified accordingly for all algorithms. The results for 30p of *Swirski* are missing because we couldn't get it to work on that resolution. It is important



(a) **(b)** **(c)**

Figure 4: *Performance as it relates to different factors. Cumulative mean error distribution for indoor and outdoor videos of the 5 algorithms (a). The x-axis describes the detection error in pixels, while the y-axis describes the percentage of detections that had an error equal to or lower than the corresponding x-value. A similar cumulated error distribution for the data that include glasses, makeup or neither (b). Performance of each algorithm for images scaled to different resolutions (c). The x-axis states the height in pixels of the resolution used (ratio of 4:3 is fixed). The y-axis describes the percentage of detections with normalised error smaller than 0.02 of the corresponding resolution.*

to note that in the implementations of the *Gradient* and *Isophote* detector the input image was by default already downsampled to 80×35 pixels. Thus the performance for those algorithms remains constant, except for the smallest resolutions. As one can see, the other algorithms all start to drop significantly in performance at some point while decreasing the resolution, until the performance becomes equal or worse to the previously mentioned method. Interestingly, the performances of *Swirski* and *ExCuSe* improved when downsampling from 480p to 240p. This indicates that 240p resolution is already enough for those methods, and higher resolution can harm performance, possibly due to increased image noise.

4 Discussion

In this paper we presented a novel dataset for the development and evaluation of pupil detection algorithms. Our goal was to collect a comprehensive set of unconstrained high-quality recordings in realistic day-to-day environments and to go beyond the difficulties provided by other existing datasets. Also, we evaluated the performance of state-of-the-art algorithms on our dataset. As the evaluation has shown, none of the examined algorithms performs well on all parts of the dataset. The detection accuracy in at least half of all cases was not sufficient to ensure a good eye tracking performance. This finding highlights the general difficulty of pupil detection in day-to-day environments and indicates the need to improve upon current algorithms. In addition, we identified several key challenges in those environments that facilitate further research. Namely, the presence of glasses and makeup, and the presence of strong sunlight, were shown to be severe problems for current algorithms. Further, the influence of image resolution has been evaluated. Given its high quality, size and difficulty, we believe our dataset serves as a good benchmark for evaluating new algorithms.

5 Conclusion

We presented labelled pupils in the wild (LPW), a novel dataset of eye region videos for the development and evaluation of pupil detection algorithms. Our dataset includes 66 ground-truth-annotated, high-quality videos (130,856 frames) recorded from 22 participants in everyday locations at about 95 FPS; it is one order of magnitude larger than existing datasets. Performance evaluations on the dataset demonstrated fundamental limitations of current pupil detection algorithms and highlighted key challenges of head-mounted pupil detection due to lighting, image resolution, and vision aids.

Acknowledgements

This work was funded, in part, by the Cluster of Excellence on Multimodal Computing and Interaction (MMCI) at Saarland University, the Alexander von Humboldt Foundation, as well as a JST CREST research grant.

References

HANSEN, D. W., AND JI, Q. 2010. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32*, 3, 478–500.

JESORSKY, O., KIRCHBERG, K. J., AND FRISCHHOLZ, R. W. 2001. Robust face detection using the hausdorff distance. In *Audio-and video-based biometric person authentication*, 90–95.

KASSNER, M., PATERA, W., AND BULLING, A. 2014. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Adj. Proc. UbiComp 2014*, 1151–1160.

MANSOURYAR, M., STEIL, J., SUGANO, Y., AND BULLING, A. 2016. 3D Gaze Estimation from 2D Pupil Positions on Monocular Head-Mounted Eye Trackers. In *Proc. ETRA*.

ŚWIRSKI, L., BULLING, A., AND DODGSON, N. 2012. Robust real-time pupil tracking in highly off-axis images. In *Proc. ETRA*, 173–176.

TIMM, F., AND BARTH, E. 2011. Accurate eye centre localisation by means of gradients. In *Proc. VISAPP*, 125–130.

VALENTI, R., AND GEVERS, T. 2012. Accurate eye center location through invariant isocentric patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence 34*, 9, 1785–1798.

W. FUHL, T. C. KBLER, K. S. W. R. E. K. 2015. Excuse: Robust pupil detection in real-world scenarios. In *Proc. CAIP 2015*.