Journal of Neuroscience Methods 324 (2019) 108307

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

DeepVOG: Open-source pupil segmentation and gaze estimation in neuroscience using deep learning

Yuk-Hoi Yiu[a], Moustafa Aboulatta[b], Theresa Raiser[a,d], Leoni Ophey[a], Virginia L. Flanagin[a], Peter zu Eulenburg[a,c], Seyed-Ahmad Ahmadi[a,*]

[a] German Center for Vertigo and Balance Disorders (DSGZ), Feodor-Lynen-Str. 19, 81377 Munich, Germany
[b] Faculty of Informatics, Technical University of Munich, Boltzmannstr. 3, 85748 Garching, Germany
[c] Department of Neurology, Klinikum der Universität München, Marchioninistr. 15, 81377 Munich, Germany
[d] Graduate School of Systemic Neurosciences, Ludwig-Maximilians Universität München, Großhadern Str. 2, 82152 Planegg, Germany

ARTICLE INFO

Keywords:
Video oculography
Pupil segmentation
Gaze estimation
blink detection
Deep learning
Convolutional neural networks

ABSTRACT

Background: A prerequisite for many eye tracking and video-oculography (VOG) methods is an accurate localization of the pupil. Several existing techniques face challenges in images with artifacts and under naturalistic low-light conditions, e.g. with highly dilated pupils.

New method: For the first time, we propose to use a fully convolutional neural network (FCNN) for segmentation of the whole pupil area, trained on 3946 VOG images hand-annotated at our institute. We integrate the FCNN into DeepVOG, along with an established method for gaze estimation from elliptical pupil contours, which we improve upon by considering our FCNN's segmentation confidence measure.

Results: The FCNN output simultaneously enables us to perform pupil center localization, elliptical contour estimation and blink detection, all with a single network and with an assigned confidence value, at framerates above 130 Hz on commercial workstations with GPU acceleration. Pupil centre coordinates can be estimated with a median accuracy of around 1.0 pixel, and gaze estimation is accurate to within 0.5 degrees. The FCNN is able to robustly segment the pupil in a wide array of datasets that were not used for training.

Comparison with existing methods: We validate our method against gold standard eye images that were artificially rendered, as well as hand-annotated VOG data from a gold-standard clinical system (EyeSeeCam) at our institute.

Conclusions: Our proposed FCNN-based pupil segmentation framework is accurate, robust and generalizes well to new VOG datasets. We provide our code and pre-trained FCNN model open-source and for free under www.github.com/pydsgz/DeepVOG.

1. Introduction

Many disciplines in clinical neurology and neuroscience benefit from the analysis of eye motion and gaze direction, which both rely on accurate pupil detection and localization as a prerequisite step. Over the years, eye tracking techniques have been contributing to the advancement of research within these areas. Examples include the analysis of attentional processes in psychology (Rehder and Hoffman, 2005) or smooth pursuit assessment in patients with degenerative cerebellar lesions (Moschner et al., 1999). One important area of application for eye tracking is vestibular research, where measurements of the vestibulo-ocular reflex (VOR) and nystagmus behavior are essential in the diagnostic pathway of balance disorders (Ben Slama et al., 2017). Beyond neuroscientific applications, eye-tracking was also utilized by autonomous driving industry for driver fatigue detection (Horng et al., 2004). Other than that, the trajectories and velocities of eye movements over a viewing task can serve as individual biometric signature for identification purpose (Bednarik et al., 2005; Liang et al., 2012). In consumer-behaviour research, eye-tracking has been used to study the dynamics and locations of consumers' attention deployment on promoted products in order to improve the design of advertisement (Lohse, 1997; Reutskaja et al., 2011). It is clear that pupil detection and tracking techniques build a fundamental block for eye movement analysis, enabling advancement in neuroscientific research, clinical assessment and real life applications.

Despite their importance, robust, replicable and accurate eye

[*] Corresponding author.
E-mail address: ahmad.ahmadi@med.uni-muenchen.de (S.-A. Ahmadi).

https://doi.org/10.1016/j.jneumeth.2019.05.016
Received 28 February 2019; Received in revised form 28 May 2019; Accepted 28 May 2019
Available online 06 June 2019

tracking and gaze estimation remain challenging under naturalistic low-light conditions. Most of the gaze estimation approaches, such as Pupil-Centre-Corneal-Reflection (PCCR) tracking (Guestrin and Eizenman, 2006) and geometric approaches based on eye shapes (Krafka et al., 2016; Chen and Ji, 2008; Yamazoe et al., 2008; Yang and Saniie, 2016; Ishikawa et al., 2004) depend on inferring gaze information from the pupil's location and shape in the image. However, the pupil is not always clearly visible to the camera. As summarized in (Schnipke and Todd, 2000), the pupil appearance can suffer from occlusion due to half-open eyelids or eyelashes, from reflection of external light sources on the cornea or glasses, from contact lenses or from low illumination, low contrast, camera defocusing or motion blur. All these artifacts pose challenges to pupil detection, and eye tracking algorithms which were not specifically designed with these artifacts in mind, may fail or give unreliable results under these circumstances.

In medical image analysis and computer vision, dramatic improvements in dealing with such artifacts have been achieved in recent years due to the introduction and rapid advancement of deep learning, specifically convolutional neural networks (CNN). An important distinction to hand-designed algorithms is that a CNN can achieve robust pupil segmentation, by automatically learning a sequence of image processing steps which are necessary to optimally compensate for all image artifacts which were encountered during training.

1.1. Related work

Conventional gaze estimation is often based on the Pupil-Centre-Corneal-Reflection (PCCR) method (Guestrin and Eizenman, 2006), which requires accurate localization of the pupil centre and glints, i.e. corneal reflections. Localization algorithms for the pupil and glints are often based on image processing heuristics such as adaptive intensity thresholding, followed by ray-based ellipse fitting (Li et al., 2005), morphological operators for contour detection (Fuhl et al., 2015a), circular filter matching (Fuhl et al., 2015b), Haar-like feature detection and clustering (Świrski et al., 2012), or radial symmetry detection (Kumar et al., 2009). It is important to note that most of these approaches assume the pupil to be the darkest region of the image (Fuhl et al., 2015a), which is susceptible to different illumination conditions and may require manual tuning of threshold parameters (Satriya et al., 2016; Kumar et al., 2009; Santini et al., 2017). Previous to our approach, several deep-learning based pupil detection approaches have been proposed to improve the robustness to artifacts by learning hierarchical image patterns with CNNs. PupilNet (Fuhl et al., 2016) locates the pupil centre position with two cascaded CNNs for coarse-to-fine localization. In Chinsatit and Saitoh (2017), another CNN cascade first classifies the eye states of "open", "half-open" and "closed", before applying specialized CNNs to estimate the pupil centre coordinates, based on the eye state. However, current CNN approaches output only the pupil centre coordinates, which alone are not enough to determine the gaze direction without calibration or additional information from corneal reflection. Some studies focus on end-to-end training of a CNN, directly mapping the input space of eye images to the gaze results (Krafka et al., 2016; Naqvi et al., 2018), but they are confined to applications in specific environment, such as estimating gaze regions on the car windscreen (Naqvi et al., 2018) or mobile device monitors (Krafka et al., 2016), which are not suitable for clinical measurement of angular eye movement.

1.2. Contribution

In this work, we propose DeepVOG, a framework for video-oculography based on deep neural networks. As its core component, we propose to use a fully convolutional neural network (FCNN) for segmentation of the complete pupil instead of only localizing its center. The segmentation output simultaneously enables us to perform pupil center localization, elliptical contour estimation and blink detection, all with a single network, and with an assigned confidence value. We train our network on a dataset of approximately four thousand eye images acquired during video-oculography experiments at our institute, and hand-labeled by human raters who outlined the elliptical pupil contour. Though trained on data from our institute, we demonstrate that the FCNN can generalize well to pupil segmentation in multiple datasets from other camera hardware and pupil tracking setups. On consumer-level hardware, we demonstrate our approach to infer pupil segmentations at a rate of more than 100 Hz. Beyond pupil segmentation, we re-implement a published and validated method for horizontal and vertical gaze estimation and integrate it as an optional module into our framework (Świrski and Dodgson, 2013). We show that the integration of gaze estimation is seamless, given that our FCNN approach provides elliptical pupil outline estimates. We further show that by considering ellipse confidence measures from our FCNN output, the accuracy of the gaze estimation algorithm can be increased. Our implementation is fully Python-based and provided open-source for free usage in academic and commercial solutions. Our code, pre-trained pupil segmentation network and documentation can be found under: www.github.com/pydsgz.

2. Materials and methods

2.1. Datasets

For this study, we acquired three datasets at the German Center for Vertigo and Balance Disorders, two for training validation of the pupil segmentation network and one for validation of the gaze estimation. Training sequences were acquired in a challenging environment, i.e. inside a MRI scanner, during a neuroscientific experiment setup involving VOG and simultaneous functional MRI (fMRI) assessment. Both training datasets were collected on a cohort of healthy young adults. Training dataset A was acquired from 35 subjects (16 male, 19 female, age 28.1 ± 4.0 years) in a fully darkened MRI scanner room. Training dataset B was acquired from 27 subjects (8 male, 19 female, age 25.5 ± 3.7 years) in a scanner room with normal illumination. For MRI-compatible recording, we utilized a commercial VOG system (NNL EyeTracking Camera, NordicNeuroLab AS, Bergen, Norway), yielding VOG videos at a 320 × 240 pixel resolution and a framerate of 60 Hz. Both datasets A and B contained video sections with challenging pupil appearance, leading to high dropout rates of eye tracking and mis-localizations of the pupil center with two commercial eye tracking software solutions (ViewPoint EyeTracker, Arrington Research, Arizona, USA; EyeSeeCam, EyeSeeTec GmbH, F+1/4rstenfeldbruck, Germany). Typical example images from both datasets are shown in Fig. 1. Pupil detection failures occur e.g. due to highly dilated pupils in dark environments, dark circular borders from ocular mount gaps, heterogeneous illumination and pupil occlusion from ocular borders, eyelids and eyelashes.

