UNIVERSITY OF THESSALY
SCHOOL OF SCIENCES
DEPARTMENT OF COMPUTER SCIENCE AND BIOMEDICAL INFORMATICS

Object segmentation in image and video using deep learning

Nousias Georgios

THESIS
Supervisor
Delibasis Konstantinos
Associate Professor

Lamia, 2020-2021

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΪΑΤΡΙΚΗ

Τμηματοποίηση αντικειμένων σε εικόνες και βίντεο μέσω νευρωνικών δικτύων encoder decoder βαθείας μάθησης

Νούσιας Γεώργιος

ΠΤΥΧΙΑΚΗ

Επιβλέπων
Δελήμπασης Κωνσταντίνος
Αναπληρωτής Καθηγητής

Λαμία, 2020-2021

Στον πατέρα μου

ΕΥΧΑΡΙΣΤΙΕΣ

Στο σημείο αυτό, θα επιθυμούσα να ευχαριστήσω τα άτομα τα οποία με βοήθησαν και με υποστήριξαν να πραγματοποιήσω την παρούσα πτυχιακή εργασία. Αρχικά, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα της πτυχιακής εργασίας, Δρ. Κωνσταντίνο Δελήμπαση, αναπληρωτή καθηγητή του τμήματος Πληροφορικής με Εφαρμογές στη Βιοϊατρική του Πανεπιστημίου Θεσσαλίας, για την καθοδήγηση και την έμπνευση, την περίσσεια γνώσεων και χρόνου που διέθεσε για την πραγματοποίηση της εργασίας. Για την συνεχή υποστήριξη, την πίστη και την ικανότητά του να με παρακινεί στο επόμενο βήμα, ως ένας μέντορας.

Επιπλέον, θα ήθελα να ευχαριστήσω την οικογένειά μου που με στηρίζει καθ' όλη την διάρκεια της φοιτητικής και όχι μόνο ζωής μου, αλλά και τα άτομα που βρίσκονται δίπλα μου, με πιστεύουν και με στηρίζουν, συνοδοιπόροι σε κάθε όμορφη και άσχημη στιγμή, όλον αυτό τον καιρό.

Τέλος, θα ήθελα να ευχαριστήσω και να αφιερώσω την παρούσα πτυχιακή στον εκλιπών πατέρα μου που πίστεψε και με παρότρυνε να ακολουθήσω αυτό το μονοπάτι στη ζωή μου.

Νούσιας Γεώργιος,
Λαμία 2021

«Με υπομονή και επιμονή τα πάντα είναι εφικτά»

Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις [1], που προβλέπονται από τις διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. *Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί **χωρίς να τα περικλείω σε εισαγωγικά** και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.*

2. *Δέχομαι ότι η αυτολεξεί **παράθεση χωρίς εισαγωγικά**, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π. χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.*

3. *Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια*

4. *Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.*

Ημερομηνία: 29/4/2021

Ο Δηλ.

Νούσιας Γεώργιος

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.

Τμηματοποίηση αντικειμένων σε εικόνες και βίντεο μέσω νευρωνικών δικτύων encoder decoder βαθείας μάθησης

Νούσιας Γεώργιος

Τριμελής Επιτροπή:

Δελήμπασης Κωνσταντίνος, Αναπληρωτής Καθηγητής (επιβλέπων)

Ιακωβίδης Δημήτριος, Καθηγητής

Πλαγιανάκος Βασίλειος, Καθηγητής

Contents

Content of Figures

Content of Tables

Abstract

Blink detection can provide a very useful clinical indicator, because of its relation with many neurological and ophthalmological pathologic conditions. In this thesis, a system is proposed to automatically detect and classify blinks as "*complete*" or "*incomplete*" in image sequences. This method utilizes iris and eyelid segmentation in both eyes from the acquired images, using state-of-the-art neural network (DeepLabv3+), U-net and Segnet deep learning encoder-decoder neural architectures - DLEDs. The sequence of the segmented frames is post-processed to calculate the distance between the eyelids of each eye (palpebral fissure height) and the corresponding iris diameter. These quantities are temporally filtered. Their fraction is subjected to adaptive thresholding to identify blinks and determine their type, on each eye independently. Two DLEDs, of the same architecture, were trained with manually segmented images of iris and eyelids, respectively. The post-process was parameterized using a 4-minute video. The proposed system was tested on eight (8) subjects, each with a 4-10-minute video. Several metrics of blink detection and classification accuracy were calculated against the ground truth, which was generated by three (3) independent experts, whose differences were resolved by a senior expert. Results show that the proposed system achieved blink detection and classification accuracy between 79.8% and 98.7% for each of the 8 subjects. It outperformed all three (3) experts in terms of accuracy for 3 participants and two of the three experts for 2 of the remaining participants. The proposed system was proven robust in handling unexpected participant movements and actions, as well as glares and reflections from the spectacles. Also, the trained DLEDs were acquired and tested on RGB videos, that were captured by a common web camera. Despite the fact that training dataset did not included any images with those lighting conditions, the trained neural networks were able to detect and segment iris and sclera.

Keywords

Deep learning, medical image segmentation, neural networks, encoder-decoder, blink detection and classification

1 Introduction

1.1 Artificial Intelligence, Machine Learning and Deep Learning

Artificial Intelligence, known as **AI**, is a general term used to show that computers can deal with some tasks and "think" as a human brain.

