SeTA: Semiautomatic Tool for Annotation of Eye Tracking Images

Andoni Larumbe-Bergera
Public University of Navarre
Pamplona, Spain
andoni.larumbe@unavarra.es

Sonia Porta
Public University of Navarre
Pamplona, Spain
sporta@unavarra.es

Rafael Cabeza
Public University of Navarre
Pamplona, Spain
rcabeza@unavarra.es

Arantxa Villanueva
Public University of Navarre
Pamplona, Spain
avilla@unavarra.es

ABSTRACT

Availability of large scale tagged datasets is a must in the field of deep learning applied to the eye tracking challenge. In this paper, the potential of Supervised-Descent-Method (SDM) as a semiautomatic labelling tool for eye tracking images is shown. The objective of the paper is to evidence how the human effort needed for manually labelling large eye tracking datasets can be radically reduced by the use of cascaded regressors. Different applications are provided in the fields of high and low resolution systems. An iris/pupil center labelling is shown as example for low resolution images while a pupil contour points detection is demonstrated in high resolution. In both cases manual annotation requirements are drastically reduced.

CCS CONCEPTS

• **Applied computing** → **Annotation**; • **Computing methodologies** → **Supervised learning by regression**.

KEYWORDS

image annotation, eye tracking, Supervised-Descent-Method

ACM Reference Format:
Andoni Larumbe-Bergera, Sonia Porta, Rafael Cabeza, and Arantxa Villanueva. 2019. SeTA: Semiautomatic Tool for Annotation of Eye Tracking Images. In *2019 Symposium on Eye Tracking Research and Applications (ETRA '19), June 25–28, 2019, Denver , CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3314111.3319830

1 INTRODUCTION

Typically, the problem of estimating gaze has been divided into two issues, namely, eye tracking and gaze estimation. Eye tracking is related to the algorithms focused on processing the acquired eye image to obtain image features e.g. iris or pupil center, glints, eyelids, etc. while gaze estimation covers the challenge of finding gaze from the image. As in many other computer vision problems deep learning techniques can be used to solve both problems as demonstrated in some of the works published in the last few years [Krafka et al. 2016] [Zhang et al. 2018] [Park et al. 2018].

One of the basic requirements for any deep learning procedure is the availability of large scale labelled datasets to be used during the training stage. The procedure to obtain these datasets is not trivial, moreover the possibility of having the images labelled is not completely solved yet. In the case of gaze estimation methods, the labelling procedure consists in tagging each one of the images with gaze data, e.g. 2D PoR value or 3D LoS. This operation can be carried out by employing previously agreed gaze points or gaze directions according to the case. Basically, the user is asked to gaze known grids of points. It is assumed that the subject gazes the corresponding point for a dwell time. In this manner, the images are tagged automatically provided that a synchronization procedure is established between the displayed points and the image recording thread. In the case of eye tracking methods, labelling is a broader problem since the required marks vary depending on the algorithm, ranging from pupil or iris center to iris contour, eye corners or eyelids among others. Moreover, the labelling procedure is not straightforward. The most obvious but tedious way to solve the labelling problem is to carry out a manual marking procedure for which dedicated tools can be designed [Fuhl et al. 2017]. Considering the required size of the datasets for deep learning this is not a practical solution. In fact, there are companies devoted to image tagging tasks being this a business showed up as result of the higher demand of labelled datasets to be employed in deep/machine learning fields. In this manner, human effort is translated into dollars. More practical proposals based on using synthesized images are found in the bibliography [Sugano et al. 2014]. One of them is data augmentation, properly defined as the process of increasing the number of data/images of datasets by means, generally, of artificial techniques. This can involve changes such as image rotation, introducing lighting variations in the images, varying the degree of noise conditions, etc., generating different sub-samples from the same original image. Moreover simulators can be employed in which camera, user, gazed points and light sources are simulated. Thus, the image is artificially generated and the labels corresponding to image features are known by construction. Examples devoted to high and low resolution eye tracking can be found in the literature [Świrski and Dodgson 2014] [Wood et al. 2016]. Finally, employing image processing techniques to partially automatize the annotation process has also been proposed in the bibliography [Tonsen et al. 2016].