For network training, we randomly sampled 3946 eye image frames from both datasets (training set A: 1539 frames, training set B: 2416 frames). Five human raters segmented the pupil in all images (one rater per image), using a custom labeling tool implemented in Python, by manual placement of at least five points on the visible part of the pupil boundary, followed by least-squares fitting of an ellipse to the boundary points.

Validation of the pupil segmentation performance was done on the test images from datasets A (959 frames) and B (1043 frames), i.e. images that were not seen by the network during training. In order to test the generalization capability of our network to entirely novel eye appearances, we tested its pupil detection performance on previously unseen third-party data, including Delhi Iris Database (Delhi) (Kumar and Passi, 2010), Labelled Pupil in the Wild (LPW) (Tonsen et al., 2015) and Multimedia-University Iris (MMU Iris)

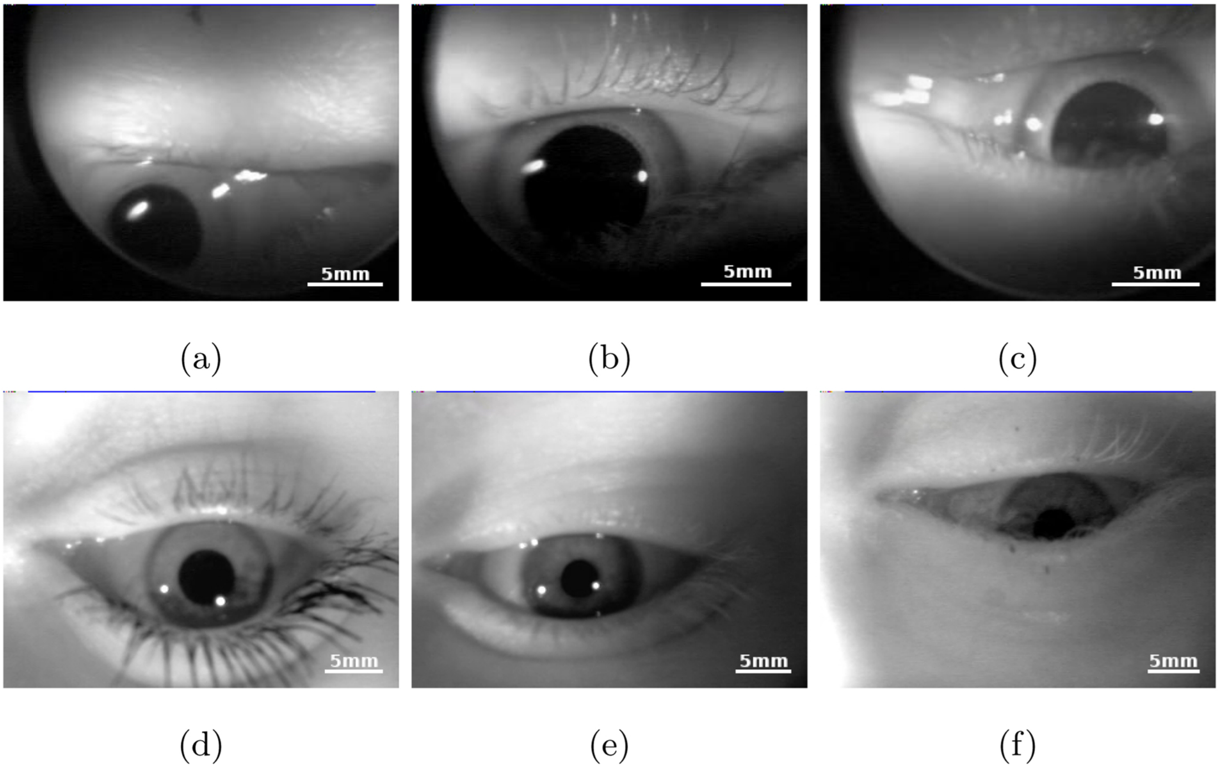


Fig. 1. Examples of input eye images in training dataset A (a, b, c) and dataset B (d, e, f). Observations and artifacts include: (a, b) pupil merging with dark background from ocular edge, (b) motion blur, (c) eyelid occlusion, (d) clear pupils with glint reflection, (e) inhomogeneous illumination, (f) half occluded pupil with dark iris. All images have a resolution of 320 × 240 in pixel.

Table 1
(a) Difference between predicted and manually labelled pupil regions, measured in terms of median values (inter-quartile range) of Dice's coefficient, Euclidean distance between pupil centers, and Hausdorff distance between pupil contours. Results on different datasets (blink images excluded for datasets A and B) are shown. Notably, the DeepVOG network was only trained on datasets A and B, but is able to generalize with high pupil segmentation accuracy to the five other datasets. (b) Blink detection rate analyzed on dataset A and B. Pupil ellipses with confidence < 0.96 were classified as blink.

(a)

Datasets	Dice's coefficient	Euclidean distance (px)	Hausdorff distance (px)
Dataset A &B (1892)	0.966 (0.948–0.976)	1.0 (0.6–1.6)	2.8 (2.0–4.0)
Dataset C (323)	0.97 (0.958–0.978)	0.9 (0.6–1.3)	2.8 (2.0–3.2)
Delhi (763)	0.978 (0.971–0.983)	0.8 (0.5–1.2)	2.8 (2.0–3.0)
LPW (466)	0.938 (0.914–0.957)	0.9 (0.6–1.4)	3.2 (2.2–4.0)
MMU Iris (167)	0.958 (0.947–0.968)	1.0 (0.6–1.4)	2.2 (2.1–3.0)
Blender (361)	0.965 (0.901–0.982)	1.8 (1.3–2.4)	3.6 (2.0–8.2)

(b)

Datasets	Accuracy (%)	Sensitivity (%)	Specificity (%)
Dataset A &B (2002)	0.93	0.94	0.93

(Multimedia-University, 2019). Furthermore, we acquired one more set of video sequence data (Dataset C) at our institute, for the quantitative validation of gaze estimation (see below). LPW and MMU Iris were provided with labels of pupil centre coordinates in their images, which enables quantitative analysis on our network's performance of localizing pupil centers. To quantitatively validate the segmentation accuracy on the other unlabelled datasets, we labeled pupils in a small subset of each dataset (cf. Table 1a). As a preprocessing step, all images or video frames with different resolution were resized to 320 × 240 in pixel for the analysis.

Dataset C, the validation set for gaze estimation, was acquired from 9 healthy subjects (5 male, 4 female, age 33.8 ± 5.9 years), in the neuro-ophthalmological examination laboratory of the German Center for Vertigo and Balance Disorders. The setup included a commercial system for clinical video oculography (EyeSeeTec GmbH,[1] F + 1/4rstenfeldbruck, Germany; video resolution 320 × 240 pixels, 120 Hz framerate). To calibrate each subject's gaze, a gold standard calibration was performed using a projector-assisted five-point calibration paradigm (8.5° horizontal and vertical gaze extent). For 3D eye model fitting using our framework, each subject additionally performed two projector-free, unassisted calibrations with three trials each (for details, see Section 2.5). To validate the accuracy of gaze estimates, each subject underwent an oculomotoric examination comprising four clinical tests: saccade test, fixation nystagmus, smooth pursuit and optokinetic nystagmus. We then compare gaze estimates from our framework to

[1] www.eyeseecam.com.

estimates from the clinical gold standard calibration and system.

2.2. Pupil segmentation network

The general structure of deep CNNs features stacked convolutional image filters and other signal processing layers (e.g. for image resampling), arranged in a sequence of processing layers (LeCun et al., 2015). These filters are trained towards optimally fulfilling a defined goal given an input image, in this study the robust segmentation of pupils, despite challenging appearance. Initially, i.e. before training, these filters extract meaningless information from images and fail to achieve the segmentation task. During training, the network is repeatedly presented with different pupil images and corresponding human pupil segmentations as a gold standard, including highly challenging examples. By means of back-propagation of the residual segmentation error through the network (LeCun et al., 2015), the filters in the network weights gradually get adapted to compensate for artifacts and iteratively make better and better estimates for the optimal pupil area.

2.3. CNN architecture and pupil segmentation

Numerous architectures for medical image segmentation have been proposed to date, and surveys like (Litjens et al., 2017) provide a good entrypoint into this fast growing field. A well-established architecture is the U-Net (Ronneberger et al., 2015), which has shown to be adaptable to many segmentation problems in various medical imaging modalities. For this work, we adapt a basic U-Net architecture for pupil segmentation in 2D greyscale images. A U-Net consists of multiple layers of feature extraction, which are arranged in a down-sampling path on the left side and an up-sampling path on the right side. Horizontal skip connections relay high-resolution image features from the down-sampling path into the up-sampling path, in order to preserve high-frequency image features and achieve a sharp segmentation output.

Our architecture is depicted in Fig. 2. Compared to the original U-Net, we previously proposed several architectural changes as V-Net (Milletari et al., 2016), which we partly adopt in this work as well. At each stage of the up- and down-sampling path, we utilize a convolutional layer with 10 × 10 filters which outputs feature maps with the same size of the input by appropriate padding. The down-sampling path reduces the size of the feature maps and increases the size of receptive fields of convolutional filters at each stage, such that more complex features in a larger context can be extracted. Compared to (Ronneberger et al., 2015), which performs down- and up-sampling through pooling operations, we utilize strided convolution (2 × 2 filter kernel size, stride 2) for downsampling and transposed convolution (Noh et al., 2015) for upsampling, which is more memory efficient (Springenberg et al., 2014; Milletari et al., 2016) and able to learn optimal down-/up-sampling filters.

The final output layer has two output maps for pupil and background, with the same size as the input layer (320 × 240 pixels). We employ a softmax layer (Litjens et al., 2017) to perform smooth maximum activation across regions. The prediction yields a probabilistic output, to which we can fit an elliptical contour representing the pupil center and eccentricity of the boundary. To determine the contour points for ellipse fitting, we incorporate a simple post-processing on the network's probabilistic prediction of the pupil foreground: the prediction posterior is thresholded at a probability of $p > 0.5$, denoised through morphological closing (Soille, 2003), and the largest connected component is extracted to reject small false positive regions (cf. Fig. 3c). For both pupil area and center, a detection confidence value can be determined by computing the average prediction confidence within the detected pupil area. Finally, a blink can be detected if this confidence falls under a pre-defined threshold.

