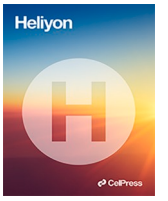
Heliyon 10 (2024) e34308

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Automated analysis for glaucoma screening of retinal videos acquired with smartphone-based ophthalmoscope

Fabio Scarpa [a,*], Alexa Berto [a,b], Nikos Tsiknakis [c], Georgios Manikis [c], Dimitrios I. Fotiadis [d,e], Kostas Marias [c,f], Alberto Scarpa [b]

[a] Department of Information Engineering, University of Padova, Padova, 35131, Italy
[b] D-Eye Srl, Padova, 35131, Italy
[c] Institute of Computer Science, Foundation for Research and Technology Hellas (FORTH), Heraklion, 70013, Greece
[d] Department of Biomedical Research, Institute of Molecular Biology and Biotechnology, FORTH, Ioannina, 45115, Greece
[e] Department of Materials Science and Engineering, Unit of Medical Technology and Intelligent Information Systems, University of Ioannina, Ioannina, 45110, Greece
[f] Department of Electrical and Computer Engineering, Hellenic Mediterranean University, Heraklion, 71004, Greece

ARTICLE INFO

Keywords:
Glaucoma
Smartphone ophthalmoscopy
Optic disc segmentation
Optic cup segmentation

ABSTRACT

Widespread screening is crucial for the early diagnosis and treatment of glaucoma, the leading cause of visual impairment and blindness. The development of portable technologies, such as smartphone-based ophthalmoscopes, able to image the optical nerve head, represents a resource for large-scale glaucoma screening. Indeed, they consist of an optical device attached to a common smartphone, making the overall device cheap and easy to use. Automated analyses able to assist clinicians are crucial for fast, reproducible, and accurate screening, and can promote its diffusion making it possible even for non-expert ophthalmologists. Images acquired with smartphone ophthalmoscopes differ from that acquired with a fundus camera for the field of view, noise, colour, and the presence of pupil, iris and eyelid. Consequently, algorithms specifically designed for this type of image need to be developed.

We propose a completely automated analysis of retinal video acquired with smartphone ophthalmoscopy. The proposed algorithm, based on convolutional neural networks, selects the most relevant frames in the video, segments both optic disc and cup, and computes the cup-to-disc ratio. The developed networks were partially trained on images from a publicly available fundus camera datasets, modified through an original procedure to be statistically equal to the ones acquired with a smartphone ophthalmoscope. The proposed algorithm achieves good results in images acquired from healthy and pathological subjects. Indeed, an accuracy $\geq$95 % was obtained for both disc and cup segmentation and the computed cup-to-disc ratios denote good agreement with manual analysis (mean difference 9 %), allowing a substantial differentiation between healthy and pathological subjects.

1. Introduction

Glaucoma is the third leading cause of irreversible blindness and a leading cause of visual impairment worldwide. It is

* Corresponding author. Via Gradenigo 6/b, Padova, 35131, Italy.
E-mail address: fabio.scarpa@unipd.it (F. Scarpa).

https://doi.org/10.1016/j.heliyon.2024.e34308
Received 7 December 2023; Received in revised form 12 June 2024; Accepted 8 July 2024
Available online 9 July 2024

characterized by optic nerve damage and visual field loss. Glaucoma affects about 80 million people worldwide. The disease remains largely asymptomatic as it progresses so that >50 % of individuals are unaware of diagnosis until glaucoma reaches advanced stages [1]. Widespread screening is therefore critical for early diagnosis, treatment, and limiting the incidence of glaucoma-associated blindness.

The pervasive diffusion of smartphones represents a resource for glaucoma screening. Indeed, smartphones can be reliably used as ophthalmoscopes with the help of cheap and portable optical devices [2–6]. Images of the optic nerve head can be easily acquired and subjected to ophthalmoscopic examination, which is crucial in the diagnosis and management of glaucomatous patients (Fig. 1). Particularly, the vertical cup-to-disc ratio (VCDR) is an important index in the screening and follow-up of patients with glaucoma [7] and has been found to correlate with visual field indexes [8]. The quality of the images acquired with smartphone ophthalmoscopy is lower than the one of images acquired with conventional fundus cameras. However, previous studies demonstrated that expert ophthalmologists can compute VCDR, and thus screen for glaucoma, from images acquired with smartphone ophthalmoscopy [9].

There has been extensive growth in recent years in the development of algorithms for the automated assessment of glaucoma from fundus cameras. Early attempts for glaucoma detection were mostly based on thresholding, pre-determined shape matching (e.g. Hough transform) combined with active contours [10,11] or, more recently, on handcrafted combinations of feature extraction techniques and supervised or unsupervised machine learning classifiers [12]. However, their accuracy was limited due to the application of manually designed features, which are unable to comprehensively characterize the large variability of the disease appearance. Deep Learning techniques, on the contrary, automatically learn these characteristics by exploiting the implicit information of large training sets of annotated images outperforming pre-existing algorithms in the medical field [13] and ophthalmology [14–17].

In the last years, several neural network architectures have been proposed for glaucoma classification and optic disc (OD) and optic cup (OC) segmentation with performance comparable to that of an experienced clinician [18–20]. Most of the recent methods are based on modifications to the original u-shaped convolutional neural network (U-Net) [21]. This is because U-Net can achieve good results even when trained using a relatively small number of images. However, these methods provide excellent results on images acquired with fundus cameras, which have higher resolution, contrast, and quality with respect to that acquired with smartphone ophthalmoscopes. In addition, the field of view of smartphone ophthalmoscopes is very limited (about 20°) with respect to the one of fundus camera (from 35 to 150°). Furthermore, a consistent portion of the typical smartphone ophthalmoscope image depicts pupil, iris, eyelid, or eyelashes that represent noise, which does not appear in conventional fundus camera images. Therefore, specific algorithms need to be developed for the analysis of smartphone ophthalmoscopy images. Some works have been recently proposed with interesting results, but they require the manual selection of a frame from the acquired video [22,23].

In this work, we propose a completely automated analysis of retinal fundus videos, able to select the most relevant frames, to segment both optic disc and cup and thus compute clinically useful indexes for glaucoma diagnosis, such as VCDR. The visual (OD and OC contours) and quantitative (clinical parameters) results can be used by clinicians to improve glaucoma screening. The proposed method is based on u-shaped convolutional neural networks, a deep learning technique.