In this paper a Semiautomatic Tool for Annotation of Eye Tracking images, SeTA, based on Supervised-Descent-Method (SDM) [Xiong and la Torre 2014] is proposed. SDM can be used as feature detection algorithm based on a training stage and has demonstrated

ETRA '19, June 25–28, 2019, Denver , CO, USA

ACM ISBN 978-1-4503-6709-7/19/06.
https://doi.org/10.1145/3314111.3319830

to provide highly accurate results in low resolution eye tracking images [Larumbe et al. 2018]. In our work, we propose to use dedicated SDMs for individual sessions. The SDM tracker is pre-trained using a reduced number of selected images of a tracking session, so that detection is adapted to the particular characteristics of that session. Two possible applications are shown: first, the potential of this technique to label iris center for low resolution images employing MPIIGaze [Zhang et al. 2018] and I2Head [Martinikorena et al. 2018] datasets is shown. Second, a high resolution application is demonstrated by labelling pupil contour and center in images obtained from a head mounted eye tracker, such as the ones provided in Labelled Pupils in the Wild (LPW) dataset [Tonsen et al. 2016]. We would like to make clear that the terms high and low resolution refer to the number of pixels in the pupil area. In the next section, the basics of SDM and the algorithm proposed are presented. To follow, the databases and the implementations proposed in each one of the cases are described. Finally, results and conclusions are shown.

2 SUPERVISED DESCENT METHOD

Supervised-Descent-Method (SDM) is a minimization technique for Nonlinear Least Squares (NLS) problems proposed by X. Xiong and F. de la Torre [Xiong and la Torre 2014] in the field of computer vision. Later, Z. Feng et al. [Feng et al. 2015] proposed a modification of the SDM algorithm by means of including an adaptive scheme for scale-invariant updates and a regularization term. During training, SDM learns a cascaded regressor (**CR**) from a set of L images labeled with a group of ground truth landmarks $\{\mathbf{x}_*^i\}(i = 1, \cdots, L)$. A cascade regressor **CR** is formed by K weak regressors $\mathbf{r}_k$ in cascade. Each weak regressor is represented by a descent map $\mathbf{R}_k$ and a bias term $\mathbf{b}_k$ which are computed in each iteration k by minimizing the expected loss between the ground truth landmarks $\mathbf{x}_*^i$ and the previous iteration predicted landmarks $\mathbf{x}_{k-1}^i$, given by

$$\sum_{i=1}^{L} \|\mathbf{x}_*^i - \mathbf{x}_{k-1}^i + \mathbf{R}_k \mathbf{h}(\mathbf{d}^i(\mathbf{x}_{k-1}^i)) - \mathbf{b}_k\|_2^2 + \lambda \|\mathbf{R}_k\|_F^2, \qquad (1)$$

where $\mathbf{d}$ is the image, $\mathbf{h}$ is a nonlinear feature extraction function (Histogram of Oriented Gradients in our case) and λ is the weight of the regularization term. Once $\mathbf{R}_k$ and $\mathbf{b}_k$ have been estimated, each sample $\mathbf{x}_{k-1}^i$ is updated to its new location $\mathbf{x}_k^i$ as follows:

$$\mathbf{x}_k^i = \mathbf{x}_{k-1}^i - \mathbf{R}_k \mathbf{h}(\mathbf{d}^i(\mathbf{x}_{k-1}^i)) + \mathbf{b}_k. \qquad (2)$$

After an update, we recompute a new descent map $\mathbf{R}_{k+1}$ and a new location $\mathbf{x}_{k+1}^i$. In testing, landmarks locations $\mathbf{x}_k^i$ are updated recursively using Equation 2, starting from a landmark initialization $\mathbf{x}_0^i$ and employing the learned **CR** [Larumbe et al. 2018].

3 SEMIAUTOMATIC LABELLING PROPOSALS

We decide to apply the annotation tool, SeTA, to two different scenarios, namely, low and high resolution tracking frameworks. An application for labelling pupil/iris center in low resolution is shown while a pupil contour and center annotation algorithm is presented for high resolution.

