COGAIN Symposium
Wuppertal, August 21st 2017.

CrowdPupil: A crowdsourced, pupil-center annotated image dataset

David Gil de Gómez Pérez
School of Computing, University of Eastern Finland

Roman Bednarik
School of Computing, University of Eastern Finland

We present a dataset of pupil images and associated hand-annotated pupil centers, obtained through the method of crowd-sourcing. Acquisition of the points is explained and the dataset is presented. We present a comparison of two state-of-the-art pupil detection algorithms as a proposal towards public benchmarking of pupil detection algorithms. We invite the eye tracking community to test their own algorithms, share the results, and thereby advance the domain systematically. We present our plans for organizing a public pupil-detection challenge.

Keywords: eye-tracking, feature annotation, crowdsourcing, dataset

Introduction

A need for an accurately annotated pupil dataset raises at early stages of any pupil detection algorithm development project. Traditionally this crucial step is approached by a single annotator, using a point-and-click interface. What is not the most engaging of tasks quickly can become straining. Such method can have serious methodological pitfalls. The annotator's perception and motor skills can introduce biases that, at the end, may introduce errors. Annotations created by a single person lack inter-rater agreement and uncertainty is hard to estimate. The fact of it being a monotonic, repetitive task can cause that the annotator gets tired over time when confronting long annotation sessions, which can be another source of lack of concentration and accuracy.

Human error can affect both the annotation and the ultimate accuracy measurements and evaluation of the pupil tracking algorithms. We have recently proposed to involve multiple annotators recruited from Internet volunteers (Gil de Gómez Pérez & Bednarik, 2017). This enabled us to reduce those pitfalls to the point of elimination, by getting input from a large number of different users for a single image.

Existing Datasets

Existing available datasets of annotated pupil centers include those involved in development of the ExCuSe (Fuhl, Kübler, Sippel, Rosenstiel, & Kasneci, 2015) and Else (Fuhl, Santini, Kübler, & Kasneci, 2016) algorithms, and the dataset developed at the Max Planck Institute (Tonsen, Zhang, Sugano, & Bulling, 2016). Neither of them addresses the pitfalls of traditional annotations that we try to eliminate by distribution of microtasks.

CrowdPupil: A public dataset

The mechanisms of engaging a large number of online users to annotate a pupil dataset were described in (Gil de Gómez Pérez & Bednarik, 2017). In brief, the users enter the system through an on-line portal, which tracks the contributions of each user, engages them to contribute more through a competition, and stores the input points into a database. When enough points for each image are collected, the final results in a form of a point cloud are filtered for outliers and the system calculates the final centroids.

Image Preprocessing

The images used in this case were obtained from the UTIRIS image database (Hosseini, Araabi, & Soltanian-Zadeh, 2010). We used a subset containing the images that made use of the near infrared (NIR) lighting. The subset comprises of 792 images from 79 individuals taken from both eyes. The original images are stored as grayscale bitmaps, and can be downloaded at `https://utiris.wordpress.com/`.

The images in the original dataset are in a high resolution, detrimenting the user experience by increasing loading times. As one of the main goals was to maximize the number of point-inputs, we tried to reduce such discomfort by compressing the images in JPG format without apparent change. The compressed images are provided with the dataset.

User Input

For each image, we initialized the input set with 8 random points, two in each quadrant of the plane surrounding an estimated center. The initialization points were not taken into consideration when calculating the final centroids. This enabled us to reject some points directly after the user input them, and boost the amount of valid points that reached the dataset.

The task was explained to the user before the start of the operation, and consisted of annotating the center of the pupil by using a single click, according to the user's best perception. The number of users participating in the annotation was, on average, 13.97 per image (SD = 1.12), with

a minimum of 8, a maximum of 16. The mechanism to select the next image for annotation was designed to maximize the number of different people annotating each image. This mechanism was transparent to the user.

Each image was annotated a mean of 1.11 times per user per image, some images were annotated by the same person more than once. On 156 images, every click was done by a different person. Considering that each image was annotated at least 15 times, these results were considered as an indicator of eliminating personal biases on the pupil center perception.

The annotated dataset can be downloaded from `http://cs.uef.fi/pupoint/`.