The main contribution of this paper is the development of a completely automated procedure specifically designed for the analysis of retinal video acquired with smartphone ophthalmoscopy, and the demonstration that the derived clinical indices can be useful for glaucoma screening. The automated analysis of these videos could be a pivotal boost for the widespread screening for glaucoma, since it could support the use of smartphone-based ophthalmoscopes (i.e., cheap and portable ophthalmoscopic devices), even by non-expert

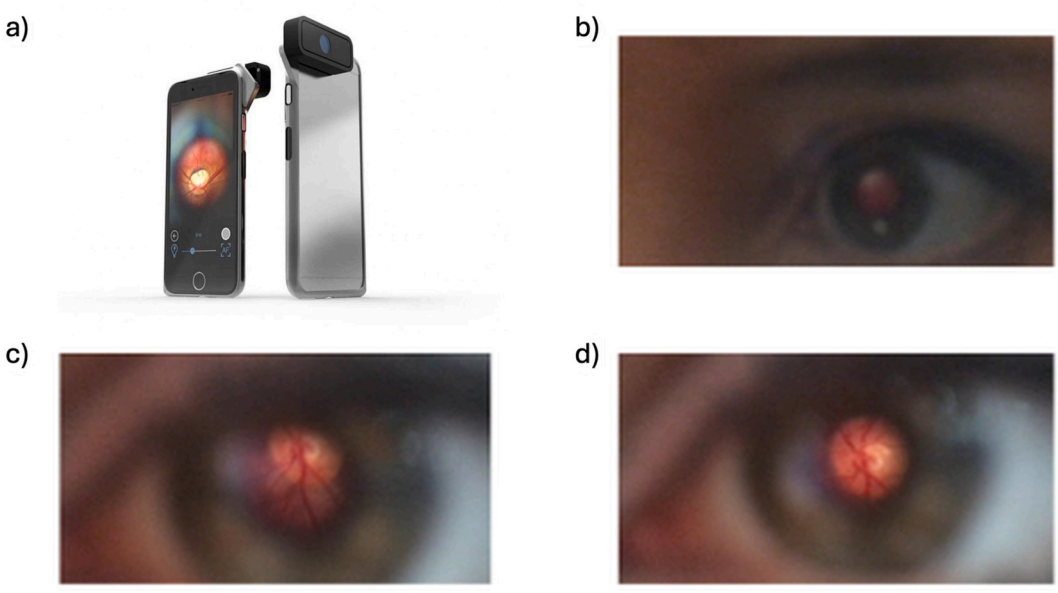


Fig. 1. Acquisition system (a) and representative examples of acquired frames (b, c, d).

ophthalmologists (e.g., general doctors, paramedics, etc.) in a telemedicine setting [5,6]. This might be especially crucial for individuals living in rural or remote areas, who have limited access to optometrists or ophthalmologists, hence to glaucoma standard tests. In addition, the procedure implemented to obtain images statistically equal to the ones under analysis can be easily adapted to other applications, in which images are acquired by different devices.

2. Material

In this study, videos acquired with the D-Eye adapter (D-Eye S.r.l., Padova, Italy) attached to iPhone 5 and 6 (Apple Inc., Cupertino, CA) have been considered. Fig. 1a shows the acquisition system. Each video has a duration of about 30 s, 30 frames per second and a resolution of 1080 $\times$ 1920 pixels.

2.1. Dataset for frames selection and OD segmentation

A dataset composed of 1686 images manually selected from 216 videos acquired with the D-EYE device was considered for Frames Selection (section III.B) and OD segmentation (section III.C). On these images, pixels corresponding to the background, retina (i.e., region of interest) or optic disc were manually identified, if present. In addition, annotations consisting of the x, y (center coordinates), width and height of retina and disc minimum bounding boxes, if present, were derived from the manual segmentation.

It is well known that CNNs need a large dataset in the training process. The dataset for training was quadrupled through data augmentation. In particular, we combined random rotation and translation ranging respectively from 0° to 5° and from 0 % to 30 % of image size, and horizontal flipping of the image. All the transformations were performed preserving the presence of the retina and OD in the image. Indeed, if the center of the object (retina or disc) in the transformed image gone outside the image size, the transformation was redone.

2.2. Dataset for OC segmentation

Images from publicly available datasets were considered for optic cup segmentation. 750 images were from RIGA dataset [24] and 800 images were from REFUGE dataset [18]. In these images, the optic disc and cup were manually segmented by expert ophthalmologists and used as ground truth during the training and validation process. Original fundus images, acquired with traditional fundus cameras, were processed to make them similar to the images acquired with the D-Eye device (III.A). We doubled this dataset by flipping the images horizontally. Finally, the manually segmented optic disc is used to blacken the images outside the optic disc. Classes are background, disc and cup.

2.3. Dataset for algorithm evaluation

Videos acquired from 15 healthy subjects and from 15 subjects affected by glaucoma were used to evaluate the performance of the overall algorithm, testing both the frame selection and segmentation accuracy, and consequently the reliability of VCDRs estimation and glaucoma detection.

3. Methods

This section describes the procedure that we implemented for dataset adjustment (A) and all the steps of the algorithm that we developed for the automated analysis of a video acquired with the D-Eye device (III.B and III.C). The algorithm performs the frames selection, the segmentation of OD and OC on the selected frames, and finally, the computation of clinical indexes, providing visual (OD and OC contours on the most meaningful frames) and quantitative information (e.g. VCDR) useful to assist clinicians in glaucoma screening. The algorithm was implemented in Python using the Tensorflow framework, and the analysis of different frames is performed in parallel on the available cores. The analysis of a single frame requires about 1 s and the analysis of a complete video requires about 2 min running on a laptop equipped with a processor Intel i5 with 4 cores (Intel, Santa Clara, California, USA).

The following paragraphs describe the network architectures that we developed and tested for frames selection, and OD and OC segmentation.

3.1. D-Eye like tool

Since it is not available OC manual annotation by expert ophthalmologists for a large set of images acquired with the D-Eye device, we decided on using public datasets composed of fundus camera images (II.B). However, these images look very different from D-Eye images, not only about the different fields of view but also about magnitude, focus, noise, and colour. Thus, we have implemented a procedure that automatically modifies fundus camera images to make them statistically and visually similar to D-Eye images.