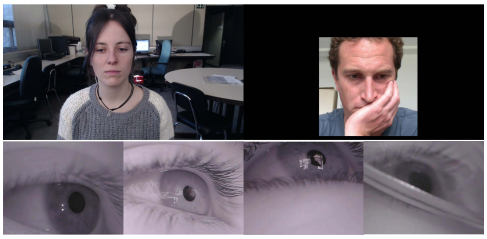

Figure 1: Samples extracted from I2Head (upper-left), MPIIGaze (upper-right, https://bit.ly/1HO7MoR) and LPW (lower, https://bit.ly/2TUqjYw) datasets.

3.1 Low resolution: MPIIGaze & I2Head

MPIIGaze and I2Head are two databases conceived, designed and built under the paradigm of low resolution eye tracking and gaze estimation. Both datasets contain images from several users (15 and 12 users, respectively). A large number of images per user is available, so that the approach proposed in this case consists in training independent SeTAs for each one of the users. To this end, a bunch of labelled images from a solely user is assigned to train that individual's model, which is going to be used later to detect pupil center in the remaining images from the same user. The underlying assumption is that the SDM will get adapted to the user's own characteristics and it will be able to perform the automatic annotation process on the non-trained images.

MPIIGaze contains about 200,000 images exhibiting a great variety of lighting, blurring and positioning conditions. The number of images per user is variable, as it is the accuracy of the iris center manual labels provided for a subset of 10,000 images. To overcome the problems arisen from unreliable labels a random selection of 40 images per user was performed, and three different contributors provided their six labels per image, namely, two corners and the pupil center for both eyes. Mean values were considered to build a subset of 40 annotated images for each user: 20 annotated images devoted to training the respective model, and 20 devoted to estimating the model accuracy in the testing stage. A second subset of 100 alternative images (also randomly chosen but not annotated) per user is constructed with the aim of providing a visual evaluation.

I2Head dataset contains images of 12 users, gazing at alternative points in a screen from varying spatial positions. Images were recorded in four centered sessions using static and free head scenarios, plus four other sessions including user displacements. Only the centered sessions are considered for this work at a rate of only one image per gazed point (although 10 images per point were acquired), up to 164 images per user. Though labels for gazed points are provided, this dataset does not supply image landmarks so, again, a team of three individuals generated six labels per image, marking both eye corners and pupil centers, as it had been done for the MPIIGaze annotated subset. Mean values were considered to build this subset of 164 annotated images for each user. A subset of 17 images per user is used for training while the remaining 147 images are employed in the testing stage (see figure 1 (top)).

The SeTA proposed for these datasets trains a personalized SDM tracker using the subset of annotated images where the eye corners are accurately known. The adjusted model is then applied upon

the user's remainder images to detect pupil centers. Some smart strategies are added in the annotation stage for both datasets:

- A well known data augmentation procedure is implemented by flipping the images contained in the training subset [Krizhevsky et al. 2012].
- The initial shape $\mathbf{x}_0^i$ is set according to the ground truth labels of eye corners attached to images included in the training subset.
- Robustness against inaccurate initialization is achieved generating from each original image a 100 copies where random noise in the eye landmarks is introduced. In this way the SDM learns the minimization strategy in a higher number of starting conditions.
- The hyperparameters are tuned resulting in five weak regressors ($K = 5$) and a regularization term of 5 ($\lambda = 5$).

