The iBUG Eye Segmentation Dataset

Bingnan Luo
Intellignet Behaviour Understanding Group, Imperial College London, United Kingdom
bingnan.luo16@imperial.ac.uk

Jie Shen
Intellignet Behaviour Understanding Group, Imperial College London, United Kingdom
jie.shen07@imeprail.ac.uk

Yujiang Wang
Intellignet Behaviour Understanding Group, Imperial College London, United Kingdom
yujiang.wang14@imeprail.ac.uk

Maja Pantic
Intellignet Behaviour Understanding Group, Imperial College London, United Kingdom
m.pantic@imeprail.ac.uk

Abstract

This paper presents the first dataset for eye segmentation in low resolution images. Although eye segmentation has long been a vital preprocessing step in biometric applications, this work is the first to focus on low resolutions image that can be expected from a consumer-grade camera under conventional human-computer interaction and/or video-chat scenarios. Existing eye datasets have multiple limitations, including: (a) datasets only contain high resolution images; (b) datasets did not include enough pose variations; (c) a utility landmark ground truth did not be provided; (d) high accurate pixel-level ground truths had not be given. Our dataset meets all the above conditions and requirements for different segmentation methods. Besides, a baseline experiment has been performed on our dataset to evaluate the performances of landmark models (Active Appearance Model, Ensemble Regression Tree and Supervised Descent Method) and deep semantic segmentation models (Atrous convolutional neural network with conditional random field). Since the novelty of our dataset is to segment the iris and the sclera areas, we evaluate above models on sclera and iris only respectively in order to indicate the feasibility on eye-partial segmentation tasks. In conclusion, based on our dataset, deep segmentation methods performed better in terms of IOU-based ROC curves and it showed potential abilities on low-resolution eye segmentation task.

2012 ACM Subject Classification Computing methodologies → Image segmentation

Keywords and phrases dataset, eye, segmentation, landmark, pixel-level

Digital Object Identifier 10.4230/OASIcs.ICCSW.2018.7

Category Main Track

1 Introduction

Eyes not only are the most vital sensory organ but also play a crucial role in conveying a person's emotional state and mental wellbeing [5]. Although there have been numerous works on blink detection [1, 8, 10], we argue that accurate segmentation of sclera and iris can provide much more information than blinks alone, thus allowing us to study the finer details of eye movement such as cascade, fixation and other gaze patterns. As a pre-processing step in iris recognition, iris segmentation in high resolution expression – less frontal face

2018 Imperial College Computing Student Workshop (ICCSW 2018).
Editors: Edoardo Pirovano and Eva Graversen; Article No. 7; pp. 7:1–7:9
OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

images have been well studied by the biometric community. However, the commonly used Hough-transform-based method [14] does not work well on low-resolution images captured under normal human-computer interaction (HCI) and/or video-chat scenarios, when the boundary of eyes and iris are blurry and the shape of the eye can differ greatly due to pose variation and facial expression. To our knowledge, this work presents the first effort in solving the eye segmentation problem under such challenging conditions.

To investigate the topic of eye segmentation in low-resolution images, the first problem we need to address is the lack of data. Albeit both biometric community and facial analysis community published an abundance of datasets over the years, none can be used as is for our purpose because the former category only contains high resolution eye scans while the latter category lacks annotation of segmentation masks for sclera and iris. Therefore, during the course this work, we created a sizable eye segmentation dataset by manually annotating images selected from HELEN[9], 300VW[12], CVL[11] and Columbia Gaze[13] datasets.

After establish our dataset, it is necessary to evaluate how good performances are based on two types of ground truths. Therefore, deformable and deep segmentation models were chosen. Active Appearance Model(AAM)[3], Ensemble Regression Tree(ERT)[7] and Supervised Descent Method(SDM)[15] were compared with deep semantic segmentation methods(DeepLab[2] proposed): Atrous Neural Network with Conditional Random Field (ACNN+CRF). For deformable models, the segmentation of non-frontal faces is a big challenge because of occlusion, shape deformation and initializations. Therefore, the deep segmentation methods can relatively compensate this shortcoming, whose performance can be more stable especially on non-frontal faces. Otherwise, since we also want to know performances on iris-only and sclera-only segmentations, all models were utilized and trained by iris-background and sclera-background data samples. Finally, all segmentation results are evaluated and discussed based on Interaction over Union(IOU) and Receiver operating characteristic(ROC) curves. Based on that, the model with the best performance will be considered as the potential model in eye segmentation researches.