Machine Learning (**ML**) represents a set of algorithms trained on various data.

Deep Learning (**DL**), is a subset of Machine Learning inspired by the biological structure of the human brain. In order to solve any task, DL is based on a structure similar to the biological neurons. So, DL uses a multi-layered structure of algorithms, known as neural network. Alike the human brain, neural network can be trained to recognize patterns and classify different types of data. The relation between those three terms is illustrated on Fig. 1.

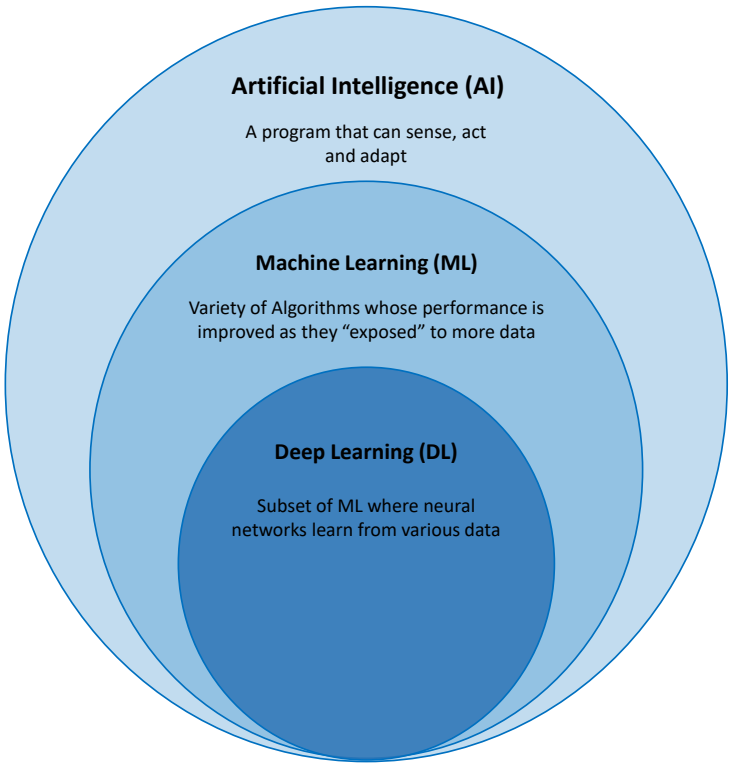


Figure 1: Relationship of AI, ML and DL

1.2 Neural Networks, Encoder-Decoder architecture

1.2.1 Neural networks

Neural network is a set of algorithms applied to recognize underlying relationships in a set of data. A neural network is composed of multiple hidden layers. Each layer is created by numerous neurons. A neuron is a mathematical model which takes inputs (i_1,

i_2, …, i_n), multiplies them by their corresponding weights (w_1, w_2, …, w_n) and then passes the sum through a function (f), called activation function, to the other neurons. The structure of a neuron is illustrated on Figure 2.

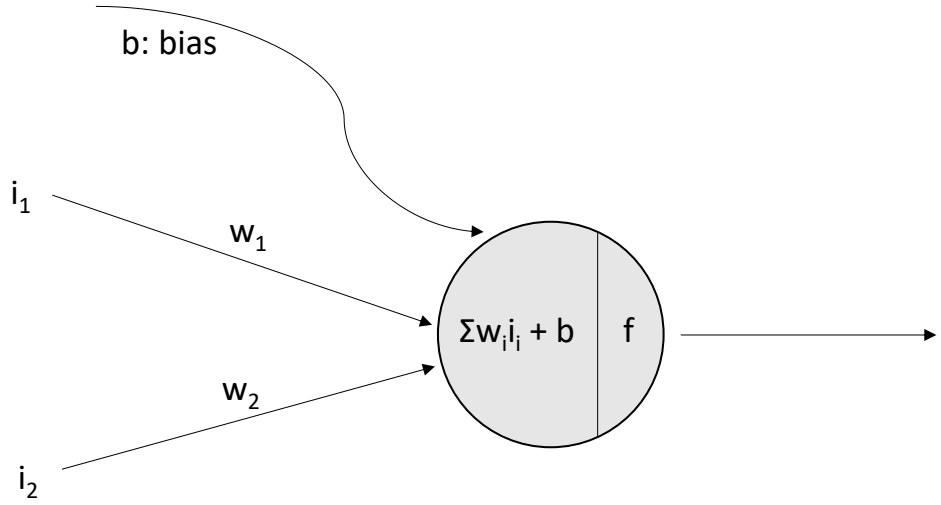


Figure 2: A simple neuron

There are various neural networks based on the number and architecture of each layer of the multilayer neural network, such as the U-net, SegNet, Alexnet, MobileNet, Deeplabv3+ etc. In this thesis, the architecture of U-net, Segnet and Deeplabv3+ will be explained. The results that have been conducted in Section 3 (Results) are using the Deeplabv3+ neural network.

1.2.2 Encoder-Decoder Architecture

The encoder-decoder networks have been successfully applied to many computer vision tasks. Typically, the encoder-decoder network contains:

i. an encoder module, also known as a down-sampling path, that gradually reduces the features maps, capturing higher semantic information
ii. a decoder module or an up-sampling path that retrieves the spatial information using transposed convolutions.

