Journal of Intelligent & Fuzzy Systems 41 (2021) 5359–5365
DOI:10.3233/JIFS-189858
IOS Press

Automated segmentation of key structures of the eye using a light-weight two-step classifier

Adish Rao[a,*], Aniruddha Mysore[a], Siddhanth Ajri[a], Abhishek Guragol[a], Poulami Sarkar[a] and Gowri Srinivasa[b]
[a]*PES Center for Pattern Recognition, PES University Electronic City Campus, Bengaluru, India*
[b]*PES Center for Pattern Recognition and the Department of Computer Science and Engineering, PES University, Bengaluru, India*

Abstract. We present an automated approach to segment key structures of the eye, viz., the iris, pupil and sclera in images obtained using an Augmented Reality (AR)/ Virtual Reality (VR) application. This is done using a two-step classifier: In the first step, we use an auto encoder-decoder network to obtain a pixel-wise classification of regions that comprise the iris, sclera and the background (image pixels that are outside the region of the eye). In the second step, we perform a pixel-wise classification of the iris region to delineate the pupil. The images in the study are from the OpenEDS challenge and were used to evaluate both the accuracy and computational cost of the proposed segmentation method. Our approach achieved a score of 0.93 on the leaderboard, outperforming the baseline model by achieving a higher accuracy and using a smaller number of parameters. These results demonstrate the great promise pipelined models hold along with the benefit of using domain-specific processing and feature engineering in conjunction with deep-learning based approaches for segmentation tasks.

Keywords: Augmented reality, computer vision, image segmentation, image processing, virtual reality

1. Introduction

With the rapid emergence of new technologies in the Augmented Reality (AR) and Virtual Reality (VR) space, providing a good user experience has become vital for the success of consumer products that make use of such technologies [1]. Consequently, methods for improving a device's understanding of its position with respect to the user has become an important subject for research [2]. Eye tracking is one such method that has been shown to play a pivotal role in quantifying a user's attention. Many eye-tracking solutions that have been proposed [3] require accurate estimation of eye-features in two-dimensional images, typically per-pixel segmentation of the key eye regions: the sclera [4–6], the iris [7, 8], the pupil [9], and everything else (background).

Segmentation of the eye poses some key challenges on account of the diverse shapes different eye regions can take on. An added issue is the image variation caused due to factors like demographics, use of makeup and spectacles. These translate to a need for any solution to be robust to these variations. Further, noise in the images is evident in the form of overexposure or presence of bright spots. Finally, to complicate the task, the proposed solutions need to be computationally inexpensive due to the limited resources available on AR/VR devices.

The OpenEDS paper from Facebook Research presented a dataset of images of the human eye along with pixel-level annotations for key eye-regions [10]. The OpenEDS challenge hosted by Facebook on the

*Corresponding author. Adish Rao, PES Center for Pattern Recognition, PES University Electronic City Campus, Bengaluru-560100, India. E-mail: contact.adishrao@gmail.com.

CloudCV platform required us to develop a robust and power efficient eye tracking solution that performs better than the baseline model reported in the original paper.

In this paper we propound a novel approach for identifying key regions of the eye using a low complexity architecture while maintaining a high accuracy. Our solution builds upon existing examples of Deep Convolutional Neural Networks (Deep CNN's) applied to the task of segmentation, and we propose a new way of using dual, pipelined models to preserve performance while downsizing model size [11–13]. This pipeline also includes preprocessing and postprocessing modules. Preprocessing involves resizing of images prior to being input to the dual model. The solution proposed also makes use of the connected components postprocessing technique to further improve the model performance. This use of a dual model approach combined with the connected components postprocessing technique endorses the utilization of pipelined models for the task of segmentation if they take advantage of appropriate preprocessing and postprocessing techniques.

2. Related work

The Facebook OpenEDS challege provides a platform for finding a solution to this complex problem. In response to this challenge, researchers have explored numerous approaches, some of which are discussed in this section [14]. Boutros et al., presented a novel technique for substantially reducing the model size by removing less important parameters thereby reducing the size of the feature map [15]. Chaudhary et al., presented RITNET for real time segmentation by using a combination of U-Net and Densenet [16]. Perry et al., proposed the MinNet model that aims at solving the same problem using an encoder-decoder convolutional neural network with dilation in its layers [17].