2 Relative Works

Image segmentation is one of the oldest computer vision problems studied by the community. Early approaches often rely on finding edges and/or specific shapes in the image or hand-craft feature maps. Albeit dated, this kind of simple methods are still being used for eye segmentation as a pre-processing step for iris recognition [14]. In this case, the eye and iris are modeled respectively, by two parabolic curves and an ellipse. The parameters of the curves are then determined using Hough transform performed on the image's edge map. Because this approach is sensitive to image noise and even slight shape variation caused by head-pose and/or facial expressions, it cannot be use for eye segmentation in images captured by consumer-grade camera under normal HCI/video-chat conditions.

On the other hand, various sparse 2D deformable-model-based methods, such as AAM, SDM or ERT, have shown promising results in image segmentation, and in particular, facial landmark localization. These methods work by finding a local minimum in the parametric space that can optimally describe the object's shapes and appearance. Since image intensity space contains multiple local minimums so that it is uneasy to control which local minimum it lies on. Moreover, the optimizing process of deformable models starts with the mean shape of object's intensity and geometry, thus the transformation of the landmark depends on the initialization and integrity of objects. Therefore, they share many common limitations, including sensitivities to initialization, occlusion, and out-of-plane rotation. In [3, 15, 7], the profile face landmark tracking is still a challenge. Thus deformable models can experienced as a candidate in our research, but the performance can be expectedly poor to profile faces.

Table 1 dataset statistical information.

Name	Value
Total number	3161
Non-frontal faces proportion	18.35%
low-resolution image proportion	66.97%
Number of bad illumination samples	10
Number of samples with glasses	185

The methods mentioned above rely on the prior knowledges (such as the number of point, landmark shapes or curves expressions) so that they cannot adapt variant images and head poses of eyes (profile faces or occlusions). To lighten the influences in the wild (illuminations and etc.) and adapt to multiple situations (tricky head poses or occlusions), more recently, various deep learning techniques have achieved impressive results in semantic segmentation of natural images, which is widely-used because of its better adaptability of performances in variant environments. In particular, DeepLab[2, 4] uses atrous convolutional neural network based on VGG-16 and ResNet-101 architectures to generate segmentation masked, refined by a fully-connected CRF layer with mean-field approximation for fast inference. The atrous convolutional neural network is the innovation of DeepLab in order to increase the ability of extracting global image features. They not only reduced the computational cost but also improved the generalization.

In previous works, there is no existed low-resolution eye dataset. However, facial datasets provide us a good sources to obtain eye images in different illuminations and pose variations. HELEN dataset[9] contains high resolution facial images in different situations (multi-faces, indoor/outdoor and etc.). 300VW[12] is a low-resolution facial video dataset captured in the wild. CVL[11] and Columbia Gaze[13] are two facial dataset technically captured in lab environment. Although these datasets cannot be utilized directly in our research, they can also regard as sources of our proposed dataset.

3 Data Description

We create the iBUG eye segmentation dataset by annotating a total of 3161 images selected from HELEN[9], 300VW[12], CVL[11] and Columbia Gaze[13] datasets. The dataset contains faces under various poses. Specifically, faces who look ahead and with slight rotations are frontal, the others are annotated sas non-frontal faces. We primarily focus on low resolution images, but a small number of high resolution images are also included for completeness. The distribution of eye-patch height is illustrated in Figure 1. Note that we use eye-patch height as a measure for image resolution because the widely-used interocular distance measure can be easily biased by face yaw. Last but not least, the dataset contains a small number of examples featuring partial occlusion and bad illumination. The detailed statistics can be found in Table 1.

Some examples of the annotated images are shown in Figure 2. The first row shows the source image and the location of the control points, while the second row visualizes the segmentation mask. Some extra statistical information has been presented in Table 1.