3.2 High resolution: LPW

Eye tracking using head mounted systems in everyday and outdoor activities is still a challenge. Labelled Pupils in the Wild (LPW) dataset [Tonsen et al. 2016] is a high resolution dataset containing 66 videos recorded from 22 participants in everyday situations employing a dark-pupil head-mounted eye tracker in which labels are provided for the pupil center. LPW includes subjects with different ethnicities, varying illumination, strong reflections occluding the pupil and moved images. In figure 1 (bottom) samples extracted from LPW are shown. Detecting pupil contour points is key for accurate eye tracking in order to refine pupil center detection or for those gaze estimation methods based on the shape of the pupil. The SeTA proposed in this scenario is to train independent SDM trackers for each one of the videos to detect pupil contour points. The high degree of variation between videos strengthen the idea of creating dedicated trackers for each one of the sessions. In other words, the SDM tracker is trained to detect 20 homogeneously distributed points in the pupil contour. Given a video of 2000 frames, 10 uniformly distributed frames are selected to be annotated manually. The annotator is asked to mark 8 points in the pupil contour to which an ellipse is adjusted and 20 landmarks are extracted. Using this subset of images the model is constructed. Once the SDM-tracker is trained for the specific session it is applied to the whole video to get the frames automatically labelled. The annotation strategy presents the following characteristics:

- The initial shape, $\mathbf{x}_0^i$ for a frame, is set according to the landmarks detected on the previous frame (the first frame is assumed to be always manually annotated).
- The 20 landmarks are geometrically coherent with respect to the image, i.e. the first point is always the one with the highest vertical coordinate.
- In order to increment the robustness against an inaccurate initialization, systematic noise is introduced in the contour points during the training. For each one of the training frames 50 samples are created by introducing a controlled perturbation in the pupil ellipse, allowing SeTA to learn the minimization strategy in a higher number of initialization conditions. Instead of perturbing each tracked point independently the parameters of the ellipse are perturbed trying to simulate potential movements of the pupil between frames.
- The annotation procedure needs to be supervised by an operator as the labelling progresses throughout the video. When the operator detects a bad annotation, the process can be interrupted and the non correctly annotated image is labelled manually. The new labelled image is used to update the model in an online fashion. A higher number of initializations (200) is set for the image the first time the sample is included into the model to increment its weight. Once the model is updated, the annotation procedure continues.
- Every time the contour points are detected, an ellipse is fitted in order to eliminate possible outliers and 20 landmarks are re-calculated according to the equation of the ellipse as the final labels for that image.
- The selected CR is formed by two weak regressors ($K = 2$) and the regularization term used is unity ($\lambda = 1$).
- Data augmentation is not performed in this case.

4 RESULTS

In this section SeTA results are evaluated according to the low resolution and high resolution tests carried out. In the experiments regarding MPIIGaze and I2Head datasets quantitative and qualitative results are shown. Quantitative values can be extracted comparing the centers provided by the manual annotation and the ones obtained by SeTA. To evaluate accuracy the absolute error is calculated as the Euclidean distance between pupil center estimates and ground truth values provided, then it is normalized relative to the inter-pupillary distance (ground truth). This is formulated by $e_{max} = \frac{max(d_{left}, d_{right})}{\omega}$ where d_{left} and d_{right} are the absolute errors for the eye pair, and ω is the inter-pupillary distance. The maximum between d_{left} and d_{right} after normalization is defined as *maximum normalized error* e_{max}. Accuracy is calculated as the percentage of images for which the error is below specific e_{max}. In figure 2 the distributions of the error considering subjects belonging to each one of the datasets are shown. The variability between annotators is provided in green for both datasets as representative of the best accuracy potentially achievable. The closeness between both curves resembles the good performance of SeTA.

Regarding qualitative evaluation, 100 random images per user are annotated using SeTA. Three different observers evaluate the results and decide whether the center was accurately estimated or not. For MPIIGaze they estimate that 7.26±7.36 (average ± standard deviation) of labels per user are inaccurately calculated, meaning that less than 5% of the cases would need to be relabelled. We would like to emphasize that even in non correctly initialized images, SeTA is able to retrieve an accurate annotation in the majority of the cases demonstrating its robustness to initial conditions. For I2Head the qualitative evaluation shows that only 3.66±3.91 of labels can be considered to be inaccurately annotated. The remarkable improvement with respect to MPIIGaze is due to the fact that the variability among the images belonging to the same user in I2Head is considerably lower compared to MPIIGaze. A sample of the results regarding MPIIGaze and I2Head is available in our **webpage**.