2.1. Other eye segmentation tasks

Many other segmentation techniques have been developed for myriad tasks over the years [18]. Stember et al., use the popular U-Net architecture to monitor differences, if any, in the results of segmentation between images that were annotated using Eye tracking hardware, as opposed to human annotation [19]. Luo et al. proposed a segmentation on low resolution images, from the HELEN, 300VW, CVL and Columbia Gaze datasets that were manually annotated to have 2 ground truths: 30 point landmarks and pixel-level segmentation [20]. The segmentation model itself was a combination of fast RCNN and ridge regression models used for eye-region detection followed by use of landmark models (Active Appearance Model (AAM), Ensemble Regression Tree (ERT) and Supervised Descent Method (SDM) respectively) to generate iris and pupil boundaries, further followed up by deep segmentation models Atrous Convolutional Neural Network with Conditional Random Field (ACNN + CRF).

Yiu et al., tackled the task of single class, pixel segmentation and from the segmented pixel region, performed gaze estimation using a model termed DeepVOG that utilised a modified U-Net and V-Net FCNN architecture [21]. Rot et al., made use of a model with no fully connected layers and the Decoder was an inverted VGG16 model [22].

Ramlee et al., use extensive pre-processing for segmentation of non-circular and abnormal pixels by making use of filtering to remove noise, followed by illumination removal, logarithmic transform, power transform, morphology operator and object removal to remove smaller objects with values less than specified number of pixels [23]. The resultant image of black and white pixels serves as a mask that is overlapped with the original image to form a boundary around the region of interest.

3. The OpenEDS dataset

The dataset used in this paper has been created by and released as a workshop challenge at ICCV by Facebook. It has been pre-split into three sets, comprising 8916 train images, 2403 validation images and 1440 test images respectively. The labels have been included for the train and validation sets, along with their respective masks. The masks for the images are of dimensions 640×400 to represent which class each individual pixel in the image belongs to. The values of each element in the array are: a value of 0 for the regions exterior to the eye, a value of 1 for the sclera region, a value of 2 for the iris region, and a value of 3 for the pupil region as given in Table 1.

The images in the dataset revealed extensive variations, as seen in Fig. 1. Few of the images were of eyes that were fully open, making it easy to identify the different regions, while few of the images range from mid-blink to completely closed. These images of mid-blink to completely closed posed a challenge

Table 1
Mask Values and Eye Regions

Value in segmentation mask	Region of the eye
0	Regions exterior to the eye
1	Sclera
2	Iris
3	Pupil

in accurately segmenting the pupil region. Further complexities in the dataset included images where the subject is wearing glasses, owing to this these images exhibited large amounts of glare from the glasses reflecting light off from the headset. This glare posed a considerable hurdle for accurate segmentation.

Finally, there were images in which there was a marked use of makeup, mainly eyeliner, which also influenced the accuracy of the model. The same variations of mid-blink to fully closed eyes were also exhibited in the images with glasses and images with makeup. Upon further inspection of the images in the test set, it was discovered that a reasonable number of images were blurred and a few images were overexposed which caused problems in accurately resolving the segmentation masks.

4. Methodology

The methodology overview has been represented in Fig. 2 and a detailed description of each step is provided below.

4.1. System configuration

The model was trained on a Kaggle Cloud provisioned Nvidia Tesla P100 machine.

4.2. Pre-processing

Due to the enormous size of the data, with a total of 12, 759 images each of size 640 × 400, the limited computing power at hand for the construction of the convolutional neural network, and knowledge of the limited resources of the platform on which the model was to be deployed, the essential and only pre-processing step was the compression of