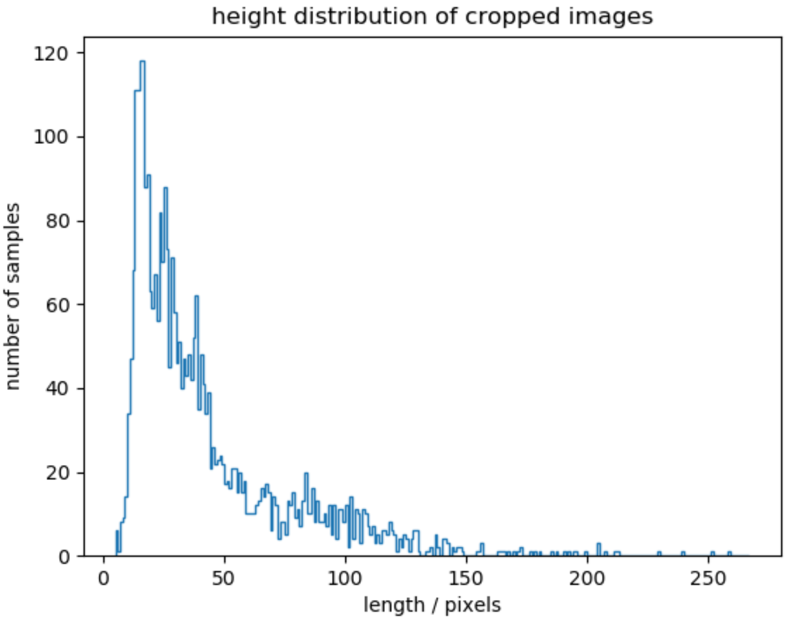


Figure 1 Height distribution of eye regions.

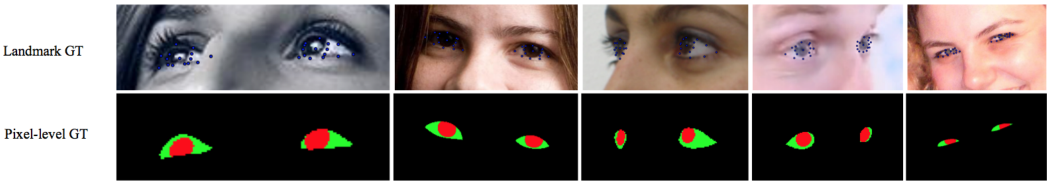


Figure 2 Some annotated images in our dataset.

4 Baseline Methods

In this work, we utilized deformable model-based methods (AAM, ERT and SDM) and DeepLab proposed Atrous CNN+CRF (CGG16+CRF and ResNet101+CRF) as the baseline method.

4.1 Landmark Models

Before we discuss details of utilizations of baseline methods, we are going to introduce some concepts about them. AAM, ERT and SDM are deformable statistic models which were generally used in object localization and alignment. The shape and texture will be transformed with a specific transformation function. Assume $x = \{x_1, x_2, ..., x_i\}$ indicates shapes of images and $g = \{g_1, g_2, ..., g_i\}$ presents textures of images. $\mathbf{x} = \bar{\mathbf{x}} + \mathbf{Q_s c_s}$ and $\mathbf{g} = \bar{\mathbf{g}} + \mathbf{Q_g c_g}$ demonstrate the transformation methods in deformable model-based methods. Q_s and Q_g are matrices describing the modes of variation derived from the training set. c_s is the shape model parameter and c_g is the appearance model parameter, which control the shape and texture gradient and the transforming directions. Thus, the aim of deformable model-based methods is to find the optimized local minimum between current image and the mean shape.

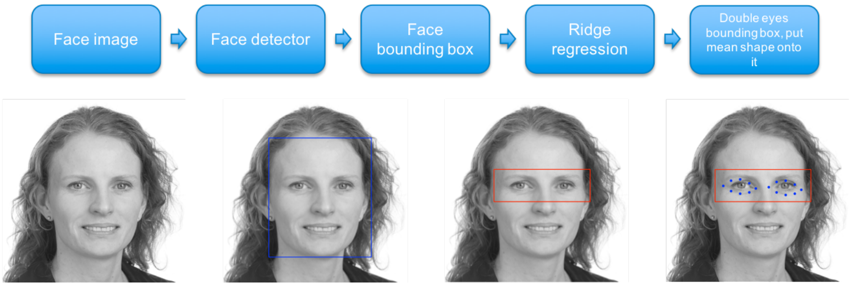


Figure 3 eye initialization generation.