This architecture is used in various neural networks, as SegNet, U-net and Deeplabv3+.

U-net is a convolutional neural network with encoder-decoder architecture with a total of 23 convolutional networks. An example of U-net's architecture is illustrated in Figure 3. The encoding path is similar to a convolutional network. Each layer of the encoder consists of the repeated applications of two (2) 3×3 convolutions, followed by

Rectified Linear Unit (ReLU) and a 2×2 max pooling operation with stride 2 for down-sampling. Every step in the decoder consists of an up-sampling of the feature map followed by a 2×2 convolution, a concatenation with correspondingly cropped feature map from the contracting path, and two 3×3 convolutions each followed by a ReLU. At the final layer a 1×1 convolution is used to map each feature vector to the desired number of classes.

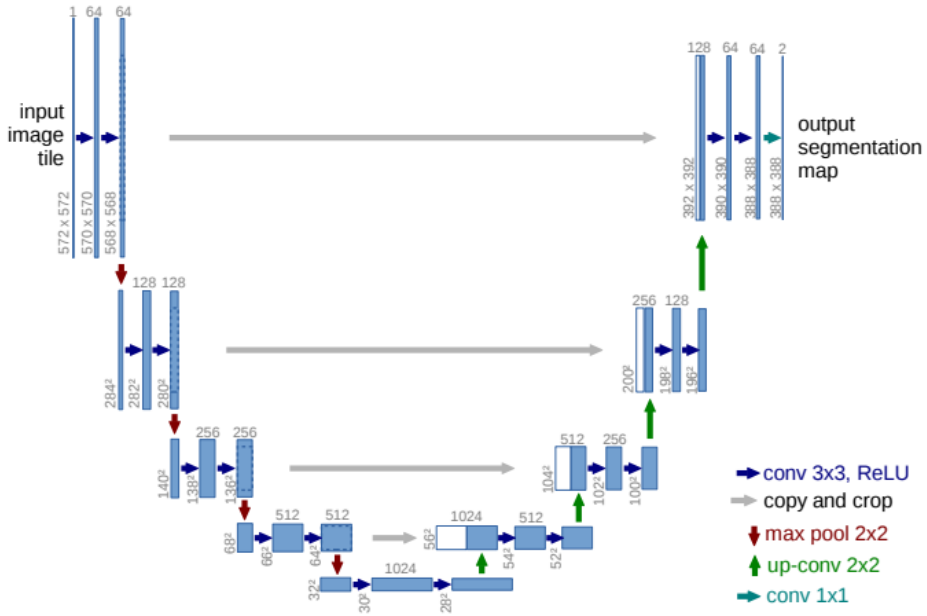


Figure 3: U-net Architecture [13]

SegNet has an encoder network and a corresponding decoder network, followed by a final pixelwise classification layer. The encoder network consists of 13 convolutional layers which correspond to the first 13 layers of VGG16 network. The final decoder output is fed to a multi-class soft-max classifier to conduct class probabilities for each pixel independently. An example of SegNet's architecture is illustrated in Figure 4.

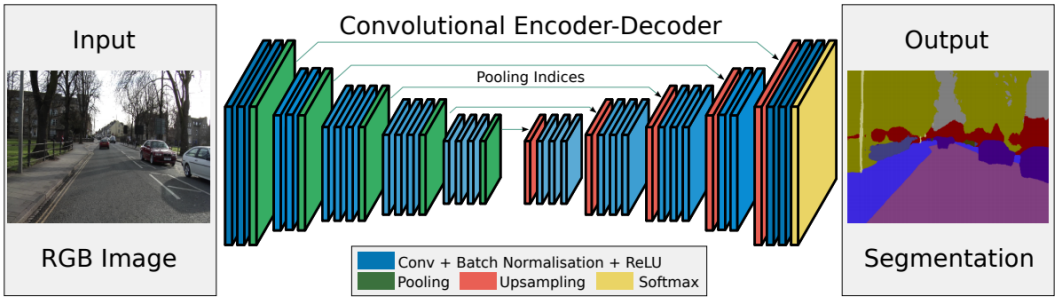


Figure 4: SegNet Architecture [14]

1.2.3 DeepLabv3+

DeepLabv3+ is a state-of-the-art neural network on image segmentation. Before the explanation of the architecture, some features should be defined, such as atrous convolutions and ASPP, to easily understand the structure of the neural network.

Atrous convolutions can explicitly control the resolution of the feature maps occurring after each layer of convolutional neural network. Also, atrous convolutions give the ability to adjust the convolutional filters in order to achieve capturing better multi-scale information. In case of a two-dimensional (2D) signal, like images, for each location ***i*** of the signal, on the output feature map ***y*** and a convolutional filter ***w***, atrous convolution is applied over the input ***x*** as follows:

$$y[i] = \sum_{k} x[i + r \cdot k] \; w[k] \tag{1}$$

where the atrous rate ***r*** determines the stride with which sampling is made at the input signal. It has to be noticed that standard convolution can be determined from Equation (2) for rate ***r*** = 1. The adjustment of the filters occurs due to the adaptively modification of the rate value.

Atrous Spatial Pyramid Pooling, also known as **ASPP**, is an atrous version of SPP in which the application of parallel atrous convolutions, with different rate at the original image, return a fused image as a result. While objects of the same class can have different scales in the image, ASPP helps to find out different object scales, improving the accuracy. In Figure 5, an example of ASPP's structure is illustrated.