Benchmarking Procedure

Another goal of this work is to generate a discussion of a procedure to benchmark different algorithms to compare their accuracy in a standardized way in order to be able to create a public pupil detection algorithm challenge with any public annotated dataset. The following considerations were made when designing the benchmarking method:

- It should provide comparable accuracy measurements regardless of the size of the images in the dataset, though a uniform image size throughout the image set can be assumed.
- The calculations involved in the final assessment should be reproducible and dependent only on the calculated pupil center and the annotated one.
- The base of the calculations should be a well-defined metric on $\mathbb{R}^2$, as the images can be considered two-dimensional. Specifically, it should give a measurement of the distance between both studied points, understanding it as the length of the path connecting them.

The Euclidean metric was chosen as the base of the calculations. The following steps provide a methodology that implements all the requirements, supposing that all the images on a given dataset have the same height and width. The term "distance" should be understood as a synonym of the result of calculating the Euclidean metric between the two bidimensional vectors defined by the given points that belong to the two-space of the image.

First, the distance in pixels between each the annotated and the calculated point is determined for each of the points, as per the standard calculation method. Then, the arithmetic mean of those distances is calculated in order to obtain the number that will statistically reflect the performance of the algorithm. This number is still dependent on the size of the original image. To avoid this effect, we normalize the number according to the size of the image by using the following formula: $SIPM = \frac{APDC}{AMIC}$.

$SIPM$ refers to a Size Independent Performance Metric, $APDC$ is the Area of the Point-Defined Circle and $AMIC$ is the Area of the Maximum Inscribed Circle.

$APDC$ is the area of the circle whose radius is the arithmetic mean distance that was calculated in the first step. For the calculation of $AMIC$ we assumed rectangular images and considered the diameter of the circle the minimum of the height and width of the image, as it defines the size of the maximum inscribed circle. If the images are not rectangular this area must be calculated using other geometric methods.

$SIPM$ provides a dimensionless ratio that can be used to compare algorithms. The smaller this ratio is, the more accurate the algorithm can be considered.

These measures can be combined with other metrics to obtain a more general overview. Other parameters, can be, for example, execution time and memory usage.

Example of Benchmarking

We choose two algorithms ExCuSe (Fuhl et al., 2015) and ElSe (Fuhl et al., 2016). As explained above, the $SIPM$ was calculated for both of them. The images in this dataset are rectangular, with a height of 776 pixels and a width of 1000 pixels. Hence the maximum inscribed circle has a diameter of 776 pixels and an area of 472948 px^2.

As the ExCuSe algorithm specifies in its definition that the optimal image has a width of 384 pixels, we scaled the images down prior to the execution of the algorithm and then we rescaled them up, projecting the points to the real image so the obtained results express optimal executions of the algorithm. For the execution, the C++ public implementation was used in both cases. The results can be seen in Table 1.

Table 1
Benchmarking for the ExCuSe and ElSe algorithms

	Mean D. (px)	$APDC$ (px^2)	$SIPM$ (%)
ExCuSe	9.56	281.12	0.061
Else	14.8	688.13	0.145

We can conclude that under these conditions ExCuSe is more accurate than ElSe. These surprising results can be due to the resizing step explained before, that only applied to ExCuSe, as ElSe already contains a resizing step.

Discussion and Conclusions

The demanding work behind annotating a pupil dataset, in our opinion, may be one of the reasons behind the lack of such publicly available datasets. This is certainly delaying the progress of the discipline. We designed a way to crowdsource annotations and annotated a case dataset, as well as proposed a benchmarking procedure. We hope to create discussion towards sound benchmarking for systematic comparison. The system can be adapted to annotate other features, such as canthi or pupil radius. We also plan to use this procedure for the design and creation of a public pupil-detection challenge.

References

Fuhl, W., Kübler, T., Sippel, K., Rosenstiel, W., & Kasneci, E. (2015). Excuse: Robust pupil detection in real-world scenarios. In *Int. conf. on computer analysis of images and patterns* (pp. 39–51).

Fuhl, W., Santini, T. C., Kübler, T., & Kasneci, E. (2016). Else: Ellipse selection for robust pupil detection in real-world environments. In *Etra'16* (pp. 123–130).

Gil de Gómez Pérez, D., & Bednarik, R. (2017). Ponline: An online pupil annotation tool employing crowd-sourcing and engagement mechanisms. *In Review*.

Hosseini, M. S., Araabi, B. N., & Soltanian-Zadeh, H. (2010). Pigment melanin: Pattern for iris recognition. *IEEE Trans. Instrumentation and Measurement*, *59*(4), 792–804.

Tonsen, M., Zhang, X., Sugano, Y., & Bulling, A. (2016). Labelled pupils in the wild: a dataset for studying pupil detection in unconstrained environments. In *Etra'16* (pp. 139–142).