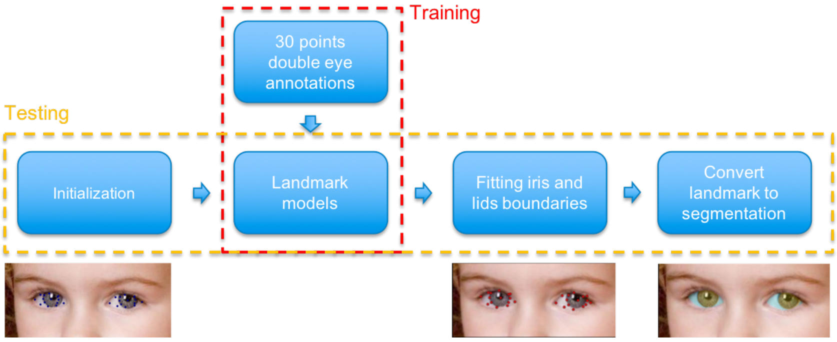

Training
30 points double eye annotations
Testing
Initialization
Landmark models
Fitting iris and lids boundaries
Convert landmark to segmentation

Figure 4 Flowchart of landmark models methodology.

Since deformable model-based methods are sensitive to initialization, the accuracy of initializations affects algorithms' performances. To generate an appropriate initialization, the first step is eye localization. The procedure is shown in Figure 3. Firstly, face detection method (fast RCNN[6]) was utilized to obtain the bounding box of each faces. Secondly, a ridge regression model was trained to predict eyes' bounding boxes based on facial bounding boxes. Finally, landmark models can use mean shape as initialization lying into above located bounding boxes. The procedure flowchart of deformable model-based method is shown in Figure 4.

4.2 Deep Segmentation Method

Atrous convolutional neural network (ACNN) effectively enlarge the field of view of filters without increasing the number of parameters or the amount of computation[2]. The method adds 'holes' to the convolution filter mask to better model the relationship between distant pixels. With our dataset, we fine-tune the ACNN network to produce the initial per-pixel probability score map from the input eye region RGB image. Note that at this stage, the entire eye region (which contains both left and right eyes) are fed to the ACNN. Because the shape and orientation of left and right eyes are highly correlated, the correlation can be learned and exploited by ACNN to produce good score map for both eyes when face is not in a frontal position so that one eye is more blurry/smaller than the other. Since the boundaries of the sclera and the iris were too blur to accurately be segmented, a CRF was utilized at the end of ACNN as post-processing in order to sharp boundaries. The procedure of the experiment is as Figure 5.

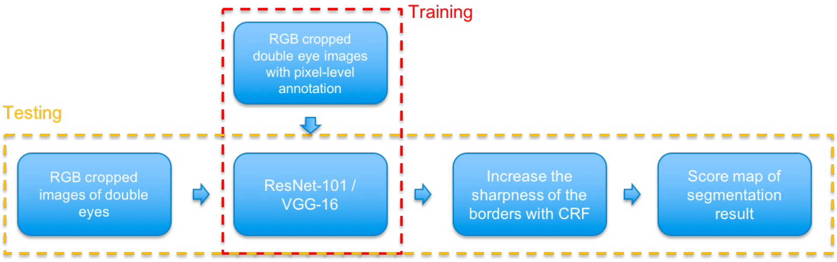
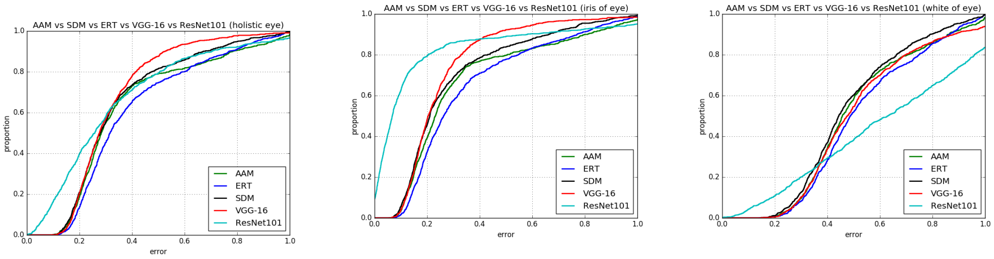

Training
RGB cropped double eye images with pixel-level annotation
Testing
RGB cropped images of double eyes
ResNet-101 / VGG-16
Increase the sharpness of the borders with CRF
Score map of segmentation result

Figure 5 Flowchart of deep model methodology.



(a) Holistic eye.

(b) Iris of eye.

(c) Sclera of eye.

Figure 6 Evaluation of all models.

5 Experiments

In this section, we evaluated deformable model-based methods and deep models according to two criteria: (a) the performance of holistic, iris-only and sclera-only segmentation; (b) robustness through comparing performances of frontal faces and non-frontal faces. Performances of landmark and deep models are evaluated in Figure 6. It is obvious that performances of deep models on holistic eye and iris-only segmentation were better than deformable models. For VGG-16 and ResNet101, ResNet101 performed better than VGG16, since ResNet101 has larger size of architecture in order to gain wispy features of the eye. On the other hand, in sclera-only segmentation, the performance of ResNet101 was relatively worse than other methods, because the dataset we built was not big enough for large-scale ResNet101. Meanwhile, the overfitting was happened during ResNet101 training.