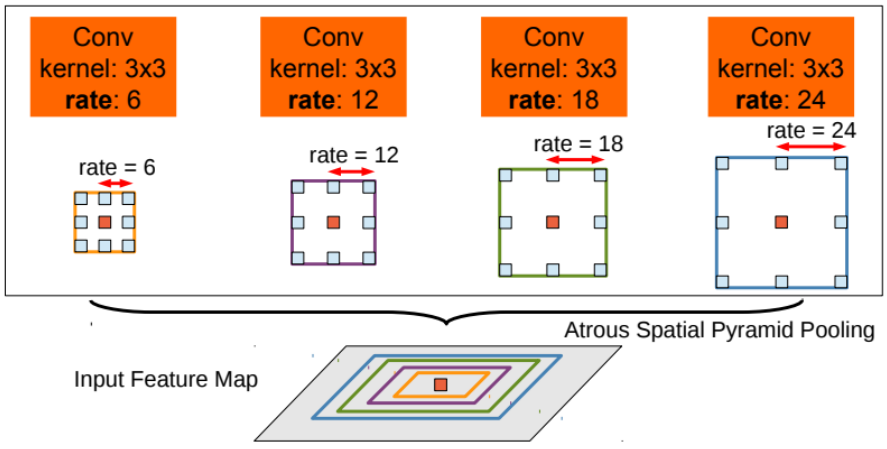
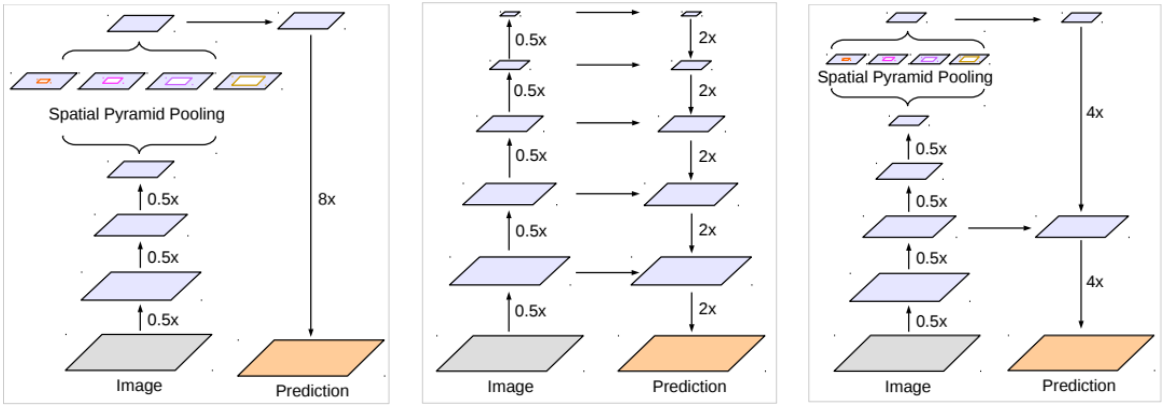


Figure 5: Atrous Spatial Pyramid Pooling (ASPP) [15]

DeepLabv3+ is based on atrous (or dilated) convolutions [15]. DeepLabv2 [16] implements a spatial pyramid pooling (ASPP) using a cascade of atrous convolutions with increasing rate, replacing the encoder – decoder architecture. It was further refined by using a parallel atrous convolution module (DeepLabv3) [17]. Finally, the Deeplabv3+ is combined with Deeplabv3, with a simple decoder module, to recover the object boundaries [18]. The evolve of DeepLab, until it reached its last state, can be observed in Figure 6, where Figure 6(a) represents first DeepLab's architecture and combining a decoder like the one of Figure 6(b), occurs the latest version of DeepLab, DeepLabv3+.



(a) (b) (c)

Figure 6: In Figure 6c is depicted DeepLabv3+ architecture [18]. This architecture is result of the combination of DeepLabv2 [16] architecture, Figure 6(a) and a simple decoder of a classic CNN, Figure 6(b)

1.3 Medical image segmentation

Medical image segmentation has a key role in computer-aided diagnostic systems with various applications. As medical imaging modalities are rapidly evolving, including microscopy, computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, X-ray and positron emission tomography, researchers' attention is triggered to develop and apply new medical image processing algorithms.

Image segmentation is considered to be the most essential image processing as it extracts the region of interest (ROI) with a semi-automatic or automatic method. The pixels of the original image are divided into areas/labels based on an algorithm.

1.4 Blink detection and classification

The assessment of blinking patterns has traditionally been of major interest due to their correlations with a series of ophthalmological and systemic diseases [1-7].

Formally, blink is defined as the provisional closure of both eyes, including the simultaneous movement of the upper and lower eyelids [8]. Blinks are classified into complete and incomplete, although there is no universal consensus exists in their definition. In theory, a complete blink requires full contact of the margins of the eyelids in one blink motion [9]. On the other hand, an incomplete blink should demonstrate partial cover of the pupil [10], cover less than two-thirds of the cornea [11], or no contact between the eyelids at all.