Four eye tracking experts select 15 random sessions from LPW dataset to be annotated employing SeTA according to the procedure explained in section 3.2. In figure 3 the average difference resulting from the Euclidean distance (pixels) between SeTA results and the

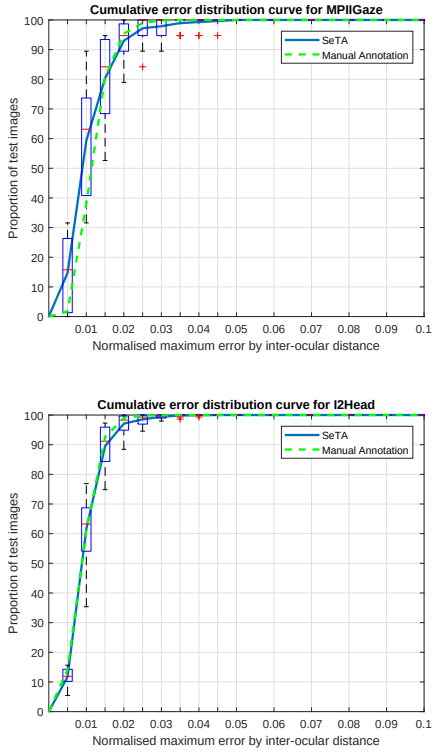

Figure 2: Cumulative error for MPIIGaze (top) and I2Head (bottom) datasets.

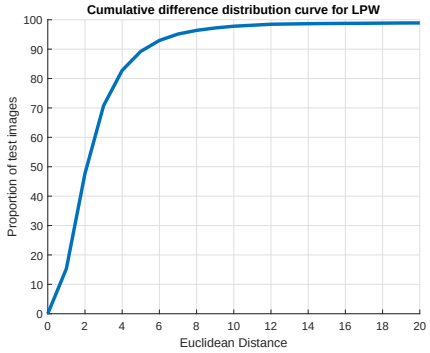

Figure 3: Cumulative difference for LPW dataset.

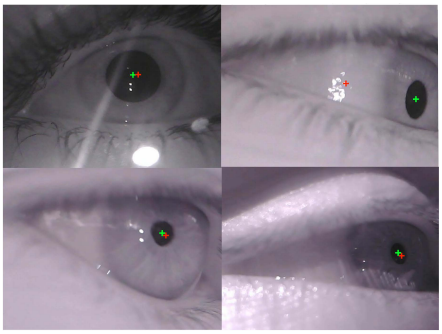

Figure 4: Samples comparing the center provided by LPW (red, https://bit.ly/2TUqjYw) and from SeTA (green).

labels provided in the dataset is shown for the selected sessions. Results shown in figure 3 should be further discussed if we take into account that the authors stated in the original paper that the labels for the pupil center were calculated employing semiautomatic image processing techniques. From the results obtained in our experiments it can be concluded that setting the pupil center as the center of the ellipse presents higher robustness and precision than those resulting from the solely pupil center annotation. In figure 4 we select some of the samples showing significant differences between SeTA annotation and the labels provided with LPW, leading us to reach the conclusion that the pupil center can be more accurately annotated if the corresponding ellipse is considered. Additional videos can be checked in our **webpage**. The accuracy values in figure 3 improve the ones found in the literature [Santini et al. 2018]. However, this comparison, although promising, is not completely fair since not all the videos have been annotated by SeTA yet.

Regarding the annotation requirements for LPW, the number of extra frames manually labelled in the online procedure is variable according to the session but the average, including the initial requirement of labelling 10 frames for a session containing 2000 images, is 23.36±13.13 manually annotated frames. Apart from accuracy values, the resulting numbers clearly show the significant human effort reduction. The annotation effort is converted into a supervision task requiring considerably less time. Supervising the annotation of an LPW image requires approximately 1 sec while it raises to 12-14 secs if a manual annotation is carried out, i.e. the spent time reduction is closer to 92%. Regarding low resolution task 6-8 secs are needed to annotate pupil centers, saving 83% of the time employed for manual annotation.