The pupil fitting procedure, including probabilistic network output, post-processing and ellipse fitting, is depicted in Fig. 3. Importantly, compared to previous approaches (Fuhl et al., 2016; Chinsatit and Saitoh, 2017; Krafka et al., 2016; Naqvi et al., 2018), our fully convolutional approach does not require a cascade of several CNNs to achieve the pupil detection, neither are we restricted to pure center localization. Pupil segmentation, center estimation and blink detection can be performed with a single FCNN, and outputs are assigned with confidence values which can be further utilized for gaze estimation and VOG evaluation.

2.4. Augmentation

To make the network more robust to expectable variations in camera pose and eye appearance, we artificially enhance the training dataset through random augmentation of image-segmentation pairs. During training, we apply random rotation within the range of ± 40°. Images were further randomly translated in the range of ± 20% of height and width. We also applied random zooming by a factor of ± 0.2 and random flipping in horizontal and vertical direction.

2.5. Gaze estimation

As derived by Świrski and Dodgson (2013), it is possible to

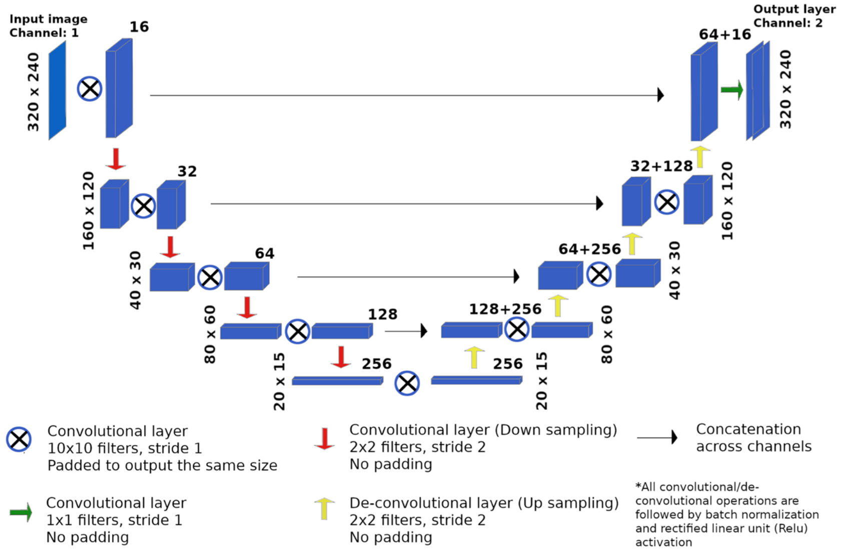


Fig. 2. Architecture of the fully convolutional neural network (FCNN) for pupil detection, inspired by U-Net (Ronneberger et al., 2015) and V-Net (Milletari et al., 2016). The network takes a single eye image as input, and produces an image-sized output with pixel-wise estimates for the pupil area. The network consists of a sequence of convolutional image filters which robustly extract features to distinguish the pupil from the background. The numbers on the top-right corner of the feature maps represent the number of filter channels. Several down-sampling operations allow the network to localize and segment the pupil at multiple image resolutions. In the up-sampling path, the low-resolution image representation gets inflated back to the original size, while reconstructing the location and shape of the pupil. Horizontal skip connections preserve high-frequency image information and sharp edges throughout the down-/up-sampling operations. During training, all image filters are tuned towards optimally compensating for image artifacts encountered in the training set.

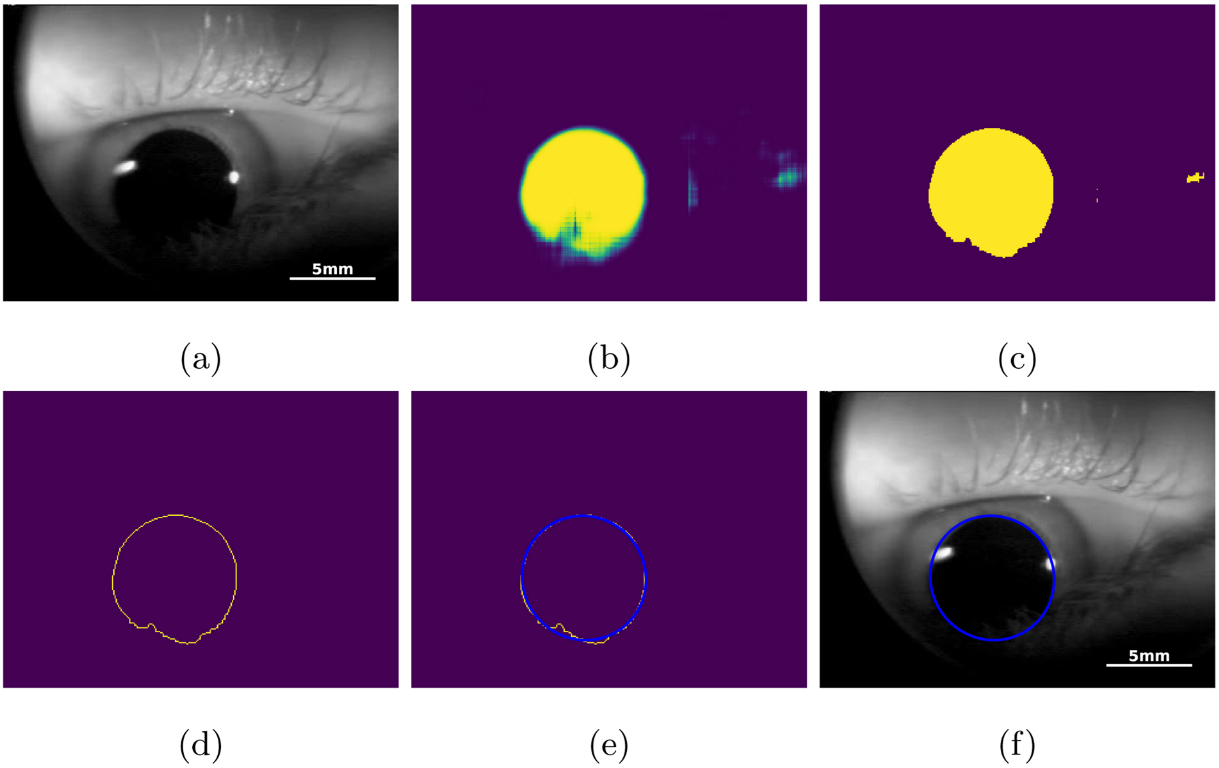


Fig. 3. Procedure of fitting an ellipse to a probability map of segmented pupil area. (a) Original input image. (b) Probabilistic pupil prediction output from the network. (c) Binarization of the output with threshold 0.5 and morphological closing for filling in small gaps. (d) Isolation of largest connected area, and extraction of perimeter points of the area. (e) Fitted ellipse (blue line) from perimeter points. (f) Fitted pupil contour on the input image. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

reconstruct a 3D model of the eyeball and perform gaze estimation, purely based on a set of estimated 2D pupil ellipses from a video sequence, and without further fixation-based or projector-assisted calibration. Using the fitted 3D model, estimates of horizontal and vertical gaze angles can then be derived in each frame of a newly recorded VOG video, using the location and eccentricity of pupil ellipses. The reconstruction of the 3D eye model is based on the un-projection algorithm by Safaee-Rad et al. (1992). Its extension to gaze estimation is described in detail in Świrski and Dodgson (2013). Notably, it is assumed that the eyeball is of spherical shape and the pupil is a perfect circular disk (with varying radius) on the sphere surface, whose projections onto the camera's image plane form the 2D ellipse shapes. For a detailed derivation of the eye model fitting theory and algorithm, we refer the reader to (Świrski and Dodgson, 2013). Likewise, our Python-based re-implementation of the method with documentation can be found in our public code repository. Here, we want to emphasize that this "self-calibration" method can be very well complemented by our FCNN model, since it directly outputs a full segmentation of the pupil and elliptical pupil outlines. In particular, we can make the 3D eye model fitting more robust and gaze estimates more accurate by incorporating the confidence estimates of our FCNN into the fitting procedure. To this end, we extend the original formulae (6) and (9) in (Świrski and Dodgson, 2013) by incorporating a confidence factor α_i for each image frame (and elliptical pupil estimate) that is considered during 3D model fitting and gaze estimation. Our confidence-weighted fitting formulae for the 3D eyeball center $\boldsymbol{C}$ and radius R can be denoted as:

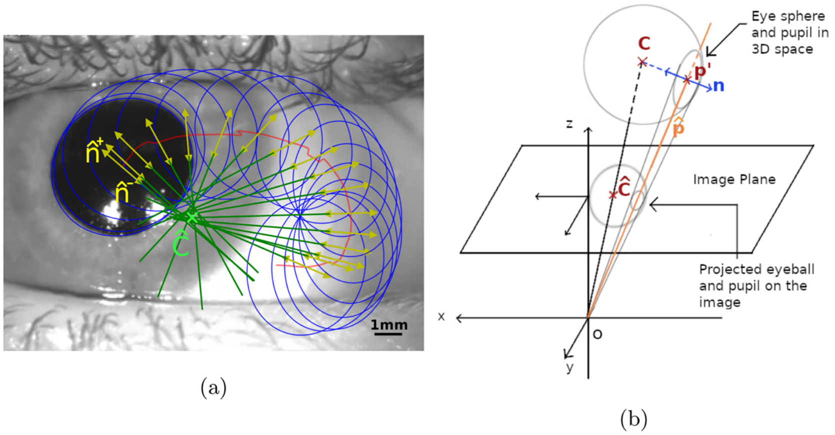
$$\alpha_i = \begin{cases} 0, & \text{confidence}_i < \theta \\ 1, & \text{confidence}_i > \theta' \end{cases} \tag{1}$$

$$\hat{\boldsymbol{C}} = \left(\sum_i \alpha_i (\boldsymbol{I} - \hat{\boldsymbol{n}}_i \hat{\boldsymbol{n}}_i^T)\right)^{-1} \left(\sum_i \alpha_i (\boldsymbol{I} - \hat{\boldsymbol{n}}_i \hat{\boldsymbol{n}}_i^T) \hat{\boldsymbol{p}}_i\right) \tag{2}$$

$$R = \frac{1}{\sum_i \alpha_i} \left(\sum_i \alpha_i \|\boldsymbol{p}'_i - \boldsymbol{C}\|\right) \tag{3}$$

Here, following the notation of (Świrski and Dodgson, 2013), $\hat{\boldsymbol{C}}$ denotes the projection of the eyeball center to the 2D image plane, in which all projected pupil normals $\hat{\boldsymbol{n}}_i$ have to intersect, given projected pupil centres $\hat{\boldsymbol{p}}_i$ estimated by the FCNN. The symbol $\boldsymbol{p}'_i$ is the intersection between $\hat{\boldsymbol{p}}_i$ and the parameterized line ($\boldsymbol{C} + \boldsymbol{n}t$), as illustrated in Fig. 4b.