Blink-associated parameters attempt to address the unmet need for reliable, reproducible and non-invasive biomarkers that facilitate both the diagnosis and the follow-up of chronic diseases. For example, incomplete blinking is associated with decreased tear break-up time parameter (TBUT) [12], which is prevalent in dry eye disease (DED). However, since clinical measurement of TBUT is an invasive technique, that requires direct application of fluorescein on the ocular surface, incomplete blinking may be considered as an alternative measure of mild-to-moderate DED assessment.

Thus, automatic eye-blink detection and classification systems may assist the diagnosis of these diseases, as well as assess therapeutic schemes, both in clinical and research settings. This task is often tackled using computer vision techniques, since image acquisition and analysis can be performed continuously and non-invasively.

The field of machine learning and deep neural networks has facilitated several computer-vision tasks, such as object segmentation and detection of Regions of Interest (ROI) in images. Deep learning encoder-decoder (DLED) neural architectures, such as U-net [13] and Segnet [14] are commonly used to segment medical images. DLED neural architectures consist of an encoder (down-sampling) part and a decoder that gradually up-samples the encoder's output by transposed convolution.

Many approaches have been proposed for eye blink detection, mostly without considering the classification of complete and incomplete blinks. In the study of Drutarovsky and Fogelton [19], an eye blink detection algorithm was suggested utilizing the vertical motions in the eye region, reaching mean accuracy 99% on public datasets. A very similar method was introduced using the Gunnar–Farneback tracker in the eye region and a finite state machine for each eye [20].

An approach based on multi- scale and orientation Gabor filtering [21] for blink detection reported precision of 84.62%. Additionally, a method using Haar wavelets and HOG (Histogram of Oriented Gradient) features, combined with SVM classifier

[22] reported blink detection with accuracy of 92.5% when tested using standard databases and 86% when tested under real world conditions. A similar approach utilizing Haar-like features, for face detection and template matching [23], reported eye-blink detection accuracy of 95% and 77% for good and poor illumination, respectively.

In the study of Choi et al. [24], a blink detection method was suggested using an AdaBoost classifier, achieving accuracy of 96% on their own dataset. Al-gawwam and Benaissa [25] proposed a novel facial landmark position estimation and used the vertical distance between the upper and lower eyelids and Savitzky–Golay (SG) filter to detect blinks with precision of 96.65% on standard datasets. A statistical Active Appearance Model (AAM), to track and detect eye blinking [26], achieved accuracy between 67.92% and 100%, on three different datasets. A real-time blink detector based on a SIFT GPU-implementation [27], reported detection rate of 97% on very low contrast images, acquired under near-infrared illumination.

Despite the number of researches done on simple eye blink detection, only a few techniques deal with blink classification, i.e., complete and incomplete blinks. The most recent of them [28] detects blink completeness, using Recurrent Neural Network (RNN) as a classifier. The F1 score was calculated between 0.879 and 0.976. It should be noted that the detection of the eyes in each frame was performed manually. Moreover, a method for eye blink detection and identification of five different eye states, employing color information [29], achieved 65% - 90% true positive rate.

In this work, a fully automatic eye-blink detection and classification blink is proposed, using an embedded camera at close distance to the face. The main contributions of this thesis is:

- The automatic blink classification into “complete” and “incomplete”
- The use of state-of-the art DLED (DeepLabv3+) for iris and eyelid (palpebral fissure) segmentation, trained on close-up face images, acquired during clinical examination
- The post-process of the segmented images to estimate palpebral fissure height and cornea (iris) diameter and their temporal adaptive filtering to detect and classify blinks

- The methodology of ground truth generation that involves three independent experts and a senior expert to resolve disagreement, allowing comparison between the proposed method and the human experts.

The proposed algorithm has been applied to our clinical datasets and on a publicly available dataset and is compared to other state-of-the-art methods.

2 Methodology

2.1 Overview of the proposed method

The proposed blink detection and characterization method incorporate the following steps. Initially, two DLEDs were trained to segment iris and eyelids (palpebral fissure) respectively, in both eyes, based on a 536-image training set. The trained DLEDs were applied on each frame of a given video and generated the corresponding segmented images. The blinks were detected and characterized as either complete or incomplete, using post-processing of the ratio of the current palpebral fissure height over the temporal median value of iris diameter of the corresponding eye. The post-process has been parameterized using a test video. Those steps are graphically depicted in Figure 7.

2.2 Eyelid and iris segmentation using Deep Learning

A web camera was used to acquire images during the clinical examination. Two different instances of DeepLabv3+ [17] neural network, using the MobileNetv2 [30] as backbone, were trained to segment a) the palpebral fissure, including sclera and iris, also called *eyelid segmentation*, and b) the iris, respectively. A typical example of the segmentation for both neural networks is shown in Figure 8. It has to be mentioned that the images contain only part of the face, in order to provide adequate spatial resolution for the characterization of the blink. This is the reason why, most of the available face detectors do not operate satisfactorily.