We computed various statistical indexes from 1000 images acquired with a fundus camera and 1000 images acquired with D-Eye device. For both sets of images, the OD region derived by manual annotation was considered. We selected the indexes that best described the differences between images acquired with the two modalities: OD diameter to quantify the magnification, standard deviation for noise, saturation and value (from HSV colour space [25]) for colour, and the gradient energy for focus. The gradient energy, GRAE [26], represents the degree of focus of an image. It is computed as the squared mean intensity difference between

adjacent pixels, along both rows and columns. We chose GRAE between all possible focus measures because it has a simple mathematical formulation and is easy to compute. As equation (1) shows, it considers the differences between adjacent pixels, both vertically and horizontally.

$$\mathrm{GRAE} = \left(\left[\frac{\sum_{i}^{n^\circ \text{ cols}} \left(I_{x_i} - I_{x_{i-1}} \right)}{n^\circ \text{ cols}} \right]^2 + \left[\frac{\sum_{i}^{n^\circ \text{ rows}} \left(I_{y_i} - I_{y_{i-1}} \right)}{n^\circ \text{ rows}} \right]^2 \right) \Big/ 2 \tag{1}$$

All indexes were derived only by the OD pixels and were chosen because the probability distributions (histograms) computed by the two different sets of images were slightly different, thus well describing the differences in the two types of images.

The developed tool consists of a multi-step procedure, in which each step modifies a property of the image described by one of the previous indexes, to be consistent with D-Eye images. Indeed, for the considered index, a random value inside the probability distribution of D-Eye images is selected, and the fundus camera image is opportunely processed as to finally have the selected value of that index. Therefore, the probability distributions of the indexes obtained by the modified images overlap the corresponding distributions of the indexes derived by images from D-Eye. In addition, the tool also compensates for the different Fields of View (FoV) and the presence of the iris in D-Eye images.

Fig. 2 summarizes the main steps to produce D-Eye-like images. For each fundus camera image (Fig. 2a), the software identifies the OD and performs a resizing to bring the diameter of the OD in the range of values computed by D-Eye images (between 360 and 440 pixels, uniformly distributed). In particular, the algorithm detects the OD bounding box from the mask image and if the OD manual segmentation is not available, it tries to automatically segment the OD using a U-Net specifically pre-trained for this task. If the OD is not found, the image is discarded from further analyses.

A crop of 640 × 640 pixels centered in the OD is got (Fig. 2b). Then, the tool adjusts the focus of the resized and cropped image. Among the various focus measures, we chose the GRAE because it well highlights the differences between the two types of images and, moreover, it is possible to obtain a new image having a desired GRAE by simply modifying the size of a Gaussian filter used to blur the image itself (the greater the filter size is, the more the image is blurred) (Fig. 2c). After that, we considered the HSV colour space.

Since the saturation and value of D-Eye images are different from that of fundus camera images, D-Eye-like tool modifies these indexes, reducing them to values comparable to that of D-Eye images (Fig. 2d). Then, to reproduce the typical noise that affects D-Eye images, the tool adds Gaussian noise to the processed image. Therefore, the standard deviation computed on small regions assumes values equal to those of a D-Eye image (Fig. 2e). At this point, to reproduce the D-Eye images FoV, a white-ellipse with a size similar to that of D-Eye FoV was drawn in a black image. In addition, a gaussian filter was applied to smooth the contour of the ellipse (Fig. 2f). This image and the one obtained in the previous step are multiplied (Fig. 2g). Finally, the code generates a synthetic iris. The creation starts by drawing filled ellipses in the image corners. A colour between brown, cyan and light green is randomly selected and used to colour the ellipses. A Gaussian filter is then applied to smooth the four ellipses (Fig. 2h). This synthetic iris and the image obtained at the previous step are summed, obtaining the final image (Fig. 2i).

Fig. 3 shows the GRAE histograms of the original images, of the D-Eye images and the new modified D-Eye like images. Good overlap between distributions obtained by D-Eye images and by the new images can be observed. Similar considerations and visual representations (useful when the input consists of a big dataset) were obtained for all the considered image properties. Analogue results were obtained for Saturation and Value distributions (Fig. 3).

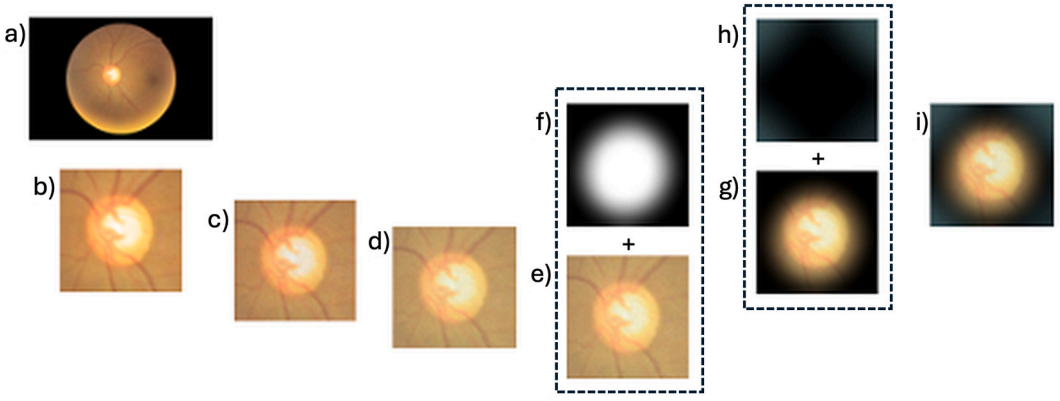


Fig. 2. Main steps of the algorithm to generate D-Eye-like images. (a) Original image. (b) Resize and crop based on OD. (c) Blurring based on GRAE. (d) Colour variation. (e) Gaussian noise. (f) FoV simulation. (g) Addition of (e) and (f). (h) Synthetic iris cre creation. (i) D-Eye-like final image obtained by adding (g) and (h).