To fit the 3D eyeball model, the camera needs to be fixed with respect to the participant's head and eye, while the participant performs some form of eye motion that yields sufficiently many image frames with pupils of elliptical appearance. This can be achieved with different unassisted calibration paradigms. In this study, we fit and compare the 3D eyeball model based on three such calibration paradigms. The motivation is to assess the robustness of gaze estimation against different calibration methods and to discuss the requirements of a correct model fitting protocol which maximizes the estimation performance. The first paradigm is "Free-looking". Here, participants need to keep their head fixed, while freely and smoothly looking all around the periphery of their visual field, thus yielding highly elliptical pupil appearances. The second is inspired by "CalibMe", a recently proposed unassisted calibration approach for eye tracking, proposed by Santini et al. (2017). Here, participants select a stationary marker in their visual field and move their heads around it in a circular motion, while keeping her gaze fixed at the selected marker. The third one is "narrow-ranged", a projector-assisted calibration approach that is commonly utilized in clinical eye-tracking systems and experiments. At our clinical center, participants focus their gaze on five sequentially presented fixation points



(a)

(b)

Fig. 4. Gaze estimation procedure, as adopted from Świrski and Dodgson (2013). Given a set of FCNN-based pupil segmentations in the camera image frame, the method reconstructs a 3D eyeball geometry that optimally explains the observed elliptic pupil projections. (a) Estimation of projected eye sphere centre $\hat{\boldsymbol{C}}$. Gaze normals (yellow arrows, $\hat{\boldsymbol{n}}^{+}$ pointing outward and $\hat{\boldsymbol{n}}^{-}$ pointing towards $\hat{\boldsymbol{C}}$) intersect in $\hat{\boldsymbol{C}}$, its location is approximated in a least-squares fashion. (b) Estimation of eye sphere radius R. The orange line denotes the possible candidates of pupil centres ($\hat{\boldsymbol{p}}$) after unprojection, which intersects with the parameterized line ($\boldsymbol{C} + \boldsymbol{n}t$) at $\boldsymbol{p}'$. The difference vector between $\boldsymbol{p}'$ and $\boldsymbol{C}$ gives information about eye radius R, and after the fitting stage, gaze direction $\boldsymbol{n}'$ (not shown here). For details, see (Świrski and Dodgson, 2013) and our documented open-source implementation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

located at the screen's centre in both horizontal and vertical direction at ± 8.5°. We investigate whether unassisted calibration is feasible under these circumstances, which is particularly relevant towards a retrospective evaluation of datasets in which large-range gaze angles were not yet considered during the calibration protocol.

3. Results

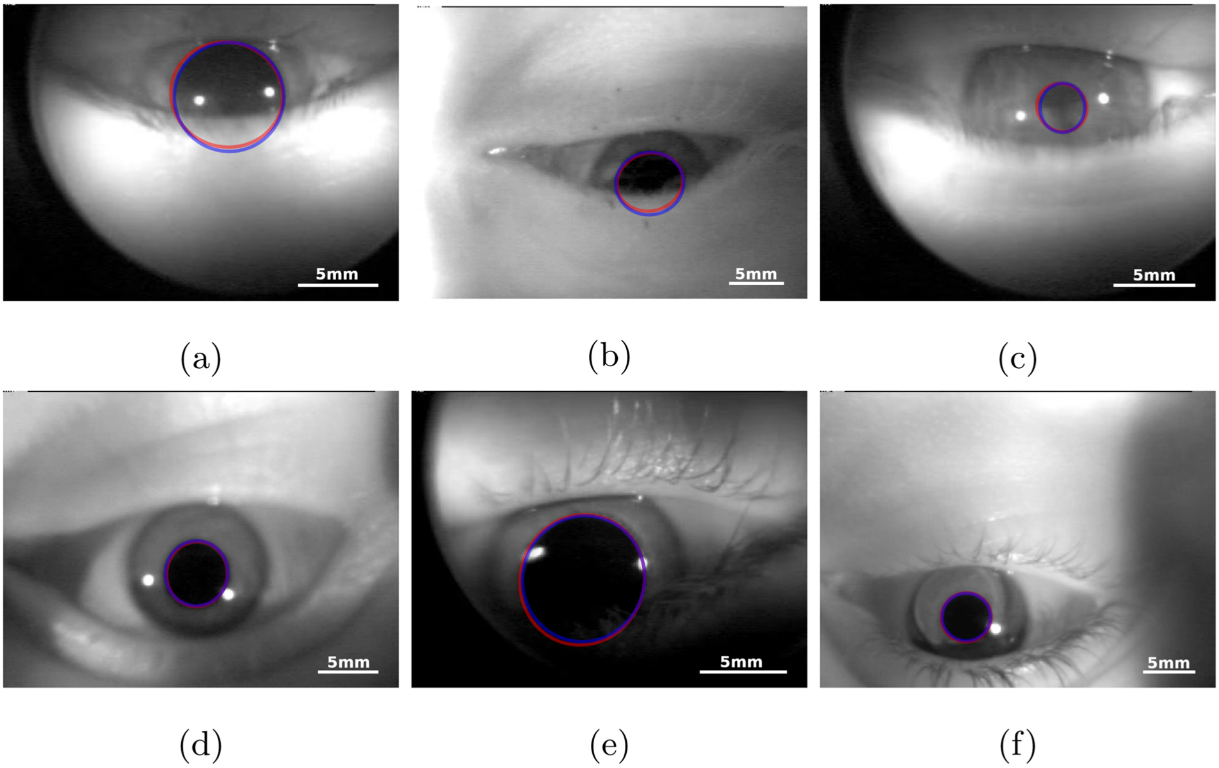
3.1. CNN-based pupil detection results

3.1.1. Robustness

Our network successfully segmented the pupil areas in images under difficult conditions such as occlusion from eyelid or eyelashes, specular reflections, non-homogeneous illumination and naturalistic low-light, leading to highly dilated pupils. Several examples are illustrated in Fig. 5. It is also robust to glint reflection, motion blur, camera de-focusing, and even to the appearance of unexpected dark edges from off-center ocular placement (see Fig. 5f), which is not a consistent feature throughout the dataset.

3.1.2. Network validation

We tested the performance of DeepVOG on the testing images of datasets A and B, which were similar to the training images but had not been seen by the network during the training process. The accuracy of pupil center detection is measured by the Euclidean distance (unit: [pixel]) between the predicted and manually labelled pupil centers. The accuracy of pupil area segmentation is computed by the Dice overlap coefficient and Hausdorff distance. The Dice coefficient computes the overlap between predicted and manually labelled areas of the pupil (range [0...1], with 1.0 indicating perfect overlap), while the Hausdorff



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 5. Pupil detection results under different difficult or noisy conditions. Red lines are pupil ellipses manually labeled by human raters. Blue lines are pupil ellipses fitted to the FCNN segmentations in DeepVOG. The resolution of images are 320 × 240 in pixel. (a, b) Eyelid occlusion and dilated pupils. (c) Eyelashes occlusion. (d) Camera de-focusing. (e) Motion blur and dark edge. (f) Contact lenses and non-homogeneous illumination. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

