

Labeled 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, xcchang, 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 with different ethnicities, 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, as well as 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, 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: I.4.9 [Image Processing and Computer Vision]: Applications;

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 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 systems are used in mobile everyday settings and eye images can therefore become subject to significant influences by changes in ambient light, corneal reflections, pupil occlusions, and shadows (see Figure 1). Despite considerable advances, we argue that methods for pupil detection on head-mounted eye trackers lack behind. When analysing current benchmark datasets, we identified two main limiting factors.

First, several existing datasets were recorded using remote cameras and only consist of 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 – that, for example, also include transitions of users between indoor and outdoor environments – or only contain 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 center. 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

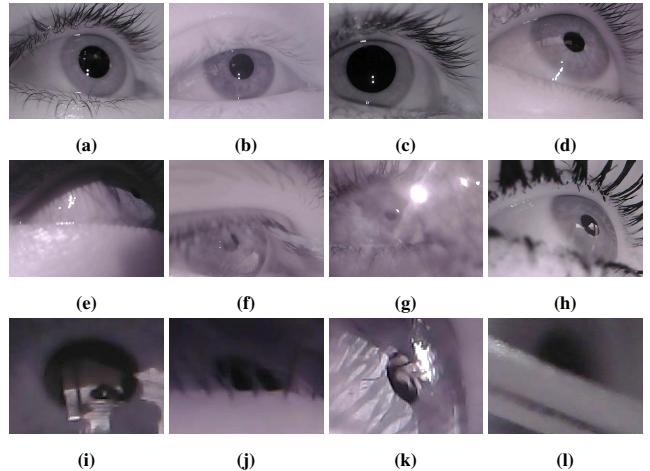


Figure 1: Example images of variability in our dataset. The first row (a) (b) (c) and (d) shows different eye appearances. The second row shows the most difficult cases according to our evaluation: (e) strong shade, (f) eyelid occlusion, (g) reflection on glasses, (h) strong makeup. The third row shows crops around pupil under challenging conditions: (i) reflection on the pupil, (j) self occluded, (k) strong sunlight and shade, (l) occlusion by glasses.

with controlled lighting conditions. A more recent dataset was introduced in [W. Fuhl 2015]. The dataset is significantly larger than the first dataset and images were recorded with a head-mounted eye tracker in uncontrolled environments, namely while driving and going shopping, but not in fully outdoor environments.

In this paper we therefore present *labelled pupils in the wild* (LPW), a novel pupil detection dataset that aims to address these shortcomings. 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 one order of magnitude larger than existing ones and covers a wide variety of realistic indoor and outdoor illumination conditions, include participants wearing glasses and eye make-up, as well as cover different ethnicities with variable skin tones, eye colours, and face shapes. All videos were manually ground-truth annotated with accurate pupil ellipse and center 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 will be made publicly available upon acceptance.

	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.

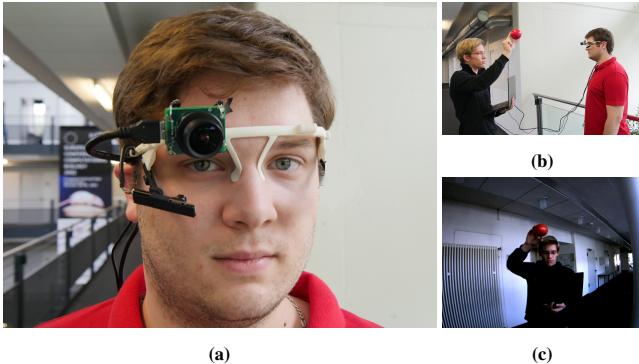


Figure 2: Data collection setting. (a) The high frame rate eye and scene cameras. (b) Participants move their eye by looking at the red ball. (c) The image captured by the scene camera.

2 Labelled pupils in the wild (LPW) dataset

We designed a data collection procedure with two main goals in mind: 1) to record samples of participants under different conditions, i.e. different lighting conditions and eye camera positions, and 2) to have a large variability in appearance of participants, such as gender, ethnicity and use of vision aids. We took each participant to a different set of locations and recorded their eye movements while looking at a moving gaze target.