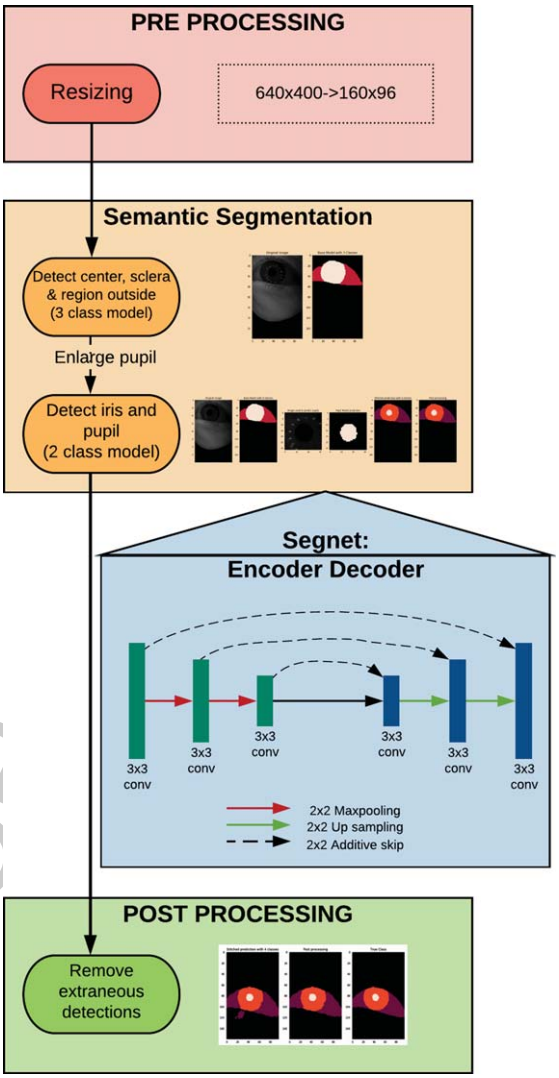

Fig. 2. A schematic diagram of the automated segmentation workflow

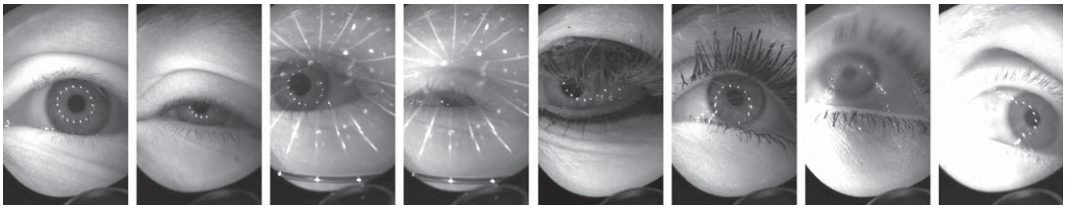

Fig. 1. Sample eye images showing the high variability in the dataset. From left to right: an open eye, a closed eye, eyes in spectacles, closed eyes in spectacles, sample with makeup, another sample with makeup, blurred image, overexposed and bright image.

the given data from images of size 640x400 into images of size 160x96 in order to maintain the image aspect ratio while still retaining an image of dimension%32. The corresponding masks for the image set are also compressed in this fashion. The reasoning behind maintaining dimensions as a multiple of 32 is due to the fact that our custom encoder-decoder network maintains neurons at each stage as a multiple of 32. However, the width of the network is kept as compact as possible while still obtaining favourable results in order to ensure low model complexity.

4.3. Pipelined encoder-decoder model with skip connections

We have used an encoder-decoder network based on the SegNet architecture to classify different key regions of the eye [24]. The network classifies each pixel of an input image into one of four classes; pupil, iris, sclera, and regions exterior to the eye. Our solution posits a two-fold implementation of encoder-decoder networks to classify each pixel of an input image. The first network is a base model that predicts 3 classes; the sclera, iris, and outside regions. This has been outlined in Table 2.

The second model, as represented by Table 3. Predicts the pixels forming the pupil of the eye inside the iris using an input image of size 32×32, the general location of which is obtained from the output of the first model. This is done by first locating the initial row and final row along with the initial column and final column within the image classified as the iris and then extracting the image of size 32×32 using the midpoint of the previously obtained rows and columns values [(initial row value + final row value)/2, (initial column value + final column value)/2]. Both models use skip connections that are interlaced between the convolution layers. The use of skip connections was necessitated by poor MIoU score as well as high model complexity when using a deeper network that did not provide alternative paths for gradient convergence. Results of experimentation with both additive as well as multiplicative skip connections showed that additive skip connections outperformed the multiplicative ones.

The final segmentation output is obtained by classifying each pixel into three categories using the base model and identifying the pupil's region using the pupil model.