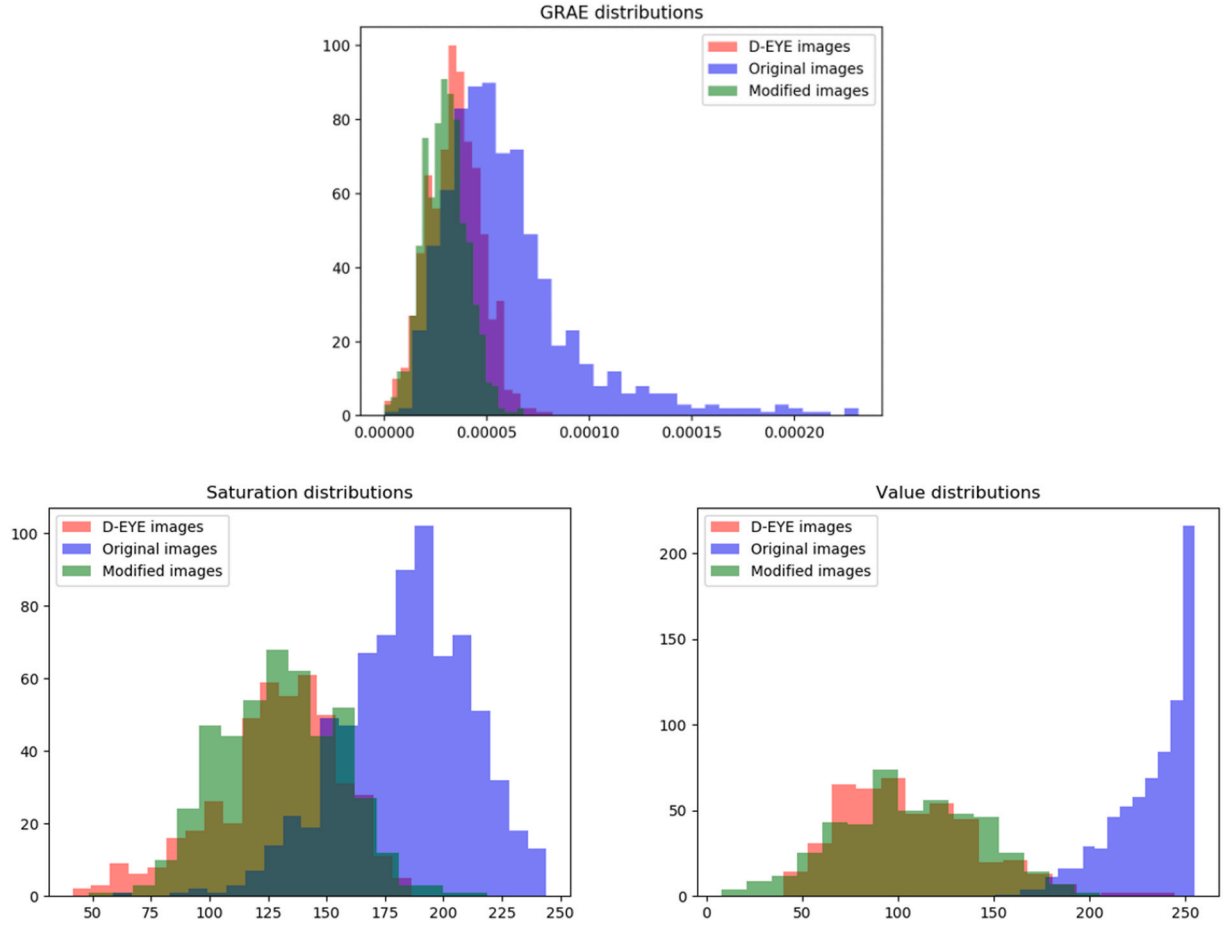


Fig. 3. GRAE, saturation and value histograms of the images.

3.2. Frames selection

The proposed algorithm analyzes all the frames of a video. Considering the horizontal orientation as a reference, vertical frames are clockwise rotated 90°. As a result, every frame consists of an RGB image with a size of 1080 × 1920 pixels (n° of rows x n° of columns). The frame is resized to 270 × 480 pixels using a linear transformation function. A CNN receives (describe in III.B.1) the resized image in input and provides an output information about the presence of an object (the retina and the OD) through a number ranging from 0 to 1 (0 = object not present, 1 = object present), the x and y coordinates of the center, normalized between 0 and 1 (0 = upper/left margin, 1 = bottom/right margin of the image) and width and height of the minimum bounding rectangle containing the retina or the OD. The algorithm selects the frames in which the OD is fully contained in the retina and, among these frames, the 10 frames with the highest values of GRAE. If less than ten frames are selected, the video is not further analyzed. On the selected frames, the normalized coordinates of the region containing retina and OD are reported to the original image size, and the image is cropped with a box of 640 × 640 pixels size centered in that region.

1) CNN for retina and optic disc detection

We tried two different architectures for this CNN. The main difference between the two architectures is that one provides the output as a single array with 5 elements (presence, x, y, width, and height) for both ROI and OD, while the other splits the output into two different arrays, the first containing an element corresponding to the presence and the second containing four elements corresponding to x and y coordinates of the center, and width and height of the minimum bounding rectangle, for both ROI and OD. Based on this difference, the CNNs have been named singleOutput and splitOutput respectively.

The same two architectures of the previous step were used. In this case, the singleOutput and splitOutput models receive as input an RGB cropped image, resized from 640 × 640 to 320 × 320. The output is a vector of 10 elements, corresponding to 5 numbers (presence, x and y center coordinates, width, and height of the bounding box) for both the retina and disc.

2) singleOutput CNN

The architecture of the singleOutput CNN consists of the concatenation of four convolutional blocks (a sequence of a 3 × 3 convolution layer, a batch-normalization layer and a 2 × 2 max-pooling layer), followed by a flatten layer and a sequence of three dense layers. This is a simplified version of a standard network for object detection, e.g., Yolo, because we have only one object to detect (the retina, if present). In order to prevent overfitting, we added a dropout layer before and after the fourth convolutional unit. We choose the number of filters and nodes as a trade-off between the necessity to describe the high variability of our data (- > high number of trainable parameters) and to reduce the computation cost of the CNN (- > low number of trainable parameters): eventually, we set 32, 64, 128, 128 filters respectively from the first to the last convolutional layer and 512, 32, 10 nodes in the dense layers. We set rectified linear unit (ReLU) activation function in each convolutional unit, the linear activation function in the first two dense layers, and the sigmoid activation function in the last dense layer. This last dense layer provides the output array made of 10 elements, 5 for the retina and 5 for OD. For both retina and OD, the first element is related to the presence, and the others to the x and y coordinates of the center, width and height of the bounding box, each ranging from 0 to 1 as expected of the sigmoid function.