Participants

Detailed information about our participants can be found in Table 2. We recruited 22 participants including 9 female through university mailing lists and personal communication. Among them are five different ethnicities: 11 Indian, 6 German, 2 Pakistani, 2 Iranian, and 1 Egyptian. In total we had five different eye colors: 12 brown, 5 black, 3 blue-gray, 1 blue-green, 1 green. Also 5 people had impaired vision, 2 wore glasses and 1 wore contact lenses. Strong eye make-up was worn by 1 person (with participant ID 22).

Apparatus

The eye tracker used for the recording was a high-speed Pupil Pro head-mounted eye tracker that record eye videos with 120 Hz [Kassner et al. 2014]. In order to capture high frame rate scene videos, we replaced the original scene camera with a PointGrey Chameleon3 USB3.0 camera recording at up to 149 Hz. The hardware set up is shown in Figure 2a and Figure 2b. It allowed us to record all videos with 95 FPS, which is a speed at which even fast eye movements last through several frames.

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. The position of the red ball in the visual field of the participant 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 the recording locations in and around of several buildings. Each location was not chosen more than once during the whole 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 participant we recorded three videos with around 20 seconds 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 center 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, we examined the annotated videos again to find wrong annotations, and corrected them by selecting 5 or more points on the pupil boundary and fitting an ellipse to them. The center of the ellipse was used as a refined pupil center position.

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. *Świrski* [Świrski et al. 2012] and *ExCuSe* [W. Fuhl 2015] are taken as examples of the 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. 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, *Świrski* 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*, *Świrski* and *ExCuSe* all return very good results in roughly 30% of all cases with less than 5px error; however their performances fall off quickly. It is worth mentioning that *ExCuSe* falls off last. The *Gradient* detector

	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	
Eye color	Brown	Blue	Brown	Brown	Brown	Black	Black	Brown	Blue-gray	Brown	Brown	
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
	2	1	2	1	2	1	1	2	2	1	2	1
	Nat	Art	Nat	Art	Nat	Art	Nat	Art	Nat	Art	Nat	Art
	3	1	3	1	2	2	2	1	3	1	2	1

	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	
Eye color	Brown	Black	Brown	Blue-gray	Green	Brown	Brown	Black	Black	Brown	Blue-gray	
Glasses	No	No	No	No	Contact lenses	No	No	No	No	No	No	
Video variability	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out
	2	1	2	1	2	1	2	1	2	1	2	1
	Nat	Art	Nat	Art	Nat	Art	Nat	Art	Nat	Art	Nat	Art
	3	1	3	1	2	1	3	1	3	1	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 videos is represented as indoor (In) and outdoor (Out), with natural (Nat) and artificial (Art) light.

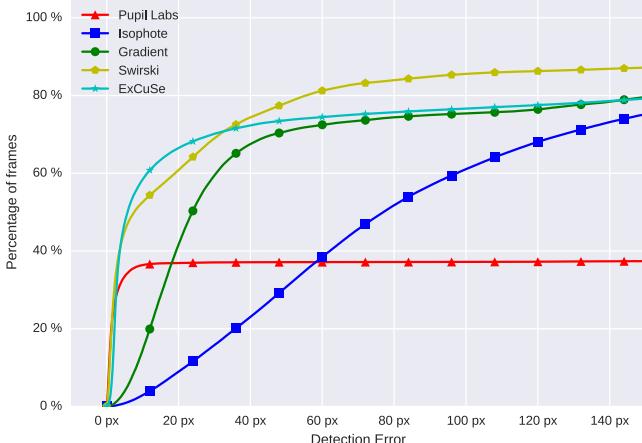


Figure 3: Cumulative error distribution of each algorithm on the entire dataset. The x-axis describes the error in pixels, while the y-axis describes the percentage of detections that achieved an error smaller or equal to the corresponding x-value.

follows a similar curve but shifted to the right, indicating a higher error on average. The *Isophote* detector's curve rises the least steep 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 there is no satisfying performance on the dataset yet 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 strong 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 intense 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 on indoor scenes roughly 60% of all detections had an error of 50px or lower, on outdoor scenes it is only about 50%.