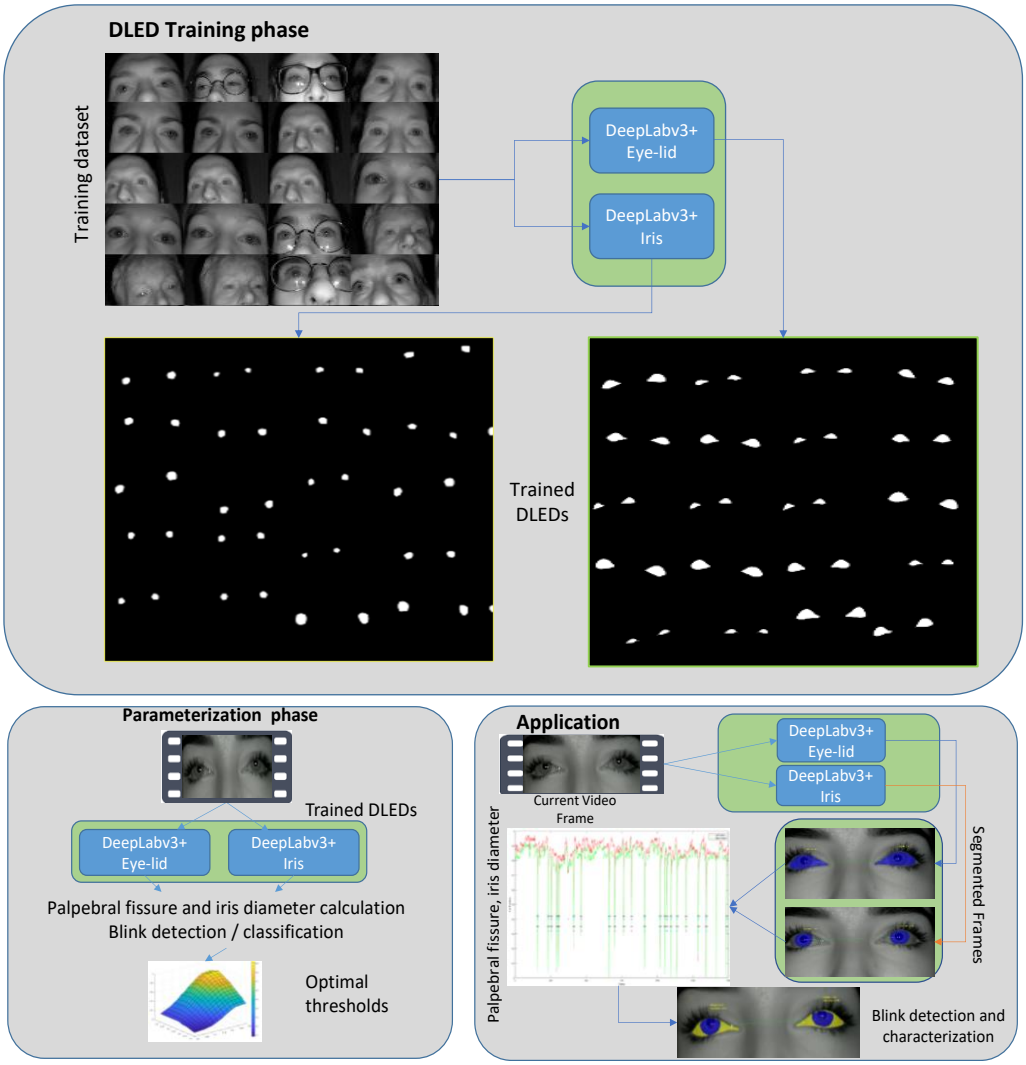


Figure 7: The main steps of the proposed algorithm

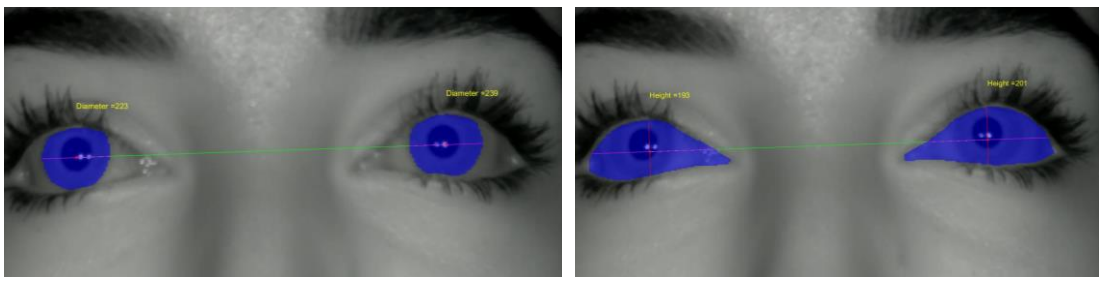


Figure 8: Eyelids segmented frame (Left) and iris segmented frame (Right)

The iris and the eyelid DLED networks were trained using a manually annotated data set of 536 RGB images for each neural network. The following parameters were used for training the DLEDs: minibatch size equal to 16, initial learning rate of 0.01 and a

maximum of 30 epochs. Training and testing dataset were acquired by the images from clinical examination videos of different participants.

2.3 Image post-process, calculation of palpebral fissure height and horizontal iris diameter

When the DLED networks are applied, the binary images with the segmented iris and eyelids are generated. The two largest connected components of each eye iris-segmented frame are identified and their centroids C_R, C_L are calculated. The line defined by C_R, C_L is lengthened to the image's borders, using the Bresenham's algorithm [32] and the two line-segments lying inside the iris regions are identified, providing the diameter of each eye d_R, d_L (in pixels), as graphically illustrated in Figure 9(b).