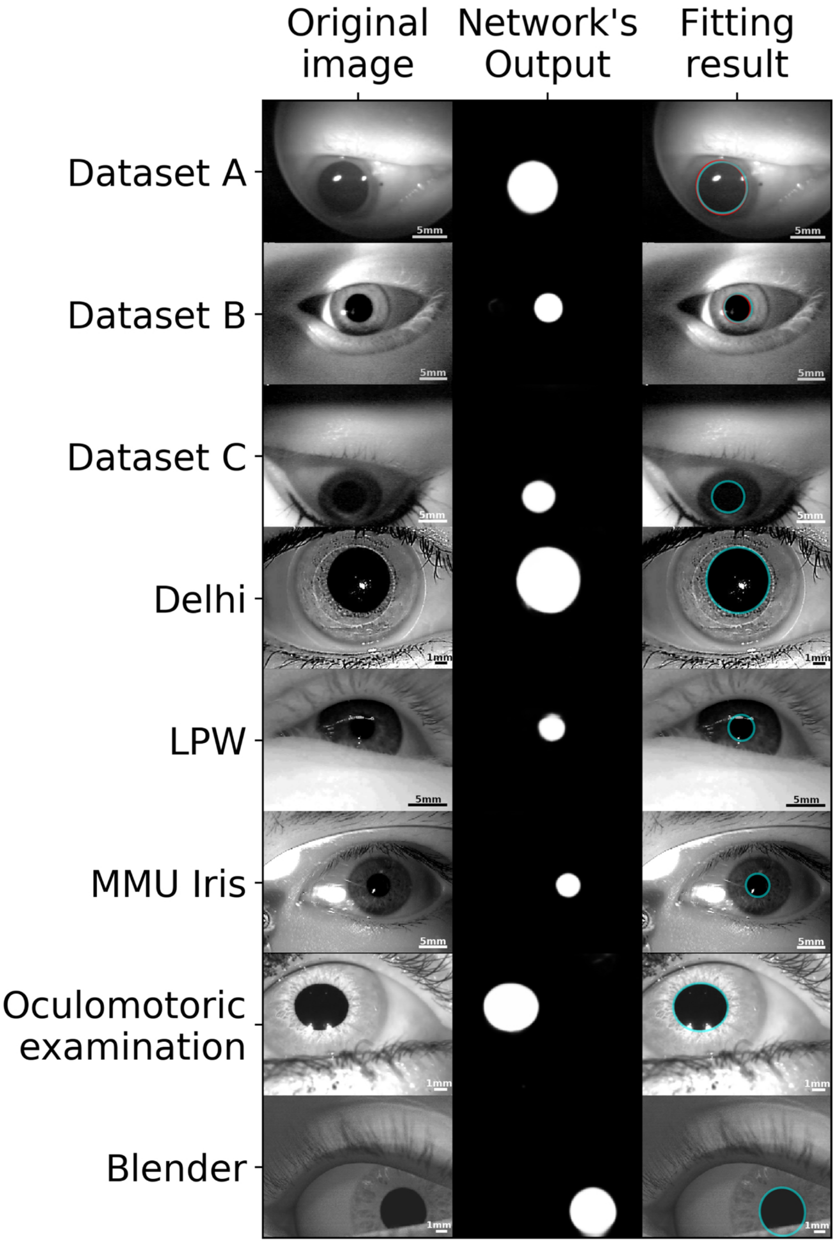


Fig. 6. Generalization capability of the FCNN pupil segmentation in DeepVOG. Examples of eye images (320 × 240 in pixel) from all datasets in this study. Left column: the original image used as an input to the network. Middle: the output of the network as a probability map of the pupil area. Right: the ellipse fitting result based on the network's output. Light blue contours denote the fitted ellipse by our model and red contours denote the manually fitted ellipse (shown only in datasets A and B). It is noteworthy that even though the FCNN was only trained on images from datasets A and B, the segmentation maps in all other datasets represent the pupil accurately and with a high confidence. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

distance (unit: [pixel]) measures the maximum pupil contour distance between prediction and manual labeling. Overall, the FCNN in DeepVOG achieved a high accuracy in pupil tracking on the 2002 unseen images of datasets A and B (see Table 1a). The median Euclidean distance was 1.0 pixel (IQR: 0.6–1.6 pixels), which is visually difficult to discern in a 320 × 240 image. The Dice coefficient (median = 0.966) and Hausdorff distance (median = 2.8 pixels) further suggest that not only the pupil centre, but the entire segmented pupil areas were highly similar to the manually labelled ground truths.

3.1.3. Generalizability to other datasets

Test images from datasets A and B were similar to training images. To test whether the network can generalize to images from other completely novel datasets, we utilized a wide variety of datasets including dataset C, Delhi, LPW, MMU Iris and artificially rendered eye images using Blender. Fig. 6 shows an example image and segmentation result for each of these datasets. Although images from Delhi, MMU Iris and LPW differ significantly with the training datasets A and B in terms of their illuminance, contrast, camera angles, shadows and reflection, our network was still able to segment a perfect pupil shape. Particularly in Fig. 6, the segmentation results are still robust against low pupil-boundary contrast in dataset C and LPW, contact lenses in Delhi and glasses in MMU Iris. On the other hand, there are a few failure cases in novel datasets if the surrounding of the pupil is shadowed and comparably dark (cf. Fig. 7). Regarding quantitative measures in Table 1a, the performance of pupil centre detection on the third-party datasets LPW (median of Euclidean distance = 0.9 pixels) and MMU Iris (median of Euclidean distance = 1.0 pixel) is on par with results on datasets A and B, which our network was trained on. The accuracy of pupil segmentation remains high on Dataset C and Delhi, as indicated by the decent quantitative results of Dice's coefficient, Euclidean distance and Hausdorff distance. This demonstrates a robust detection of pupil centres in unseen datasets, without much decrease of accuracy.

Among the datasets above, we tested the performance of DeepVOG on artificially rendered eye images which we generated with a VOG simulation (Świrski and Dodgson, 2014) in the 3D modeling software “Blender”, as well as on oculomotoric examination data from our clinical center (for details, see Sections 3.2 and 3.3). Here, a remarkable

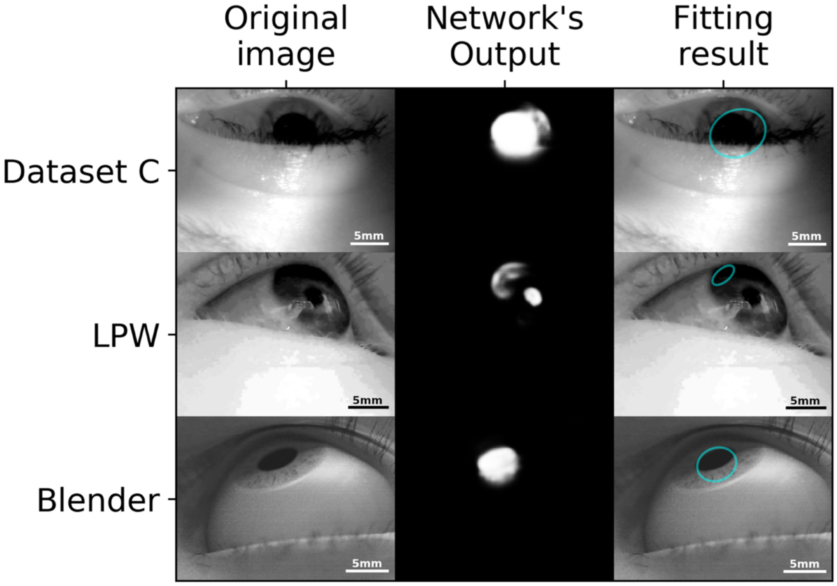


Fig. 7. Examples of failed pupil segmentation results (320 × 240 in pixel). The light blue contour denotes the fitted ellipse by our model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

accuracy can also be observed for both datasets, which adds to the impression that DeepVOG can generalize to new data distributions to a great extent.

3.1.4. Blink detection

To realize blink detection, we utilize the fact that when the FCNN is uncertain about the classification, it will yield a lower confidence estimate, which we calculate as the average confidence of the output pixel values within the fitted ellipse area. Here we analyze how reliable this confidence indicator is in identifying blinks. Table 1b shows that a straightforward confidence thresholding is able to classify blinks with high accuracy, sensitivity and specificity on dataset A and B where image frames with during blinks ($n = 110$) had also been manually labelled.

3.1.5. Inference speed

The FCNN forward-pass and inference speed is important with respect to real-time measurement of eye positions, as well as efficient offline data analysis. We tested the forward inference speed of DeepVOG network on a consumer-level workstation with a Nvidia GTX 1080 Ti graphics processing unit (GPU). The results show that DeepVOG runs at a 30 Hz detection rate if segmentations are inferred frame-by-frame (corresponding to 17 ms latency). If segmentations are computed in batches of 32 frames (i.e. 533 ms latency), inference can reach a framerate of more than 130 Hz (see Fig. 8), demonstrating the potential of fast offline data inference, and real-time pupil detection above 100 Hz and with latencies well below one second.

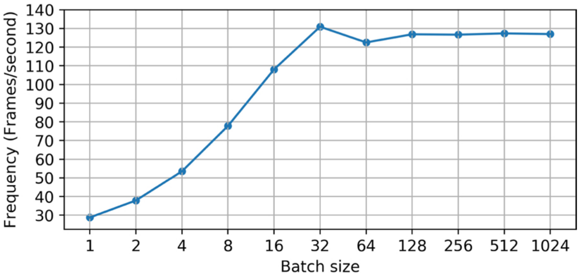


Fig. 8. Prediction speed: Forward inference frequency (in frames per second) of the FCNN in DeepVOG, with various input batch sizes at test time. At a batch size and latency of 32 frames, the FCNN can segment pupils at a rate of more than 130 Hz.

3.2. Gaze estimation on artificial data

Using the 3D modeling software “Blender”, we generated artificial gold standard eye images and gaze directions using a dedicated VOG simulation (Świrski and Dodgson, 2014). DeepVOG performed the pupil centre detection on Blender's dataset with high accuracy, as indicated by small Euclidean distances between the predicted centres and simulation ground-truths (Table 2). Further, our model can estimate gaze directions with very small angular errors at a median of around 0.5°. We further investigated the effect of using only confident pupil segmentations predicted by the network for 3D eye model fitting. If we filter out gaze estimation results based on low-confidence network's outputs, the median of angular errors is reduced by around 0.13° and its upper-quartile by 0.2°–0.4°. Fig. 9 visualizes the comparison between the predicted and ground-truth pupil centre coordinates. The visualization demonstrates a highly accurate performance on Blender images for both pupil centre detection and gaze estimation, except at the top left corner where the pupil shape becomes highly elliptical or barely visible from the camera angle (example shown in Fig. 7). With confidence thresholding, these relatively inaccurate predictions can be identified and optionally omitted (cf. Fig. 9c and d).

3.3. Gaze estimation on clinical data

3.3.1. Accuracy and precision

We evaluated the gaze estimation accuracy of the DeepVOG framework, given three calibration paradigms (Free-looking, CalibMe and narrow-ranged calibration), using the oculomotoric examination videos

Table 2
Median (inter-quartile range) of absolute angular errors for horizontal and vertical eye movements, as well as Euclidean distance between predictions and simulation ground-truth. Confident prediction includes only data with pupil segmentation confidence > 0.96 while normal prediction includes all data.