Table 2
Base Model

Layer (Type)	Shape	Parameters
input_1 (InputLayer)	160, 96, 3	0
conv_1 (Conv2D)	160, 96, 32	896
pool_1 (MaxPooling2D)	80, 48, 32	0
conv_2 (Conv2D)	80, 48, 64	18496
pool_2 (MaxPooling2D)	40, 24, 64	0
conv_3 (Conv2D)	40, 24, 32	18464
conv_4 (Conv2D)	40, 24, 32	9248
conv_5 (Conv2D)	40, 24, 32	9248
add_2 (Add)	40, 24, 32	0
conv_6 (Conv2D)	40, 24, 32	9248
add_3 (Add)	40, 24, 32	0
up_samp_1 (UpSampling2D)	80, 48, 32	0
concat_1 (Concatenate)	80, 48, 96	0
conv_7 (Conv2D)	80, 48, 64	55360
add_4 (Add)	80, 48, 64	0
add_5 (Add)	80, 48, 64	0
up_samp_2 (UpSampling2D)	160, 96, 64	0
concat_2 (Concatenate)	160, 96, 96	0
conv_8 (Conv2D)	160, 96, 32	27680
add_6 (Add)	160, 96, 32	0
add_7 (Add)	160, 96, 32	0
conv_9 (Conv2D)	160, 96, 3	99
act_1 (Activation)	160, 96, 3	0

Table 3
Pupils model

Layer (Type)	Shape	Parameters
input_2 (InputLayer)	32, 32, 3	0
conv_10 (Conv2D)	32, 32, 32	896
pool_3 (MaxPooling2D)	16, 16, 32	0
conv_11 (Conv2D)	16, 16, 32	1056
pool_4 (MaxPooling2D)	8, 8, 32	0
conv_12 (Conv2D)	8, 8, 32	9248
conv_13 (Conv2D)	8, 8, 32	1056
add_8 (Add)	8, 8, 32	0
up_samp_3 (UpSampling2D)	16, 16, 32	0
concat_3 (Concatenate)	16, 16, 64	0
conv_14 (Conv2D)	16, 16, 32	2080
add_9 (Add)	16, 16, 32	0
up_samp_4 (UpSampling2D)	32, 32, 32	0
concat_4 (Concatenate)	32, 32, 64	0
conv_15 (Conv2D)	32, 32, 32	18464
add_10 (Add)	32, 32, 32	0
conv_16 (Conv2D)	32, 32, 2	66
act_2 (Activation)	32, 32, 2	0

4.4. Architecture

Our model is based on SegNet, which uses the encoder-decoder architecture. The encoder part of the neural network takes an image as input and generates a high-dimensional feature vector as its output.

We run this feature vector through several dimensionality reduction techniques to bring down our model parameters to 148, 739 for the base-model and 32,866 for the pupil-model. One such dimen-

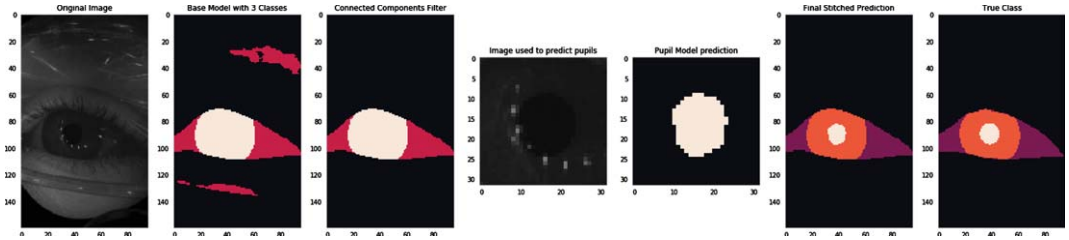

Fig. 3. Sample images showing representative results: The first-step model delineates the iris in yellow, sclera in red and the second-step model delineates the iris in red, pupil in yellow (the sclera is the same region as delineated in the first-step).

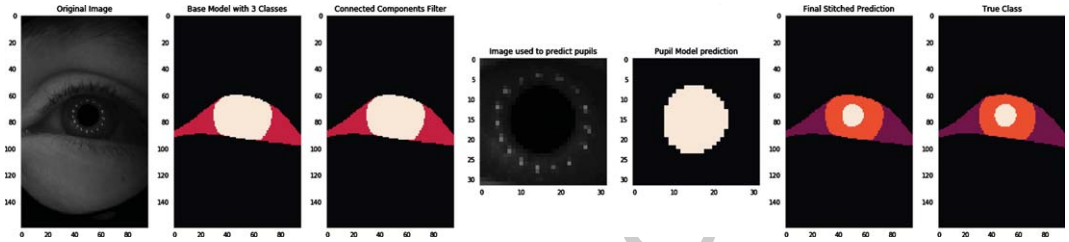

Fig. 4. Sample images showing representative results: An accurate prediction.