5 CONCLUSIONS

SeTA is a semiautomatic annotation tool for eye tracking images. In this work we have demonstrated the potential of SDM to be used for automated labelling processes. Two applications have been shown related to pupil contour points detection in high resolution systems and pupil center annotation in low resolution images. SeTA has demonstrated to reduce drastically the human effort requirements in order to get annotated datasets. We consider that the lack of large scale annotated databases prevent researchers to apply machine learning methods to eye tracking and gaze estimation fields in a proper way. Our work has tried to contribute to the challenge of moving the technology closer to this goal.

ACKNOWLEDGMENTS

We would like to acknowledge the Spanish Ministry of Science, Innovation and Universities for their support under Contract TIN2017-84388-R.

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

REFERENCES

Z.-H. Feng, P. Huber, J. Kittler, W. Christmas, and X.-J. Wu. 2015. Random Cascaded-Regression Copse for robust facial landmark detection. *Signal Processing Letters, IEEE* 22, 1 (Jan 2015), 76–80. https://doi.org/10.1109/LSP.2014.2347011

W. Fuhl, T. Santini, D. Geisler, T. Kübler, and E. Kasneci. 2017. EyeLad: remote eye tracking image labeling tool. In *12th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017)*.

K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba. 2016. Eye Tracking for Everyone. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 1097–1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

A. Larumbe, R. Cabeza, and A. Villanueva. 2018. Supervised descent method (SDM) applied to accurate pupil detection in off-the-shelf eye tracking systems. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research and Applications (ETRA'18)*. ACM, Warsaw, Poland, 7:1–7:8. https://doi.org/10.1145/3204493.3204551

I. Martinikorena, R. Cabeza, A. Villanueva, and S. Porta. 2018. Introducing I2Head database. In *Proceedings of the 7th Workshop on Pervasive Eye Tracking and Mobile Eye-Based Interaction (ETRA'18)*. Warsaw, Poland, 1:1–1:7. https://doi.org/10.1145/3208031.3208033

S. Park, X. Zhang, A. Bulling, and O. Hilliges. 2018. Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research and Applications (ETRA'18)*. ACM, Warsaw, Poland, Article 21, 10 pages. https://doi.org/10.1145/3204493.3204545

T. Santini, W. Fuhl, and E. Kasneci. 2018. PuReST: robust pupil tracking for real-time pervasive eye tracking. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research and Applications (ETRA'18)*. Warsaw, Poland, Article 61. https://doi.org/10.1145/3204493.3204578

Y. Sugano, Y. Matsushita, and Y. Sato. 2014. Learning-by-Synthesis for Appearance-Based 3D Gaze Estimation. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'14)*. IEEE Computer Society, 1821–1828. http://dblp.uni-trier.de/db/conf/cvpr/cvpr2014.html#SuganoMS14

L. Świrski and N. A. Dodgson. 2014. Rendering synthetic ground truth images for eye tracker evaluation. In *Proceedings of the 2014 ACM Symposium on Eye Tracking Research and Applications (ETRA'14)*. ACM, New York, NY, USA, 219–222. http://www.cl.cam.ac.uk/research/rainbow/projects/eyerender/

M. Tonsen, X. Zhang, Y. Sugano, and A. Bulling. 2016. Labelled Pupils in the Wild: a dataset for studying pupil detection in unconstrained environments. In *Proceedings of the 2016 ACM Symposium on Eye Tracking Research and Applications (ETRA '16)*. ACM, New York, NY, USA, 139–142. https://doi.org/10.1145/2857491.2857520

E. Wood, T. Baltrusaitis, L.-P. Morency, P. Robinson, and A. Bulling. 2016. Learning an appearance-based gaze estimator from one million synthesised images. In *Proceedings of the 2016 Symposium on Eye Tracking Research and Applications (ETRA'16)*. ACM, New York, NY, USA, 131–138. https://doi.org/10.1145/2857491.2857492

X. Xiong and F. De la Torre. 2014. Supervised descent method for solving nonlinear least squares problems in computer vision. *CoRR* abs/1405.0601 (2014). arXiv:1405.0601 http://arxiv.org/abs/1405.0601

X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. 2018. MPIIGaze: real-world dataset and deep appearance-based gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Early Access (2018). https://doi.org/10.1109/TPAMI.2017.2778103