Prediction type	Absolute horizontal angular errors (°)	Absolute vertical angular errors (°)	Euclidean distance (px)
Normal	0.61 (0.24–1.29)	0.51 (0.26–1.01)	1.78 (1.36–2.42)
Confident	0.45 (0.20–1.02)	0.38 (0.22–0.66)	1.56 (1.22–1.99)

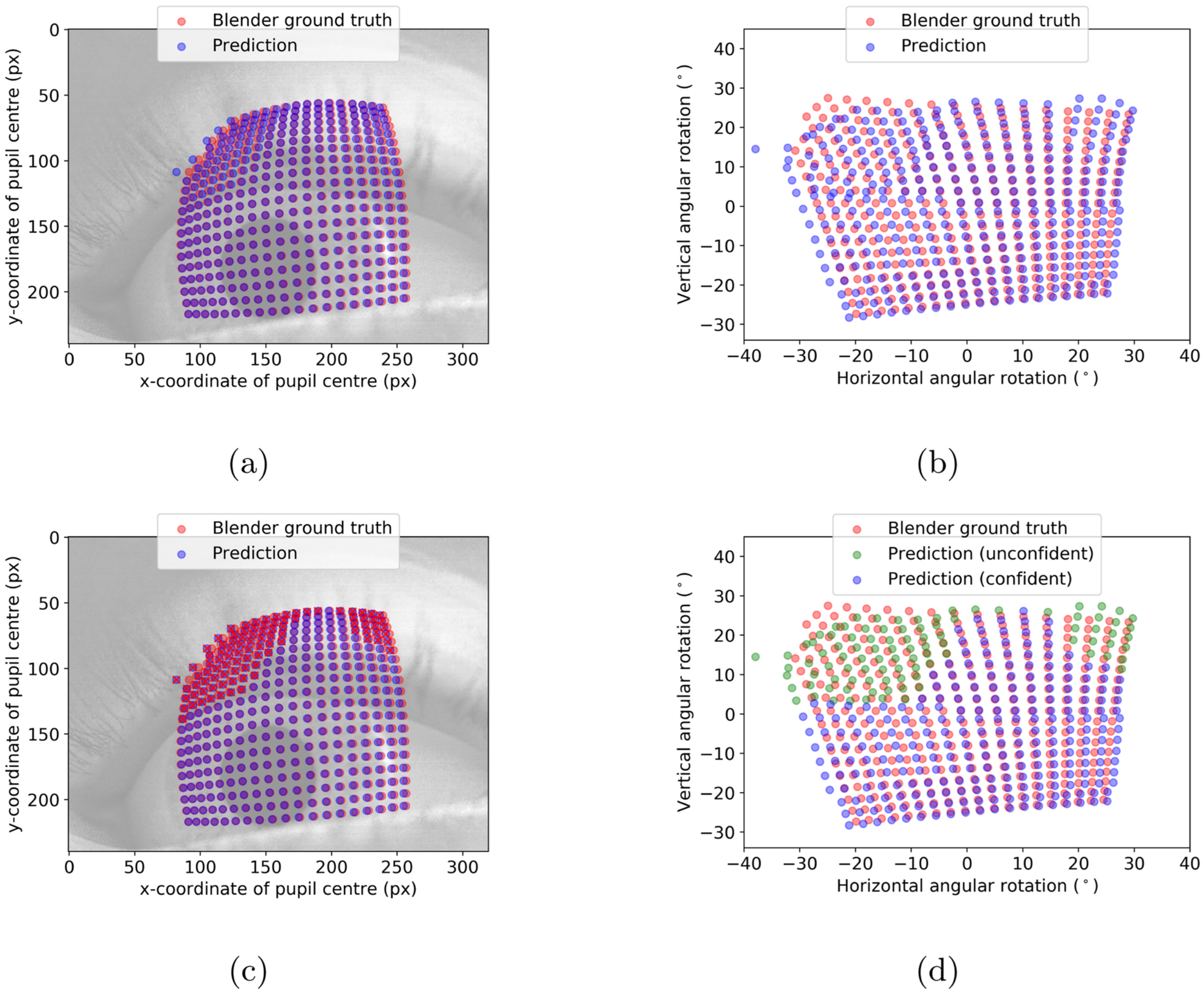


Fig. 9. Results of DeepVOG gaze estimations compared to gold-standard artificial eye images rendered with a VOG simulation (Świrski and Dodgson, 2014) in the 3D modeling software Blender. (a) Predicted pupil centre positions (blue dots) by the FCNN on images rendered by Blender, overlaid on ground-truth centre positions (red dots). (b) Predicted gaze directions by DeepVOG (blue dots) on rendered images, overlaid on ground-truth gaze directions (red dots). (c) Same figure as a, but with unconfident pupil detections crossed out (confidence < 0.96). (d) Same figure as b, but with the unconfident gaze estimation points shown as green dots. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of 9 participants (Dataset C, cf. Section 2.1). Each examination was repeated for 3 trials to assess the repeatability of each calibration paradigm. Our gaze estimation results were validated against the results of the clinical gold standard system (EyeSeeTec). An example of gaze estimation results for the oculomotor examination video sequence is shown in Fig. 10. The median angular errors range between 0.3° and 0.6° (Table 3a), which demonstrates an accurate detection of eye angular movement. This also implies that the FCNN in DeepVOG is able to generalize well to this novel dataset of oculomotoric examination videos, given that the correctness of gaze estimation is substantially determined by the success of pupil area segmentation.

3.3.2. *Robustness to self-calibration paradigm*

All three self-calibration paradigms in this study enabled sub-degree performance of angular eye movement tracking, demonstrating the robustness to all three calibration paradigms. Nonetheless, there are differences. The Free-looking and CalibMe paradigms produced less angular errors than the narrow-ranged paradigm, as indicated in the accumulative distributions of angular errors (Fig. 11a and b). Importantly, in Świrski's and Dodgson's model (Świrski and Dodgson, 2013), the 3D eyeball parameters are more accurately fitted if a large range of pupil motion and elliptical appearances is covered. Here, the free-looking and CalibMe paradigms provide observations from a wider distribution of gaze angles, while the narrow-ranged paradigm gives only small gaze angles at five distinct directions, which is not diverse enough for estimating an accurate 3D eyeball model. Consequently, this leads to larger angular deviations.

3.3.3. *Repeatability*

To assess the repeatability of 3D eye model fitting, we computed the intra-class correlation coefficient (ICC) across the measurements of three trials of each calibration paradigm (Table 2b), using ICC's two-way mixed effect, single-measures and absolute agreement model, i.e. ICC(A,1). Results revealed that the gaze predictions from the three trials achieved an excellent (ICC > 0.9, (Koo and Li, 2016)) reliability on their absolute agreement. The high reliability of measurement implies that DeepVOG can produce consistent and repeatable results across several experiment trials. Again, the Free-looking and CalibMe calibrations lead to similar ICC values, which are higher than if only Narrow-ranged gaze angles are used during 3D eye model fitting.

4. Discussion

In this manuscript, we describe DeepVOG, a novel eye-tracking framework that uses fully convolutional neural networks to perform pupil segmentation. It outputs pupil centers, pupil areas, elliptical

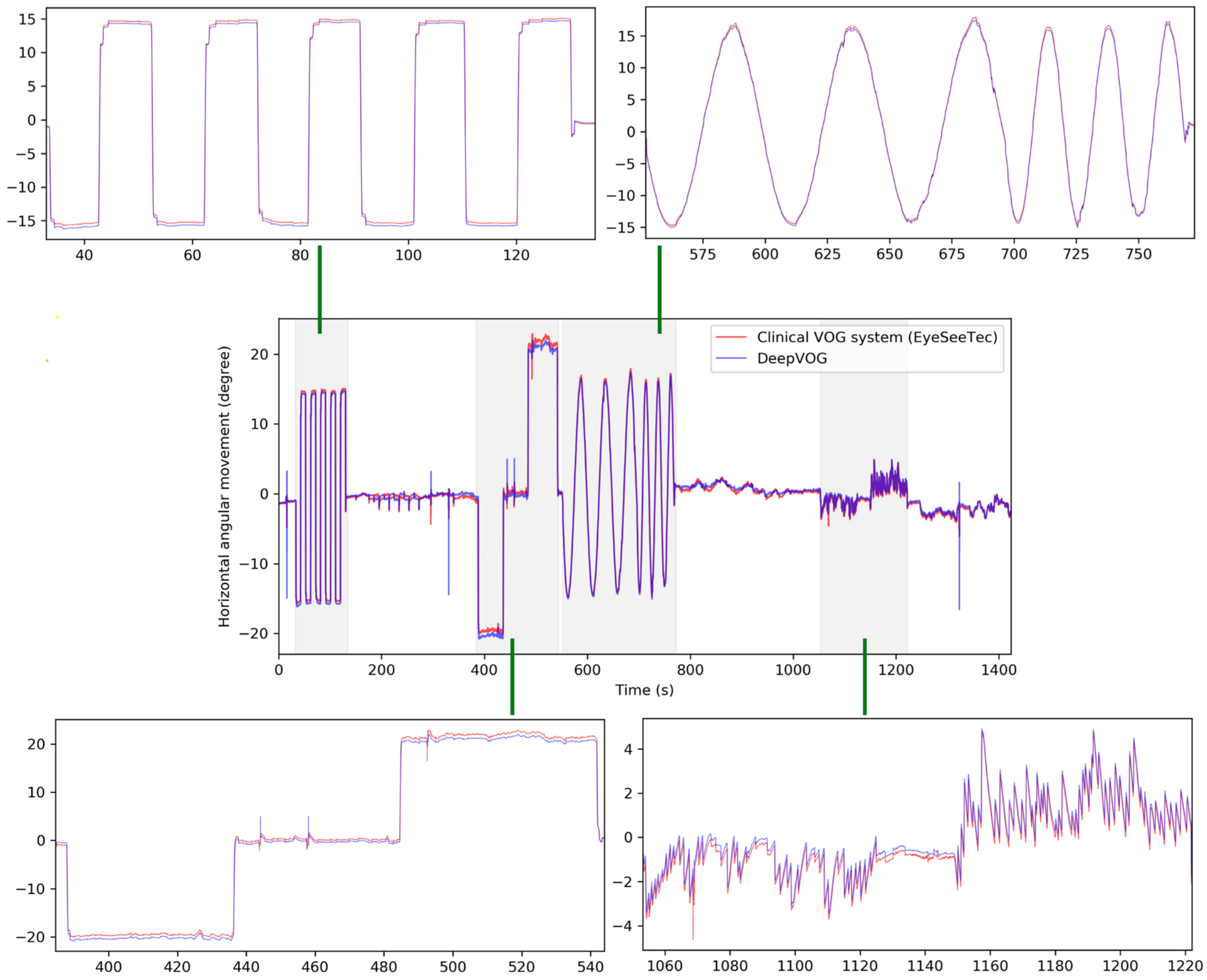


Fig. 10. Example gaze estimation results on an oculomotor examination video sequence (Dataset C). The blue line represents the estimations from DeepVOG and the red line from the gold-standard clinical VOG setup and eye tracking system (EyeSeeCam, EyeSeeTec GmbH, F + 1/4rstenfeldbruck, Germany). The oculomotor examination comprises measurement of saccades, fixation nystagmus, smooth pursuit and optokinetic nystagmus (shaded regions, from left to right). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
(a) Absolute horizontal and vertical angular errors (in °), across calibration paradigms (Free-looking, CalibMe and Clinical standard) and three trials. (b) Intra-class correlation coefficient (ICC) as a measure for the repeatability of DeepVOG method under each calibration paradigm.