3) splitOutput CNN

The first part of the splitOutput CNN consists of a main branch made, as the previous CNN, by the concatenation of four convolutional blocks with two dropout layers, followed by a flatten layer. This main branch is then split into two branches, composed of three dense layers, providing each a different output. In this CNN, considering again a trade-off between the necessity to describe the high variability of our data and to reduce the computation cost of the CNN, we tested a different number of filters compared to the previous case. We set 32, 64, 256, and 32 respectively from the first to the last 3 × 3 convolutional layer, and 512, 128 in the first two dense layers of each branch. We set 1 and 4 nodes respectively in the last dense layer of the two branches, for retina and OD. Indeed, the first branch provides an array of one number related to the presence of the object in the input image, while the second branch provides an array containing four numbers, related respectively to the x and y coordinates of the center, width, and height of the bounding box. As in the previous CNN, we set rectified linear unit (ReLU) activation function in every convolutional unit, linear activation function in the first two dense layers, and sigmoid activation function in the last dense layers.

4) Custom Loss Function

For both singleOutput and splitOutput CNN, we implemented a custom loss function used in the training process. The identification of the presence of the object and the localization of the center point and the size of the bounding box are very different problems. The presence of the object needs to be checked in every image (because not all the images contain retina and/or OD) and it is a binary decision (present/not present). The localization of the center point and size of the bounding box occurs only in images containing the retina or disc, and it consists of the estimation of four parameters: x and y coordinates, width, and height. Thus, standard loss functions based on accuracy or least squares would not lead to good results. Hence, it has been implemented a loss function that consists of customization, adapted to our specific case, of the typical loss function employed for object detection. In particular, the loss function is made by the sum of two parts, the first being the mean square error between the true (p) and the predicted ($\widehat{p}$) presence of the Retina or OD in the N images in a batch, as shown in

$$presence = \sum_{i=1}^{N} \left[(p_i - \widehat{p_i})^2 \right] \Big/ N \tag{2}$$

The second part is based on the mean squared error between the true (x, y) and the predicted ($\widehat{x}$, $\widehat{y}$) coordinates (and analogously for width and height) of the object (ROI or OD) computed only if the object is present in the image $\left(1_{object}\right)$, as shown in

$$center = \sum_{i=1}^{N} 1_{ROI} \left[(x_i - \widehat{x_i})^2 + (y_i - \widehat{y_i})^2 \right] \Big/ N \tag{3}$$

$$size = \sum_{i=1}^{N} 1_{ROI} \left[(w_i - \widehat{w_i})^2 + (h_i - \widehat{h_i})^2 \right] \Big/ N \tag{4}$$

For the singleOutput CNN, the final loss value is derived by a weighted mean of the terms. For the splitOutput CNN, each term corresponds to the loss of an output branch, but a global loss is still computed as a weighted mean of the branches. Particular attention was paid to the coefficients of this weighted average, but no significant differences were found between the various tested values. Thus, the same weight (i.e., 0.5) can be assigned to each term.

3.3. Optic disc and optic cup segmentation

The frames selected in the previous step are further processed to segment both Optic Disc and Optic Cup. OD and OC segmentation are performed by two U-Net, which have the same architecture but have been trained on a different dataset. Indeed, the first has been trained on the dataset described in section II.A, made by images acquired with the D-Eye device and by the corresponding manual annotation related to OD contour. The second U-Net has been trained on the dataset described in section II.B, composed of images from public datasets and processed with the D-Eye-like tool (described in III.A), which are similar to real D-Eye images inside the OD and

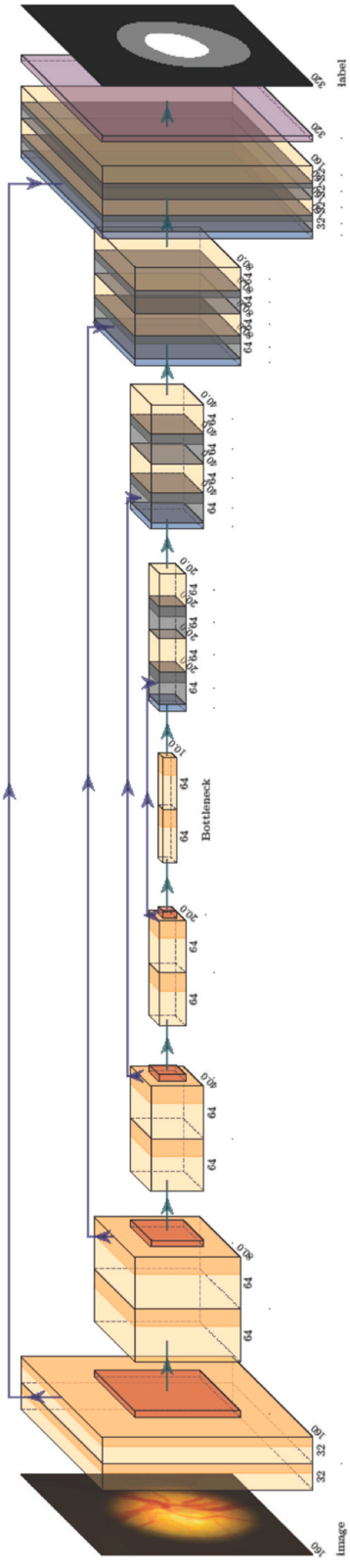


Fig. 4. U-NET architecture.

have manual segmentation of OC. To this U-net, input images were blackened outside the OD. For both architectures, the output is an image with three possible classes assigned to each pixel: background, retina, OD for the first U-Net, and background, OD and OC for the second U-Net.

1) U-Net architecture

The U-Net is made of a contracting-encoder part and an expanding-decoder part. The first part is dedicated to feature extraction, while the latter allows obtaining a label classification on every single pixel. It was originally developed by Ronneberger [21] for fast and precise segmentation of biomedical images and at the moment represents a gold standard for segmentation tasks, also in the retinal image for glaucoma detection. Two U-NET architectures (described in III.C.2) and III.C.3) were implemented and tested, both with 5 blocks in the contracting-encoder part and the expanding-decoder part. Fig. 4 shows the common structure of the two architectures.

2) Tiny U-Net

Each block of the contracting part consists of two convolutional units (3×3 convolution, batch normalization, rectified linear unit) followed by a 2×2 max-pooling layer (to down-sample the image). 32, 64, 64, 64, and 64 filters respectively from the first to the last 3×3 convolutional layer were set. The blocks of the expanding part up-sample the image, concatenate it with the corresponding one in the encoder part and then apply two convolutional units (3×3 convolution, batch normalization, rectified linear unit). Similarly, to the encoding part, we set 64, 64, 64, 64, and 32 filters respectively from the first to the last layer. This architecture, characterized by a low number of filters, was successfully used for optic disc and cup segmentation on retinal images acquired with a traditional fundus camera [27].