Glasses and Makeup

For users with impaired eyes, the possibility to wear glasses along with the eye tracker is very important. However, glasses can cause intense reflections in the images 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 ones 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. One could expect this, since all algorithms either look for large black blobs or strong edges, which both could be also created by makeup.

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

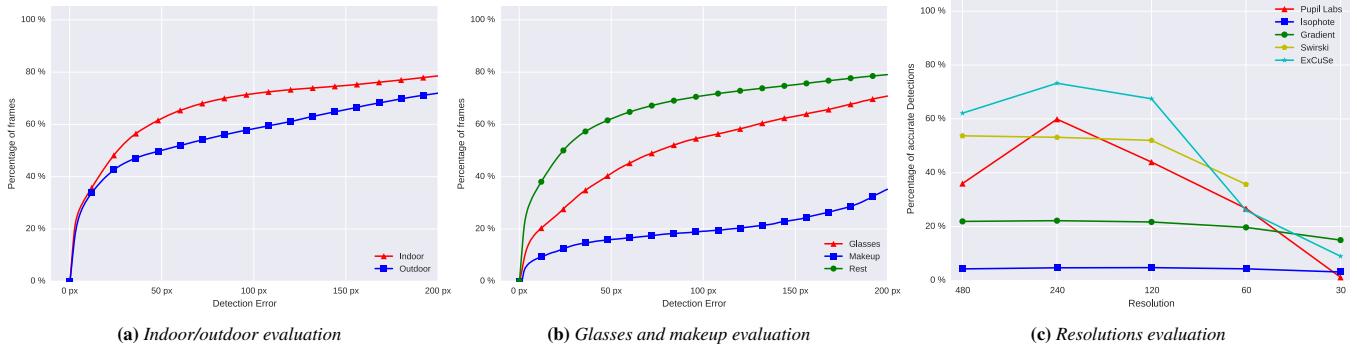


Figure 4: Performance over different factors. Cumulative mean error distribution for indoor and outdoor videos of the 5 algorithms (a). The x-axis describes the detection error in pixels, while the y-axis describes the percentage of detections that had an error equal or lower to the corresponding x-value.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

images while the others are usually for higher resolutions. In Figure 4c, we show the performance of each algorithm for different resolutions. The error is normalized by image width, and the percentage of detections with an error lower than 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 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 former mentioned method. Interestingly, the performances of *Swirski* and *ExCuSe* improved when downsampling from 480p to 240p. It indicates that 240p resolution is already enough for those methods, and higher resolution can harm the 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 performed 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 highlights the general difficulty of pupil detection in day-to-day environments and indicates the need to improve upon current algorithms. Further we were able to identify some of the key-challenges in those environments, which can give researchers an idea about what problems to focus on. Especially the presence of glasses and makeup could be shown to be a severe problem for current algorithms. Also the difficulty of performing on images recorded outdoors was highlighted in comparison to images recorded indoors. Further the influence of image resolution has been evaluated. While this identification of challenges is not yet complete, it highlights many open problems and can serve as a reference when developing new approaches. Given its high quality, size and difficulty, our dataset serves as a good benchmark for evaluating new algorithms. The videos have been recorded in realistic day-to-day environments, however the actual viewing behaviour of

the participant was controlled via a gaze target and is thus not natural. Given the videos high FPS, the development of tracking based algorithms can be considered.

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.

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. of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2014)*, 1151–1160.
- ŚWIRSKI, L., BULLING, A., AND DODGSON, N. 2012. Robust real-time pupil tracking in highly off-axis images. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ACM, New York, NY, USA, ETRA ’12, 173–176.
- TIMM, F., AND BARTH, E. 2011. Accurate eye centre localisation by means of gradients. In *VISAPP*, SciTePress, L. Mestetskiy and J. Braz, Eds., 125–130.
- VALENTI, R., AND GEVERS, T. 2012. Accurate eye center location through invariant isocentric patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 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*.