(a)

	1st Trial	2nd Trial	3rd Trial
	Free-looking	Free-looking	Free-looking
Horizontal	0.364 (0.187–0.691)	0.383 (0.211–0.763)	0.404 (0.209–0.837)
Vertical	0.54 (0.2–1.223)	0.477 (0.191–1.116)	0.476 (0.189–1.16)
	CalibMe	CalibMe	CalibMe
Horizontal	0.309 (0.158–0.557)	0.329 (0.165–0.717)	0.315 (0.158–0.574)
Vertical	0.569 (0.22–1.144)	0.591 (0.235–1.213)	0.565 (0.22–1.167)
	Narrow-ranged	Narrow-ranged	Narrow-ranged
Horizontal	0.472 (0.191–1.209)	0.547 (0.208–1.464)	0.604 (0.235–1.512)
Vertical	0.561 (0.248–1.172)	0.59 (0.255–1.21)	0.567 (0.258–1.212)

(b)

Intra-class correlation coefficient	Free-looking	CalibMe	Narrow-ranged
Horizontal	0.996	0.998	0.980
Vertical	0.996	0.998	0.958

contour estimates and blink events, as well as a measure of confidence for these values. In addition, gaze direction is estimated using an established method for 3D spherical eyeball model fitting, which we improve upon by incorporating confidence estimates from our network. Our results show our FCNN-based pupil segmentation, center localization and blink detection to be highly accurate and robust. Likewise, FCNN-based elliptical pupil contours are accurate enough to be directly used for robust, accurate and repeatable gaze estimation, which we validated with a clinical gold standard VOG system.

4.1. Utility of DeepVOG

Traditionally, and in most related works, hand-engineered steps such as thresholding, edge detection and rejection were used to segment the pupil area in images. Instead, we utilize a deep-learning model to autonomously learn the optimal image filters and segmentation rules from training data. The trained FCNN model yields robust pupil fitting results even in noisy, underexposed and artifact-ridden images. Though trained on hand-annotated data from our institute, we demonstrated a high level of generalizability to new datasets from various eye tracking setups, which is why we hope that DeepVOG can be readily used by other research labs in the community. It should be emphasized that DeepVOG can segment a pupil's shape robustly from low-contrast images with heterogenous illuminance and highly dilated pupils. This makes it particularly useful in low-light environments such as darkened

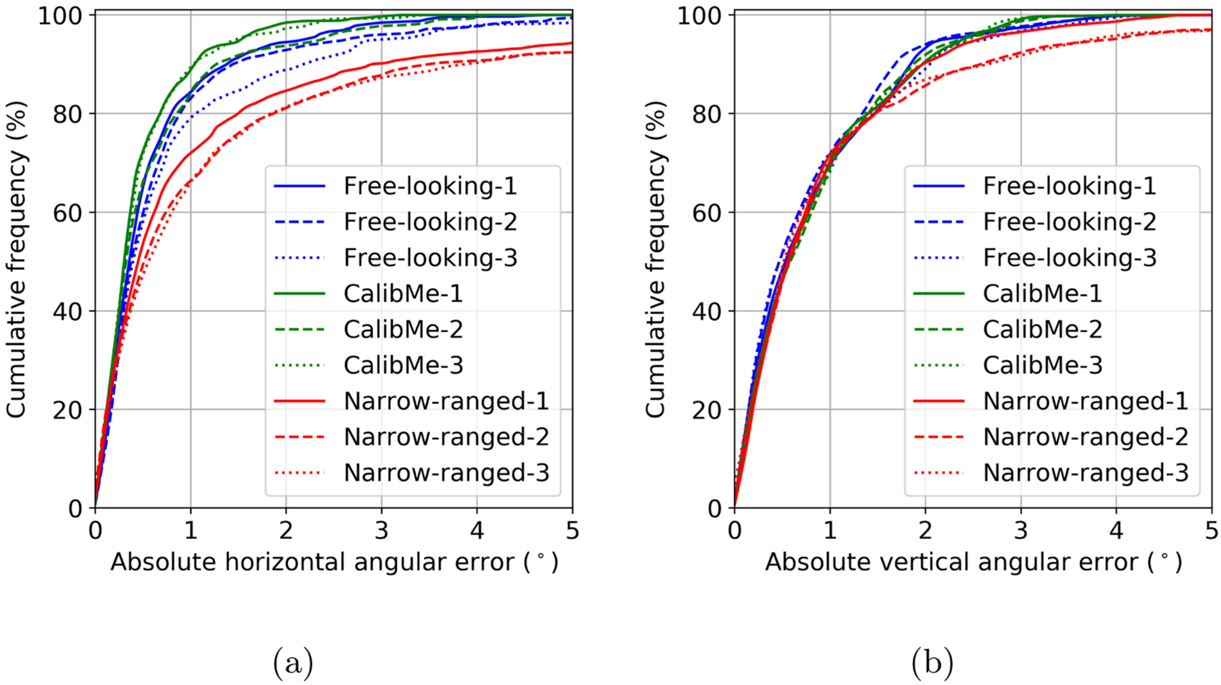


Fig. 11. Cumulative frequency plots for absolute (a) horizontal and (b) vertical angular errors, for three calibration paradigms (free-looking, CalibMe and narrow-ranged, represented by blue, green and red line respectively). To assess repeatability, each paradigm was repeated three times, represented by solid, dashed and dotted lines, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

MRI scanner rooms, a scenario for which we had difficulties performing eye tracking using established solutions. The unique feature of pupil region segmentation separates DeepVOG from previously proposed CNN-based approaches that only infer pupil centre coordinates (Fuhl et al., 2016; Chinsatit and Saitoh, 2017). This not only allows gaze inference based on pupil shape, but also enables other applications in neuroscientific research. For example, a change in pupil size is an automatic response to affective stimuli and an objective measure for emotional arousal (Bradley et al., 2008). The pupil segmentation may then aid or even replace the extra measurements of skin conductance and heart rate in some studies of emotion. Additionally, the output as a probability map informs the user about the confidence of the segmentation, which gives valuable information on data reliability, interpretation and blink detection.

The pupil ellipse estimates and confidence estimates from our FCNN lay the foundation for accurate gaze estimation with median angular errors of around 0.5°, as compared to RMSE of 1.6° in the original study of (Świrski and Dodgson, 2013), and 0.59° in EyeRecToo (Santini et al., 2017), one of the best-performing, recently proposed methods. We further show that if the network's confidence output is considered for 3D model fitting and gaze estimation, the accuracy can be further improved to angular errors around 0.38°–0.45°. Such accuracy could improve the validity of results in eye-tracking based experiments, for example, clinical assessment of vestibular and ocular motor disorders as well as visual attention studies in cognitive neuroscience. Further, DeepVOG demonstrates a high repeatability given multiple trials of two unassisted calibration paradigms, making it a stable tool for gaze data acquisition. Naturally, a projector-assisted, fixation-based calibration routine as in the neuro-ophthalmological examination laboratory of our clinical center can further improve the accuracy of gaze estimates. However, if such a procedure is impossible, for example due to hardware constraints, or in patients with fixation problems, the investigated unassisted calibration and gaze estimations in DeepVOG might be a very interesting option. Finally, we highlight the accessibility of DeepVOG as an open-source software, which does not depend on corneal reflections or stimulus-based calibrations, leaving a head-mounted low-cost camera as the only required equipment.

4.2. *Limitations and future work*

Even though DeepVOG's FCNN-based pupil segmentation can generalize well to unseen datasets, mis-segmentations still do occur (cf. Fig. 7). In particular, if videos are recorded from a longer distance, thus containing other facial features such as eyebrows or the nose, DeepVOG is likely to fail, since it did not encounter such images during training. Further, if DeepVOG is used for gaze estimation, our experiments demonstrated that a narrow-angle calibration yields inferior accuracy during unassisted calibration. Hence, study conductors should make sure that study participants cover a sufficiently wide angular range of gaze directions (e.g. larger than 20°), to achieve highly elliptic pupil shapes ideally in the entire visual periphery. A fundamental limitation of the gaze estimation method which we employ in DeepVOG is the assumption of a spherical eye model, as proposed by Świrski and Dodgson (2013). Several improvements can be made here, since the real pupil is not exactly circular, and elliptical shapes are distorted by light refraction through the cornea. To this end, in a very recent work by Dierkes et al. (2018) and Pupil Labs Research (Pupil Labs GmbH, Berlin, Germany), the Le Grand eye model (Le Grand, 1968) was employed instead, which assumes the eye to consist of two intersecting spheres, i.e. the eyeball and the cornea. The non-elliptical appearance of pupils caused by corneal refraction leads to reported gaze estimation errors similar to those observed in our experiments (cf. Fig. 9d). An improved 3D eye model fitting loss function and algorithm were proposed (Dierkes et al., 2018), which could help in further improving gaze estimates in future work. Further, DeepVOG is not applicable in eye tracking setups where no video can be recorded and provided as input to the algorithm as a video file or as a real-time video stream. Certain eye tracking systems, especially those operating at high frequencies around 1 kHz (e.g. EyeLink 1000, SR Research, Ottawa, Canada), commonly process eye tracking data internally and do not provide an interface to high-quality video data in real-time and at a high framerate.