3) Heavy U-Net

Each block of the contracting part consists of two convolutional units (3×3 convolution, rectified linear unit) followed by a 2×2 max-pooling layer (to down-sample the image). The 3×3 convolutional layer has, from the first to the last, 16, 32, 64, 128, and 256 filters respectively. The U-NET has a dropout layer before and after the last convolutional unit. The blocks of the expanding part up-sample the image, concatenate it with the corresponding one in the encoder part and then apply two convolutional units (3×3 convolution, rectified linear unit). Similarly, to the encoding part, the 3×3 convolutional layers have, from the first to the last, 256, 128, 64, 32 and 16 filters, respectively. This model consists of a widely used conventional architecture [21] that usually provides good results, but it is computationally expensive due to the high number of filters and trainable parameters.

4) Post-processing and clinical parameters estimation

A post-processing procedure is applied to the regions corresponding to the retina, optic disc and cup separately. Each region, connected component of white adjacent pixels, undergoes a sequence of morphological operations: erosion and dilation (both with a 5×5 kernel), area filtering and filling. These operations aim to remove small objects from the image while preserving the shape and size of larger objects and filling in small holes inside the object. Finally, if the region is sufficiently great (i.e., 40 pixels), the ellipse that fits (in a least-squares sense) the contour of the connected component is computed. The contour of the ellipse can be visualized on the original image for inspection of its correspondence with the real contour of the disc or cup, and clinical parameters, e.g., Vertical Cup-to-Disc Ratio (VCDR), can then be derived by cup and disc elliptical fits.

4. Results

4.1. Frames selection

The performances of the CNNs for frame selection are reported in Table 1. The number of images used for training and validation is 1518 and 168, quadrupled through data augmentation. Accuracy values are the complementary values of loss (i.e. 1-loss). The total number of epochs is the set number, or the number achieved at the EarlyStopping callback. The best epoch corresponds to the last model saved by ModelCheckpoint callback (the model checkpoints were saved every 5 epochs). As illustrated by Table 1, the splitOutput model achieves better accuracy than the singleOutput model and in a shorter time.

Table 1
CNN for retina and OD detection.

Model	Train Accuracy [%]	Validation Accuracy [%]	N° Trainable Parameters	Time for Training [hours]	Best epoch
single Output	95.0	95.0	2,898,218	56	60
split Output	99.3	96.4	10,992,106	19	25

4.2. Optic disc and cup segmentation

Table 2 reports the performances of the U-Nets for disc and cup segmentation. As described in the previous paragraphs, for both we tried two different models (III.C.2) and III.C.3)). The number of images used for training and validation is 1518 and 168 for OD and 1395 and 155 for OC, quadrupled through data augmentation. Accuracy values are the complementary values of loss. The total number of epochs was set to 200, but it stopped using the EarlyStopping callback. The best epoch corresponds to the last model saved by ModelCheckpoint callback (the model checkpoints were saved every 5 epochs). Categorical cross entropy and intersection over union (IoU) scores were both used as loss functions:

$$Categorical\ cross\ entropy = \sum_{i=1}^{N} p_i \cdot \log\left(\widehat{p}_i\right) \tag{5}$$

$$1 - IoU = 1 - \frac{\text{area of intersection}}{\text{area of union}} \tag{6}$$

They consider respectively the real (p_i) and predicted ($\widehat{p}_i$) classification of each pixel and the intersection and union between the real and the predicted areas (connected pixels belonging to the same class). The IoU increased ten times the execution time of the training process but does not provide significant improvement on the results, and thus was only computed in the validation set. Fig. 5 shows a representative example of the manual and automated optic disc and cup segmentation.

4.3. Clinical parameters estimation

The reliability of the analysis provided by the algorithm from a video has been evaluated on 30 videos acquired with D-Eye, from 15 healthy subjects and 15 subjects affected by glaucoma. The final evaluation is derived for each subject, considering the median VCDRs obtained by the ten automatically selected frames. The obtained VCDRs well correlate with that obtained manually by expert clinicians. Indeed, the Pearson correlation coefficient was 0.93 and the mean difference between the manual and automatic VCDR estimations was 9.1 %. Automated VCDRs were 0.44 $\pm$ 0.027 (mean $\pm$ standard deviation) for healthy subjects and 0.65 $\pm$ 0.032 for subjects with glaucoma, denoting a substantial difference between the two groups of subjects. Indeed, 28 subjects were correctly classified, while 2 healthy subjects were erroneously classified as pathological (accuracy 0.93, sensitivity 1, and specificity 0.88). Results obtained in our limited dataset need further investigation on a larger dataset, and the computation and analysis of other quantitative clinical indices (e.g., horizontal cup-to-disc ratio, neuro-retinal rim area, etc.) is recommended. However, the obtained results highlight the capability of the proposed methods in assisting in glaucoma screening.

We also considered, for 4 healthy subjects, two videos acquired from the same subject by different operator. The automated analysis was consistent, providing VCDR values that have a maximum difference of 0.7 % between the two videos from the same subject. The robustness of the developed method along different videos from the same subject is mainly due to the best frames selection procedure, that is able to select the frames in which the optic disc is correctly depicted. Analysing the best 10 frames per videos, with the final VCDR derived from the median of the 10 values, also reduce the variability in VCDR predictions due to differences in video acquisition.

Vertical cup-to-disc ratios derived both manually and automatically by the video of our dataset appear to be overestimated. This effect was also noticed in Ref. [9] and is probably due to an oversaturation of the acquired image around the cup (much brighter compared with the rim), making it appear larger.

5. Discussion

A reliable automated analysis of fundus videos might be a pivotal boost for the widespread screening for glaucoma. It could support the use of smartphone-based ophthalmoscopy (i.e. cheap and portable ophthalmoscopic devices) even by non-expert ophthalmologists (e.g. general doctors, paramedics, etc.) in a telemedicine setting [5,6]. Indeed, the automatic analysis can give immediate information about the quality of the acquired video, can provide preliminary clinical information, and can allow the remote analysis of a few frames instead of the entire video. This might be especially crucial for individuals living in rural or remote areas. Automated analyses proposed in literature can be applied to images acquired with conventional fundus camera instead of a smartphone. The use of a common,

Table 2
U-net for OD and OC segmentation.