According to Figure 7 in appendix, the segmentation of profile faces is worse than frontal faces. Even so, deep segmentation models still got higher performances than deformable models on profile faces. On another aspect, the accuracy reduction of deep models is milder than deformable methods, which means that deep models are more robust than deformable methods under pose variation. Theoretically, for non-frontal faces, the face shape and texture need to be transformed further than frontal faces during predictions of deformable methods, so that it is difficult to find optimized local minimum from image intensity space. With inaccuracy initializations, non-frontal faces are still challenges for landmark tracking. On the other hand, the baseline methods we utilized in this research aims to evaluate the availability of our dataset, meanwhile, the performance of experiments provide a preliminary research on low-resolution eye segmentation. Although methods above are widely-used and may not be state-of-art currently, it is enough for us to present the effectiveness of our dataset. Therefore, this research indicates: (a) eye segmentation research can reasonably work on our dataset; (b) deep models are more potential for eye segmentation compared with deformable model-based methods.

6 Conclusion

In conclusion, there are two contributions in this research. Firstly, we proposed a new dataset for low-resolution eye segmentation. Our dataset provides two types of ground truths: 30-point landmarks and pixel-level ground truth. In terms of contents, the dataset contains frontal and non-frontal faces in low resolution of eye region, under variant illuminations and with/without glasses. Secondly, in order to evaluate the usability of our dataset and provide a preliminary eye segmentation investigation on low-resolution eye segmentation, we applied deformable model-based methods (AAM, SDM and ERT) and deep semantic segmentation models (VGG16+CRF and ResNet101+CRF) as baseline methods. According to the ROC curves of IOU accuracy, deep models got a better robustness than deformable methods. Moreover, especially for non-frontal faces, performances of deep models can adapt head poses variation. Otherwise, our dataset can be utilized for iris-only and sclera-only segmentation. Based on experiments, deep models got better performances on our dataset as well. Therefore, this research indicates that researchers can put more efforts to use deep segmentation methods instead of deformable model-based methods in eye segmentation task. Otherwise, existed models did not consider the shape refinement and shape prior of the eye, thus in future researches plugging in shape prior and post-processing shape model can extremely improve segmentation performance.

References

1 Michael Chau and Margrit Betke. Real time eye tracking and blink detection with usb cameras. Technical report, Boston University Computer Science Department, 2005.

2 Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

3 Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pages 681–685, 2001.

4 Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

5 Bruce M Hood, J Douglas Willen, and Jon Driver. Adult's eyes trigger shifts of visual attention in human infants. *Psychological Science*, 9(2):131–134, 1998.

6 Huaizu Jiang and Erik Learned-Miller. Face detection with the faster R-CNN. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 650–657. IEEE, 2017.

7 Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014.

8 Marc Lalonde, David Byrns, Langis Gagnon, Normand Teasdale, and Denis Laurendeau. Real-time eye blink detection with GPU-based SIFT tracking. In *Computer and Robot Vision, 2007. CRV'07. Fourth Canadian Conference on*, pages 481–487. IEEE, 2007.

9 Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *European conference on computer vision*, pages 679–692. Springer, 2012.

10 Yuezun Li, Ming-Ching Chang, Hany Farid, and Siwei Lyu. In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking. *arXiv preprint arXiv:1806.02877*, 2018.