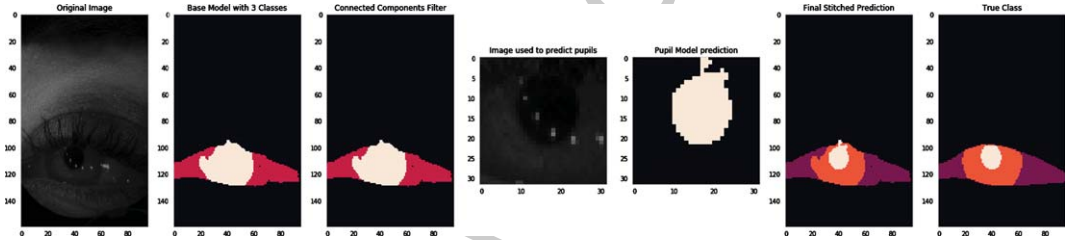

Fig. 5. Sample images showing representative results: An erroneous prediction.

sionality reduction technique is downsampling using max-pooling. The decoder part of the neural network uses upsampling between the convolution layers. In addition interleaved additive skip connections significantly contribute to improving the model performance.

4.5. Post processing

As seen in Fig. 3, the predicted output of the stitched neural network yields a significant amount of spurious detection. We use image processing techniques to detect and remove these erroneous labels. In particular, we have used connected components to find extraneous detection in the predicted output of the base model, and eliminate them based on their weighted area, before extracting the smaller image for predicting pupils.

The predictions obtained are of dimensions 160×96 as the images were initially scaled-down during the pre-processing step. These predictions need to now be scaled-up back to the original dimensions of 640×400. During the process of scaling-down, some amount of information is lost due to the lower image resolution and while scaling the image up this leads to an increase in errors.

5. Results

Our model has achieved an mIoU of 0.89353 with a model complexity of 181, 605. The score awarded to us based on the metric used for the challenge was 0.94677. The model created by the OpenEDS team attained a mIoU of 0.89478 and a model complexity of 416, 088.00000 with a score of 0.7624. Our model clearly outperforms the original model proposed by the openEDS team. Figures 4 and 5 demonstrate sample resultant predictions made by our model.

Timing analysis of the final model yielded an average inference time of 0.26 seconds per image for a dataset of 10 images. Note that this calculation reflects the inference time when each image is predicted individually by the model rather than a batch prediction, in order to better reflect the use-case of live video frames being captured by a VR headset. The prediction times were measured on a Python notebook running on Kaggle without a GPU since we expect most embedded microprocessors will not have a powerful GPU.

6. Conclusion

The solution we present to the FaceBook OpenEDS challenge outperforms the model proposed in the original paper. Akin to the original paper, we have also used skip connections. However, having taken inspiration from numerous segmentation techniques previously mentioned, we have devised a novel approach to reduce the model size while maintaining accuracy. Our approach uses two models - one model that classifies the image into non-eye, sclera, and combined pupil-iris classes, and another that classifies an enlarged image into non-pupil and pupil classes. The output of the second model is then stitched into the output of the first model. This dual model approach exhibited better results than the original model, in that it is light-weight when compared to the current state-of-the-art techniques and only marginally less accurate.

It is noteworthy that this accuracy was obtained by training the model on the raw data without the use of any augmentation techniques or any complex preprocessing which could further improve the model's performance. This shows that pipelined models still hold weight in segmentation tasks (as opposed to the non-pipelined method employed in the original paper) and given the application of the right set of preprocessing and postprocessing steps, along with image augmentation, can obtain high accuracy with a lightweight model, which would be ideal and necessary for deployment in edge devices. Extended research into the use of dual models for semantic segmentation and the appropriate preprocessing and postprocessing techniques that ought to be used in conjunction with these types of models would further reduce model complexity and enhance the accuracy of the segmentation task.

References

[1] K. Harezlak and P. Kasprowski, Application of eye tracking in medicine: A survey, research issues and challenges, *Computerized Medical Imaging and Graphics* **65** (2018), 176–190.

[2] A. Kennedy, Book review: Eye tracking: A comprehensive guide to methods and measures, *Quarterly Journal of Experimental Psychology* **69**(3) (2016), 607–609.

[3] K.A.F. Mora, F. Monay and J.-M. Odobez, EYEDIAP, in *Proceedings of the Symposium on Eye Tracking Research and Applications – ETRA '14*. ACM Press, 2014.

[4] A. Das, U. Pal, M. Blumenstein and M.A.F. Ballester, Sclera recognition – a survey, in *2013 2nd IAPR Asian Conference on Pattern Recognition*, IEEE, nov 2013.