Subsequently, the height of the palpebral fissure h_L, h_R for the left and the right eye are calculated. If the eyelids of both eyes have been segmented by the DLED, the two largest connected components of the segmented frame are identified and their centroids V_R, V_L are calculated. The line passing through V_R and is perpendicular to the line (V_R, V_L) is generated using the Bresenham's algorithm and its segment, that lies inside the corresponding binary connected component of the eyelid segmented frame, is determined. The height of the palpebral fissure h_R is set equal to the length of this segment. The same process is applied to calculate the height of the palpebral fissure h_L for the left eye. The calculations are depicted in Figure 9 for a typical frame. The current slope at frame k is calculated as shown below, providing that both eyelids have been detected,

$$\theta_k = \tan^{-1} \frac{V_{Ry} - V_{Ly}}{V_{Rx} - V_{Lx}}, \tag{3}$$

otherwise θ_k will not be defined in the current frame. The average slope at frame k, is calculated using the exponential forgetting recursive formula,

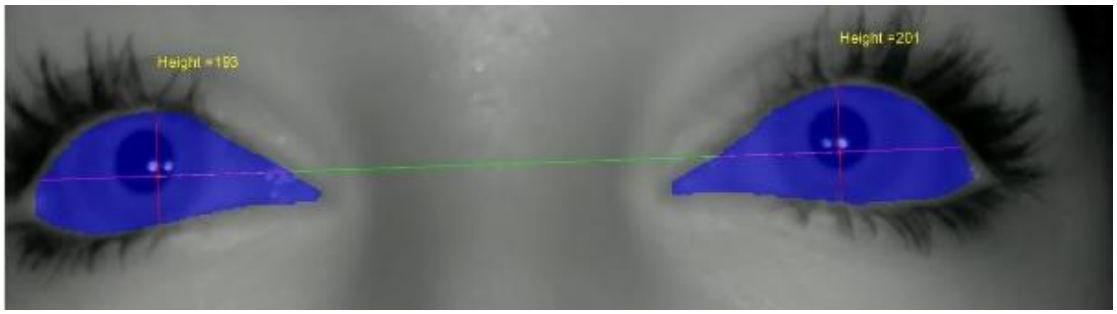
$$\bar{\theta}_k = a\theta_k + (1-a)\bar{\theta}_{k-1}, \tag{4}$$

where the parameter a is set to 0.1, providing that both eyelids have been detected in the current frame with the current value of θ_k close enough to the average slope,

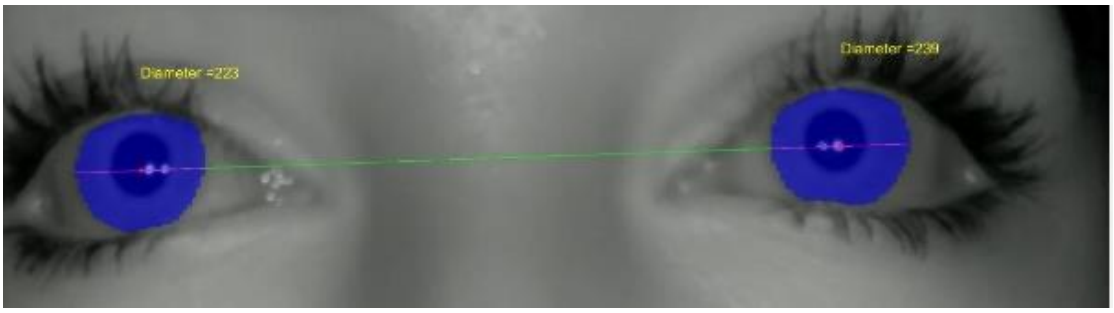
$\left|\theta_k - \bar{\theta}_{k-1}\right| = 0.2\,\text{rad}$, in order to prevent possible outliers affecting the average slope. Otherwise, the current average slope duplicates the previous average value: $\bar{\theta}_k = \bar{\theta}_{k-1}$.

If the eyelids of only one eye have been segmented, then a slightly different approach is applied. First, the binary eyelid image is rotated by an angle equal to the current value of the average slope $-\bar{\theta}_k$. Subsequently, the binary rotated image is projected horizontally and the height of the palpebral fissure h_R or h_L is set equal to the length of the non-zero projection. The aforementioned steps are shown graphically in Figure 10.

If no eyelid has been segmented, then the corresponding eye is assumed to be fully closed.



(a)



(b)

Figure 9: Eyelid (a) and iris (b) segmentation using the trained DLEDs and calculation of the palpebral fissure heights (h_R, h_L) and iris diameters (d_R, d_L)