4.3. *Conclusion*

DeepVOG is a software solution for gaze estimation in neurological and neuroscientific experiments. It incorporates a novel pupil localization and segmentation approach based on a deep fully convolutional neural network. Pupil segmentation and gaze estimates are accurate, robust, fast and repeatable, under a wide range of eye appearances. We have made DeepVOG's pupil segmentation and gaze estimation components open-source and provide it to the community as freely available software modules for standalone video-oculography, or incorporation into already existing frameworks.

In future work, we aim to incorporate a large number of images from third-party public eye datasets into training of the DeepVOG FCNN. This would extend the FCNN's generalization capability and robustness to an even wider variety of eye and pupil appearances and avoid mis-segmentations that still do occur (cf. Fig. 7). An easy-to-use graphical user interface will also be a focus of development. To this end, it is possible to integrate our segmentation part into other existing frameworks where gaze inference is based on pupil information, since DeepVOG is modularised as two parts: pupil segmentation by FCNN and gaze estimation by Świrski et al. model. Especially Pupil Labs Research (Dierkes et al., 2018), with its more realistic Le Grand eye model (Le Grand, 1968) and its Python-based open-source user interface,[2] serves as an inspiration to our next step of improvement.

Acknowledgements

The study was supported by the German Federal Ministry of Education and Research (BMBF) in connection with the foundation of the German Center for Vertigo and Balance Disorders (DSGZ) (grant number 01 EO 0901), and a stipend of the Graduate School of Systemic Neurosciences (DFG-GSC 82/3; TR: Oct. 2015–Jan. 2016; Mar. 2018–Sept. 2018).

References

Bednarik, T., Kinnunen, A., Mihaila, P., 2005. FrSnti: Eye-Movements as a Biometric, vol. 3540. pp. 780–789.

Ben Slama, A., Mouelhi, A., Sahli, H., Manoubi, S., Mbarek, C., Trabelsi, H., Fnaiech, F., Sayadi, M., 2017. A new preprocessing parameter estimation based on geodesic active contour model for automatic vestibular neuritis diagnosis. Artif. Intell. Med. 80, 48–62. https://doi.org/10.1016/j.artmed.2017.07.005.

Bradley, M.M., Miccoli, L., Escrig, M.A., Lang, P.J., 2008. The pupil as a measure of emotional arousal and autonomic activation. Psychophysiology 45 (4), 602–607. https://doi.org/10.1111/j.1469-8986.2008.00654.x.

Chen, J., Ji, Q., 2008. 3d gaze estimation with a single camera without IR illumination. 2008 19th International Conference on Pattern Recognition 1–4. https://doi.org/10.1109/ICPR.2008.4761343.

Chinsatit, W., Saitoh, T., 2017. CNN-based pupil center detection for wearable gaze estimation system. Appl. Comput. Intell. Soft Comput. https://doi.org/10.1155/2017/2017/8718956.

Dierkes, K., Kassner, M., Bulling, A., 2018. A novel approach to single camera, glint-free 3d eye model fitting including corneal refraction. Proceedings of the 2018 ACM Symposium on Eye Tracking Research Applications, ETRA'18 9:1–9:9.

Fuhl, W., Kübler, T., Sippel, K., Rosenstiel, W., Kasneci, E., 2015a. Excuse: robust pupil detection in real-world scenarios. In: Azzopardi, G., Petkov, N. (Eds.), Computer Analysis of Images and Patterns. Springer International Publishing, Cham, pp. 39–51.

Fuhl, W., Santini, T.C., Kübler, T.C., Kasneci, E., 2015b. Else: Ellipse Selection for Robust Pupil Detection in Real-World Environments. CoRR abs/1511.06575.

Fuhl, W., Santini, T., Kasneci, G., Kasneci, E., 2016. Pupilnet: Convolutional Neural Networks for Robust Pupil Detection. CoRR abs/1601.04902.

Guestrin, E.D., Eizenman, M., 2006. General theory of rermote gaze estimation using the pupil center and corneal reflections. IEEE Trans. Biomed. Eng. 53 (6), 1124–1133.

Horng, W.-B., Chen, C.-Y., Chang, Y., Fan, C.-H., 2004. Driver fatigue detection based on eye tracking and dynamic template matching. IEEE International Conference on Networking, Sensing and Control, 2004, vol. 1 7–12. https://doi.org/10.1109/ICNSC.2004.1297400.

Ishikawa, T., Baker, S., Matthews, I., Kanade, T., 2004. Passive driver gaze tracking with active appearance models, Tech. Rep. CMU-RI-TR-04-08. Carnegie Mellon University, Pittsburgh, PA (February).

Koo, T., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropr. Med. 15 (03). https://doi.org/10.1016/j.jcm.2016.02.012.

Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., Torralba, A., 2016. Eye tracking for everyone. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Kumar, A., Passi, A., 2010. Comparison and combination of iris matchers for reliable personal authentication. Pattern Recogn. 43 (3), 1016–1026.

Kumar, N., Kohlbecher, S., Schneider, E., 2009. A novel approach to video-based pupil tracking. 2009 IEEE International Conference on Systems, Man and Cybernetics 1255–1262. https://doi.org/10.1109/ICSMC.2009.5345909.

Le Grand, Y., 1968. Light, Colour and Vision, 1st ed. Chapman and Hall, London, UK.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444. https://doi.org/10.1038/nature14539.

Li, D., Winfield, D., Parkhurst, D.J., 2005. Starburst: a hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) – Workshops 79. https://doi.org/10.1109/CVPR.2005.531.

Liang, Z., Tan, F., Chi, Z., 2012. Video-Based Biometric Identification Using Eye Tracking Technique. pp. 728–733.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sßnchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88. https://doi.org/10.1016/j.media.2017.07.005.

Lohse, G.L., 1997. Consumer eye movement patterns on yellow pages advertising. J. Advert. 26 (1), 61–73.

Milletari, F., Navab, N., Ahmadi, S., 2016. V-net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In 2016 Fourth International Conference on 3D Vision (3DV) 565–571 IEEE.

Moschner, C., Crawford, T.J., Heide, W., Trillenberg, P., Kömpf, D., Kennard, C., 1999. Deficits of smooth pursuit initiation in patients with degenerative cerebellar lesions. Brain 122 (11), 2147–2158. https://doi.org/10.1093/brain/122.11.2147.

Multimedia-University, 2019. MMU Iris Dataset. (accessed 20.01.19). http://www.cs.princeton.edu/andyz/irisrecognition.

Naqvi, R.A., Arsalan, M., Batchuluun, G., Yoon, H.S., Park, K.R., 2018. Deep learning-based gaze detection system for automobile drivers using a NIR camera sensor. Sensors 18 (2). https://doi.org/10.3390/s18020456.

Noh, H., Hong, S., Han, B., 2015. Learning Deconvolution Network for Semantic Segmentation. CoRR abs/1505.04366.

Rehder, B., Hoffman, A.B., 2005. Eyetracking and selective attention in category learning. Cogn. Psychol. 51 (1), 1–41.

Reutskaja, E., Nagel, R., Camerer, C.F., Rangel, A., 2011. Search dynamics in consumer choice under time pressure: an eye-tracking study. Am. Econ. Rev. 101 (2), 900–926.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional Networks for Biomedical Image Segmentation. CoRR abs/1505.04597.

Safaee-Rad, R., Tchoukanov, I., Smith, K.C., Benhabib, B., 1992. Three-dimensional location estimation of circular features for machine vision. IEEE Trans. Robot. Autom. 8 (5), 624–640. https://doi.org/10.1109/70.163786.

Santini, T., Fuhl, W., Geisler, D., Kasneci, E., 2017. Eyerectoo: Open-Source Software for Real-Time Pervasive Head-Mounted Eye Tracking. pp. 96–101.

Satriya, T., Wibirama, S., Ardiyanto, I., 2016. Robust pupil tracking algorithm based on ellipse fitting. 2016 International Symposium on Electronics and Smart Devices (ISESD) 253–257. https://doi.org/10.1109/ISESD.2016.7886728.

Schnipke, S.K., Todd, M.W., 2000. Trials and tribulations of using an eye-tracking system. CHI'00 Extended Abstracts on Human Factors in Computing Systems, CHI EA'00. ACM, New York, NY, USA, pp. 273–274. https://doi.org/10.1145/633292.6334.52.

Soille, P., 2003. Morphological Image Analysis: Principles and Applications, 2nd ed. Springer-Verlag, Berlin, Heidelberg.

Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.A., 2014. Striving for Simplicity: the All Convolutional Net. CoRR abs/1412.6806.

Świrski, L., Dodgson, N.A., 2013. A fully-automatic, temporal approach to single camera, glint-free 3d eye model fitting. Proceedings of ECEM 2013.

Świrski, L., Dodgson, N.A., 2014. Rendering synthetic ground truth images for eye tracker evaluation. Proceedings of ETRA 2014 219–222.

Świrski, L., Bulling, A., Dodgson, N., 2012. Robust real-time pupil tracking in highly off-axis images. Proceedings of the Symposium on Eye Tracking Research and Applications, ETRA'12 173–176.

Tonsen, M., Zhang, X., Sugano, Y., Bulling, A., 2015. Labeled Pupils in the Wild: A Dataset for Studying Pupil Detection in Unconstrained Environments. CoRR abs/1511.05768.

Yamazoe, H., Utsumi, A., Yonezawa, T., Abe, S., 2008. Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions. Proceedings of the 2008 Symposium on Eye Tracking Research & Applications – ETRA'08, 1 (212) 245. https://doi.org/10.1145/1344471.1344527.

Yang, G., Saniie, J., 2016. Eye tracking using monocular camera for gaze estimation applications. 2016 IEEE International Conference on Electro Information Technology (EIT) 0292–0296. https://doi.org/10.1109/EIT.2016.7535254.

[2] https://github.com/pupil-labs/pupil.