Model object	Train Accuracy [%]	Validation Accuracy [%]	IoU score	Trainable Params	Training Time [hours]	Best epoch
Tiny Disc	96,0	94,3	0.621	806,083	18	130
Heavy Disc	95,3	94,3	0.616	1,941,139	101	105
Tiny Cup	95	94.7	0.630	806,083	5	50
Heavy Cup	94.8	93.3	0.603	1,941,139	108	170

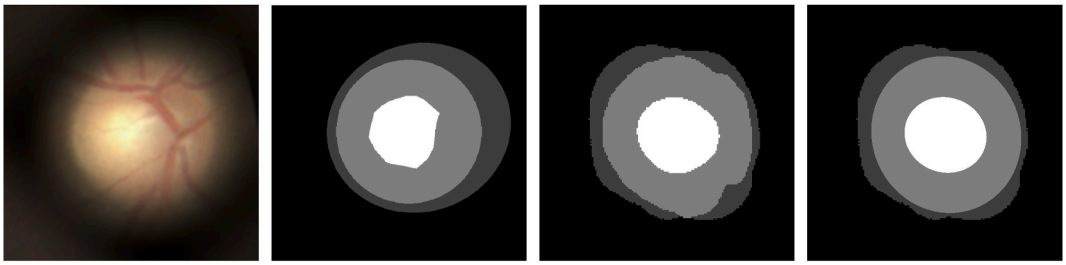


Fig. 5. From left to right: original image, manual segmentation, automated segmentation, post-processing.

portable and chip device like the smartphone can greatly improve the diffusion of glaucoma screening. Automated methods specifically designed for smartphone ophthalmoscopy still require some manual processing, for example the manual selection of a frame from the acquired video [22,23].

The proposed completely automated analysis of fundus video acquired with smartphone ophthalmoscopy, selects the most relevant frames, segments both optic disc and cup, computes the cup-to-disc ratio, and thus provides information useful for clinicians in the glaucoma diagnostic process. Good results were obtained in each main step of the algorithm. Indeed, good accuracy was obtained in the detection and localization of the region of interest and consequently in frame selection. High agreement between manual and automatic segmentation of optic disc and cup was also obtained. The network for OC segmentation was trained on fundus images modified through the D-Eye like tool. The good results obtained for OC segmentation in images acquired with the D-Eye device denote the reliability of the developed tool. Finally, in our limited test set, the automatically computed VCDRs denoted a consistent difference between healthy subjects and subjects with glaucoma.

The developed method derives a quantitative analysis from retinal videos, that can assist clinicians in glaucoma screening. The derived analysis is fast, reproducible, and completely automated. The proposed algorithm improves the ones presented in literatures, combining the analysis of each frame, to select the best ones, with the segmentation of retinal structures and the estimation of clinical parameters.

The proposed analysis can also provide real-time feedback on the quality of the acquired video, suggesting the clinician to re-acquire if no good frames have been found. Enhancements are also provided in a telemedicine context, where only the selected best frames can be streamed over the internet instead of the entire video. Further development of the proposed method will include the evaluation of a larger dataset and the investigation of other clinical parameters, for example, the area of the neuroretinal rim of the optic nerve head. The influence of different smartphone models, that may include variability in the image intensities and features, will also be investigated. The development of similar algorithms for the detection of diabetic retinopathy, age-related macular degeneration and other retinal diseases will also be considered.

6. Conclusions

Widespread screening for glaucoma is crucial for early diagnosis, treatment, and limiting the incidence of glaucoma-associated blindness. The development of smartphone-based ophthalmoscopes might improve large-scale screening for glaucoma. Indeed, these optical devices are cheap, portable, and can easily acquire images of the optic nerve head, crucial in the diagnosis and management of glaucomatous patients.

In this paper, we proposed a fast and completely automated analysis of fundus video acquired with smartphone ophthalmoscopy. The algorithm has been specifically designed for this type of image that are quite different from the conventional images acquired with fundus cameras, for the field of view, noise, colour and the presence of pupil, iris, eyelid, and eyelashes. The proposed algorithm selects the most relevant frames, segments both optic disc and cup, and computes the cup-to-disc ratio. It provides visual (optic disc and cup contours) and quantitative information (clinical parameters such as VCDR) useful for clinicians in the glaucoma diagnostic process.

The good results obtained by the proposed algorithm demonstrate that convolutional neural networks can provide fast and reliable automated analysis of fundus video acquired with a smartphone-based ophthalmoscope. Even if the investigation on a large dataset is required, these findings encourage the use of smartphone ophthalmoscopes and the further development of the proposed method.

Ethical statement

Retinal videos were acquired in different clinical centers from healthy and pathological subjects for the European H2020 research project SeeFar, that includes this study. After being properly informed about the aim, methods, and implications of the project SeeFar, subjects signed the informed consent, approved by local Ethical Committees. Each video was anonymized eliminating any personal information. Approval from an Ethical Committee for the overall collection used in this study was not required. The authors obtained permission from the subjects to publish the images that have been included in this paper.

Data availability statement

The authors do not have permission to share the images used in this study.

CRediT authorship contribution statement

Fabio Scarpa: Writing – original draft, Supervision, Methodology, Investigation, Data curation, Conceptualization. **Alexa Berto:** Writing – original draft, Methodology, Investigation, Formal analysis. **Nikos Tsiknakis:** Writing – review & editing, Investigation. **Georgios Manikis:** Writing – review & editing, Investigation. **Dimitrios I. Fotiadis:** Writing – review & editing, Supervision. **Kostas Marias:** Writing – review & editing, Supervision. **Alberto Scarpa:** Writing – review & editing, Supervision, Project administration, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was partially supported by the European H2020 specific targeted research project SeeFar: Smart glasses for multifacEted visual loss mitigation and chronic disEase prevention indicator for healthier, saFer, and more productive workplAce foR ageing population. (H2020-SC1- DTH-2018-1, GA No 826429) (www.see-far.eu). This paper reflects only the author's view, and the Commission is not responsible for any use that may be made of the information it contains.