[5] A. Das, U. Pal, M.A. Ferrer, M. Blumenstein, D. Stepec, P. Rot, Z. Emersic, P. Peer, V. Struc, S.V.A. Kumar and B.S. Harish, SSERBC 2017: Sclera segmentation and eye recognition benchmarking competition, in *2017 IEEE International Joint Conference on Biometrics (.CB)*. IEEE, oct 2017.

[6] A. Das, U. Pal, M.A. Ferrer, M. Blumenstein, D. Stepec, P. Rot, Z. Emersic, P. Peer and V. Struc, SSBC 2018: Sclera segmentation benchmarking competition, in *2018 International Conference on Biometrics (ICB)*. IEEE, feb 2018.

[7] R. Satish and P.R. Kumar, State-of-the art iris segmentation methods a survey, *International Journal of Computer Sciences and Engineering* **6**(11) (2018), 739–748.

[8] W. Sankowski, K. Grabowski, M. Napieralska, M. Zubert and A. Napieralski, Reliable algorithm for iris segmentation in eye image, *Image and Vision Computing* **28**(2) (2010), 231–237.

[9] M. Tonsen, X. Zhang, Y. Sugano and A. Bulling, Labelled pupils in the wild, in *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications – ETRA '16*. ACM Press, 2016.

[10] S.J. Garbin, Y. Shen, I. Schuetz, R. Cavin, G. Hughes and S.S. Talathi, Openeds: Open eye dataset, 2019.

[11] K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, jun 2016.

[12] J. Long, E. Shelhamer and T. Darrell, Fully convolutional networks for semantic segmentation, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, jun 2015.

[13] D.R. Lucio, R. Laroca, E. Severo, A.S. Britto and D. Menotti, Fully convolutional networks and generative adversarial networks applied to sclera segmentation, in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, IEEE, oct 2018.

[14] V.T. Huynh, S.-H. Kim, G.-S. Lee and H.-J. Yang, Eye semantic segmentation with a lightweight model, in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, IEEE, oct 2019.

[15] F. Boutros, N. Damer, F. Kirchbuchner and A. Kuijper, Eye-MMS: Miniature multi-scale segmentation network of key eye-regions in embedded applications, in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, IEEE, oct 2019.

[16] A.K. Chaudhary, R. Kothari, M. Acharya, S. Dangi, N. Nair, R. Bailey, C. Kanan, G. Diaz and J.B. Pelz, RITnet: Real-time semantic segmentation of the eye for gaze tracking,

in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, IEEE, oct 2019.

[17] J. Perry and A. Fernandez, MinENet: A dilated CNN for semantic segmentation of eye features, in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, IEEE, oct 2019.

[18] C.D. McMurrough, V. Metsis, J. Rich and F. Makedon, An eye tracking dataset for point of gaze detection, in *Proceedings of the Symposium on Eye Tracking Research and Applications – ETRA '12*. ACM Press, 2012.

[19] J.N. Stember, H. Celik, E. Krupinski, P.D. Chang, S. Mutasa, B.J. Wood, A. Lignelli, G. Moonis, L.H. Schwartz, S. Jambawalikar and U. Bagci, Eye tracking for deep learning segmentation using convolutional neural networks, *Journal of Digital Imaging* **32**(4) (2019), 597–604.

[20] B. Luo, J. Shen, Y. Wang and M. Pantic, The ibug eye segmentation dataset, 2019.

[21] Y.-H. Yiu, M. Aboulatta, T. Raiser, L. Ophey, V.L. Flanagin, P. zu Eulenburg and S.-A. Ahmadi, DeepVOG: Open-source pupil segmentation and gaze estimation in neuroscience using deep learning, *Journal of Neuroscience Methods* **324** 108307, 2019.

[22] P. Rot, Z. Emersic, V. Struc and P. Peer, Deep multi-class eye segmentation for ocular biometrics, in *2018 IEEE InternationalWork Conference on Bioinspired Intelligence (IWOBI)*, IEEE, jul 2018.

[23] R.A. Ramlee, A.R. Ramli and Z.M. Noh, Pupil segmentation of abnormal eye using image enhancement in spatial domain, *IOP Conference Series: Materials Science and Engineering* **210** 012031, 2017.

[24] V. Badrinarayanan, A. Kendall and R. Cipolla, Segnet: Adeep convolutional encoder-decoder architecture for image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.