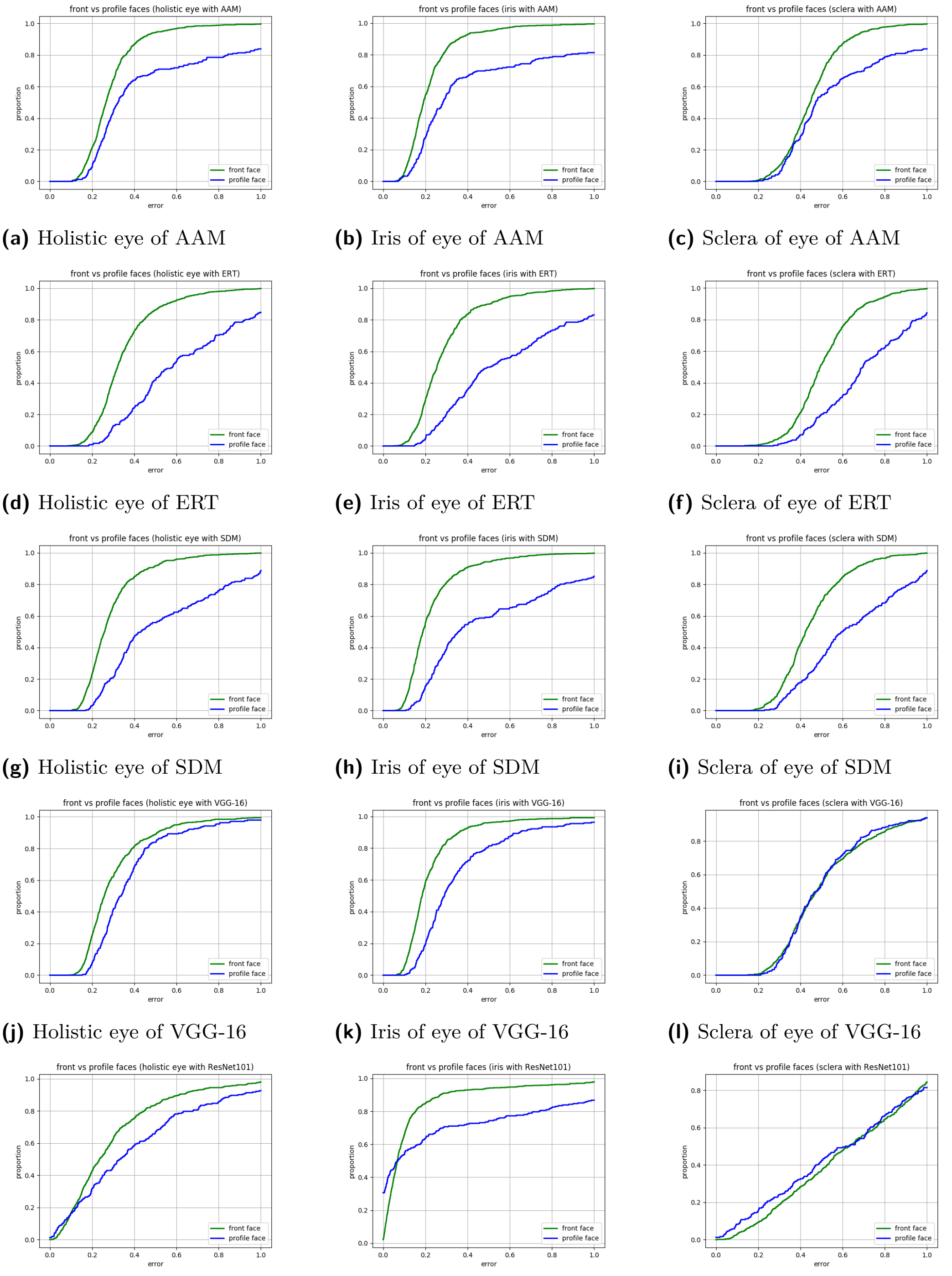
11 Peter Peer. Cvl face database. *Computer vision lab., faculty of computer and information science, University of Ljubljana, Slovenia. Available at http://www. lrv. fri. uni-lj. si/facedb. html*, 2005.

12 Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 50–58, 2015.

13 Brian A Smith, Qi Yin, Steven K Feiner, and Shree K Nayar. Gaze locking: passive eye contact detection for human-object interaction. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, pages 271–280. ACM, 2013.

14 Qi-Chuan Tian, Quan Pan, Yong-Mei Cheng, and Quan-Xue Gao. Fast algorithm and application of hough transform in iris segmentation. In *Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on*, volume 7, pages 3977–3980. IEEE, 2004.

15 Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013.



(a) Holistic eye of AAM

(b) Iris of eye of AAM

(c) Sclera of eye of AAM

(d) Holistic eye of ERT

(e) Iris of eye of ERT

(f) Sclera of eye of ERT

(g) Holistic eye of SDM

(h) Iris of eye of SDM

(i) Sclera of eye of SDM

(j) Holistic eye of VGG-16

(k) Iris of eye of VGG-16

(l) Sclera of eye of VGG-16

(m) Holistic eye of ResNet101

(n) Iris of eye of ResNet101

(o) Sclera of eye of ResNet101

Figure 7 Appendix: Robustness evaluation compared between profile and frontal faces.