References

[1] H.A. Quigley, A.T. Broman, The number of people with glaucoma worldwide in 2010 and 2020, Br. J. Ophthalmol. 90 (3) (2006) 262–267, https://doi.org/10.1136/bjo.2005.081224.
[2] A. Bastawrous, H.K. Rono, I.A.T. Livingstone, et al., Development and validation of a smartphone-based visual acuity test (peek acuity) for clinical practice and community-based fieldwork, JAMA Ophthalmol 133 (8) (2015) 930–937, https://doi.org/10.1001/jamaophthalmol.2015.1468.
[3] A. Russo, F. Morescalchi, C. Costagliola, L. Delcassi, F. Semeraro, Comparison of smartphone ophthalmoscopy with slit-lamp biomicroscopy for grading diabetic retinopathy, Am. J. Ophthalmol. 159 (2) (2015) 360–364.e1, https://doi.org/10.1016/j.ajo.2014.11.008.
[4] A. Russo, F. Morescalchi, C. Costagliola, L. Delcassi, F. Semeraro, A novel device to exploit the smartphone camera for fundus photography, J Ophthalmol. 2015 (2015) 823139, https://doi.org/10.1155/2015/823139.
[5] M.K. Adam, C.J. Brady, A.M. Flowers, et al., Quality and diagnostic utility of mydriatic smartphone photography: the smartphone ophthalmoscopy reliability trial, Ophthalmic Surg Lasers Imaging Retina 46 (6) (2015) 631–637, https://doi.org/10.3928/23258160-20150610-06.
[6] E. Bifolck, A. Fink, D. Pedersen, T. Gregory, Smartphone imaging for the ophthalmic examination in primary care, JAAPA 31 (8) (2018) 34–38, https://doi.org/10.1097/01.JAA.0000541482.54611.7c.
[7] F. Arnalich-Montiel, F.J. Muñoz-Negrete, G. Rebolleda, M. Sales-Sanz, C. Cabarga, Cup-to-disc ratio: agreement between slit-lamp indirect ophthalmoscopic estimation and stratus optical coherence tomography measurement, Eye 21 (8) (2007) 1041–1049, https://doi.org/10.1038/sj.eye.6702391.
[8] P.J. Airaksinen, S.M. Drance, G.R. Douglas, M. Schulzer, Neuroretinal rim areas and visual field indices in glaucoma, Am. J. Ophthalmol. 99 (2) (1985) 107–110, https://doi.org/10.1016/0002-9394(85)90216-8.
[9] A. Russo, W. Mapham, R. Turano, et al., Comparison of smartphone ophthalmoscopy with slit-lamp biomicroscopy for grading vertical cup-to-disc ratio, J. Glaucoma 25 (9) (2016) e777–e781, https://doi.org/10.1097/IJG.0000000000000499.
[10] A. Almazroa, R. Burman, K. Raahemifar, V. Lakshminarayanan, Optic disc and optic cup segmentation methodologies for glaucoma image detection: a survey, in: C. Costagliola (Ed.), Journal of Ophthalmology 2015 (2015) 180972, https://doi.org/10.1155/2015/180972.
[11] N. Thakur, M. Juneja, Survey on segmentation and classification approaches of optic cup and optic disc for diagnosis of glaucoma, Biomed. Signal Process Control 42 (2018) 162–189, https://doi.org/10.1016/j.bspc.2018.01.014.
[12] C. Zheng, T.V. Johnson, A. Garg, M.V. Boland, Artificial intelligence in glaucoma, Curr. Opin. Ophthalmol. 30 (2) (2019) 97–103, https://doi.org/10.1097/ICU.0000000000000552.
[13] M. Biswas, V. Kuppili, L. Saba, et al., State-of-the-art review on deep learning in medical imaging, Front. Biosci. 24 (3) (2019) 392–426, https://doi.org/10.2741/4725.
[14] P.S. Grewal, F. Oloumi, U. Rubin, M.T.S. Tennant, Deep learning in ophthalmology: a review, Can. J. Ophthalmol. 53 (4) (2018) 309–313, https://doi.org/10.1016/j.jcjo.2018.04.019.
[15] U. Schmidt-Erfurth, A. Sadeghipour, B.S. Gerendas, S.M. Waldstein, H. Bogunović, Artificial intelligence in retina, Prog. Retin. Eye Res. 67 (2018) 1–29, https://doi.org/10.1016/j.preteyeres.2018.07.004.
[16] D.S.W. Ting, L.R. Pasquale, L. Peng, et al., Artificial intelligence and deep learning in ophthalmology, Br. J. Ophthalmol. 103 (2) (2019) 167–175, https://doi.org/10.1136/bjophthalmol-2018-313173.
[17] J.H. Wu, T. Nishida, R.N. Weinreb, J.W. Lin, Performances of machine learning in detecting glaucoma using fundus and retinal optical coherence tomography images: a meta-analysis, Am. J. Ophthalmol. 237 (2022) 1–12, https://doi.org/10.1016/j.ajo.2021.12.008.
[18] J.I. Orlando, H. Fu, J. Barbosa Breda, et al., REFUGE Challenge: a unified framework for evaluating automated methods for glaucoma assessment from fundus photographs, Med. Image Anal. 59 (2020) 101570, https://doi.org/10.1016/j.media.2019.101570.
[19] E.L. Mayro, M. Wang, T. Elze, L.R. Pasquale, The impact of artificial intelligence in the diagnosis and management of glaucoma, Eye 34 (1) (2020) 1–11, https://doi.org/10.1038/s41433-019-0577-x.
[20] A.C. Thompson, A.A. Jammal, F.A. Medeiros, A review of deep learning for screening, diagnosis, and detection of glaucoma progression, Translational Vision Science & Technology 9 (2) (2020), https://doi.org/10.1167/tvst.9.2.42, 42-42.
[21] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, 2015, pp. 234–241.
[22] K. Nakahara, R. Asaoka, M. Tanito, et al., Deep learning-assisted (automatic) diagnosis of glaucoma using a smartphone, Br. J. Ophthalmol. 106 (4) (2022) 587–592, https://doi.org/10.1136/bjophthalmol-2020-318107.

[23] F. Li, Y. Su, F. Lin, et al., A deep-learning system predicts glaucoma incidence and progression using retinal photographs, J. Clin. Invest. 132 (11) (2022), https://doi.org/10.1172/JCI157968.
[24] Almazroa Ahmed, Sami Alodhayb, Essameldin Osman, et al., Retinal fundus images for glaucoma analysis: the RIGA dataset. In 10579 (2018) 105790B, https://doi.org/10.1117/12.2293584.
[25] R.C. Gonzals, R.E. Woods, Digital Image Processing, third ed., Prentice Hall, 2018.
[26] M. Subbarao, T.S. Choi, A. Nikzad, Focusing techniques, Opt. Eng. 32 (11) (1993) 2824–2836, https://doi.org/10.1117/12.147706.
[27] A. Sevastopolsky, Optic disc and cup segmentation methods for glaucoma detection with modification of U-Net convolutional neural network, Pattern Recogn. Image Anal. 27 (3) (2017) 618–624, https://doi.org/10.1134/S1054661817030269.