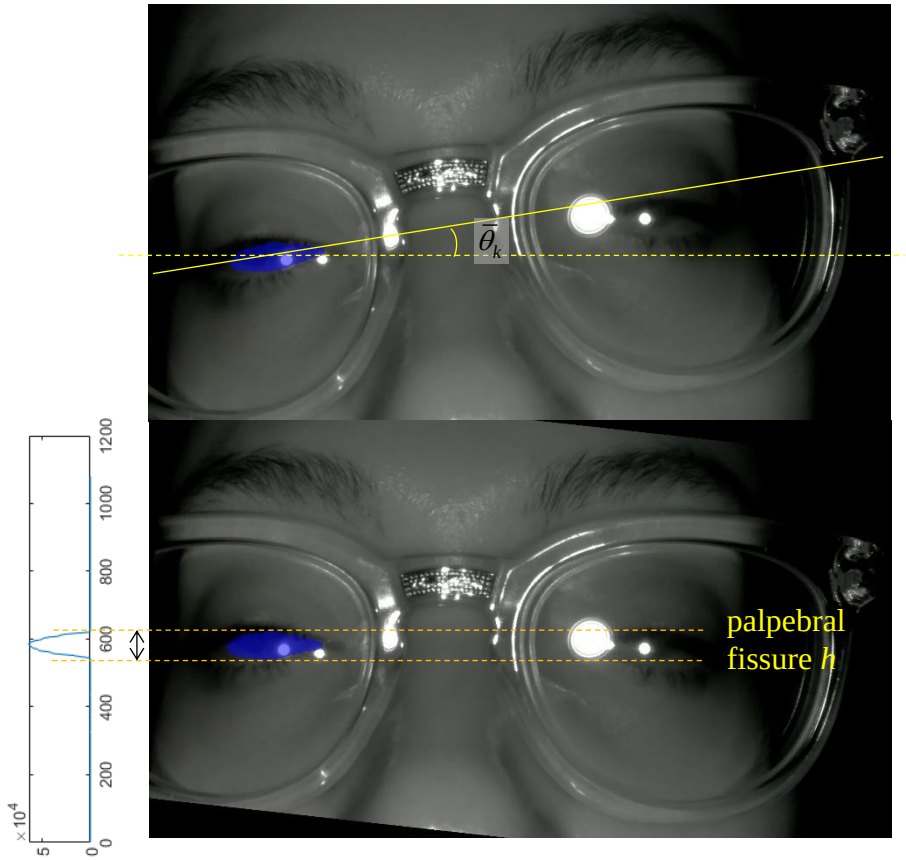


Figure 10: Calculation of palpebral fissure height for one eye only.

2.4 Identification of complete and incomplete blinks

The identification of complete and incomplete blinks is based on the previously defined parameters (d_R, d_L, h_R, h_L) and two thresholds. More specifically, for each frame k of a video, the palpebral fissure heights h_R^k, h_L^k and the moving median values $\bar{d}_R^k, \bar{d}_L^k$ of iris diameter are calculated by:

$$\bar{d}_R^k = median\left(d_R^{k-n}, ..., d_R^k, ..., d_R^{k+n}\right) \tag{5}$$

$$\bar{d}_L^k = median\left(d_L^{k-n}, ..., d_L^k, ..., d_L^{k+n}\right) \tag{6}$$

The median value of iris diameter is computed over a window of length $2n+1$ frames that includes the element in the current position, n elements backward, and n elements forward, with n set to a number of frames that span 0.6 seconds (defined experimentally). Subsequently, the following fraction is defined for each eye:

$$p_R^k = \frac{h_R^k}{\bar{d}_R^k}, p_L^k = \frac{h_L^k}{\bar{d}_L^k}. \tag{7}$$

These quantities express the fissure height as a fraction of moving median iris diameter, enabling the proposed system to handle cases of different patient-camera distance or the patient's orientation, throughout video acquisition. Furthermore, the median value of iris diameter is used in order to calculate the p_R^k, p_L^k fractions in each frame, even when the iris is partially or fully occluded in the current frame.

The moving median value of the fractions are calculated over a number of frames n_1 that cover approximately 5 seconds before and after the current frame k.

$$\bar{p}_R^k = median\left(p_R^{k-n_1}, ..., p_R^k, ..., p_R^{k+n_1} \right) \tag{8}$$

$$\bar{p}_L^k = median\left(p_L^{k-n_1}, ..., p_L^k, ..., p_L^{k+n_1} \right) \tag{9}$$

The blink is detected independently for each eye (R or L), the type of blink, as well as the starting and ending frames ($start_R$, end_R and $start_L$, end_L) are determined using the following heuristic algorithm that uses two global thresholds $0<T_2<T_1<1$.

For each frame k

If $p_R^k < T_1 \bar{p}_R^k$ and $p_R^{k-1} \geq T_1 \bar{p}_R^{k-1}$ Then $start_R$=k

If $p_R^k \geq T_1 \bar{p}_R^k$ and $p_R^{k-1} < T_1 \bar{p}_R^{k-1}$ Then end_R=k

If $\min\left(p_R^m \right) < T_2 \bar{p}_R^m, m = start_R, ..., end_R$ then

type= "complete" else type= "incomplete"

In Figure 11 it is shown an example of segmented image sequences for a complete (left) and incomplete (right) blink from patient 1. Yellow and blue colors refer to the two regions, eyelids and superimposed iris, respectively, after the application of the two neural networks. The values of $\bar{d}_R^k, \bar{d}_L^k$ and h_R^k, h_L^k are overlaid on each frame, along with the graphic illustration of the fractions $\bar{p}_R^k$, $\bar{p}_L^k$ (at the upper left corner of each frame). The values of d_R, d_L, h_R, h_L are shown in Figure 12 and Fig. 13 for one-minute of the same video.

The values of p_R^k, p_L^k are shown in Figure 14. The starting and ending frame of complete and incomplete blinks is indicated by blue symbols “o” and “+” for the proposed algorithm and black “o” and “+” for ground truth. The first two “rows” of symbols represent the left eye’s blinks and the remaining “rows” the right eye’s blinks. The thin dashed lines indicate the current values of $T_1\overline{p}_R^k$ and $T_1\overline{p}_L^k$ (red and green respectively).

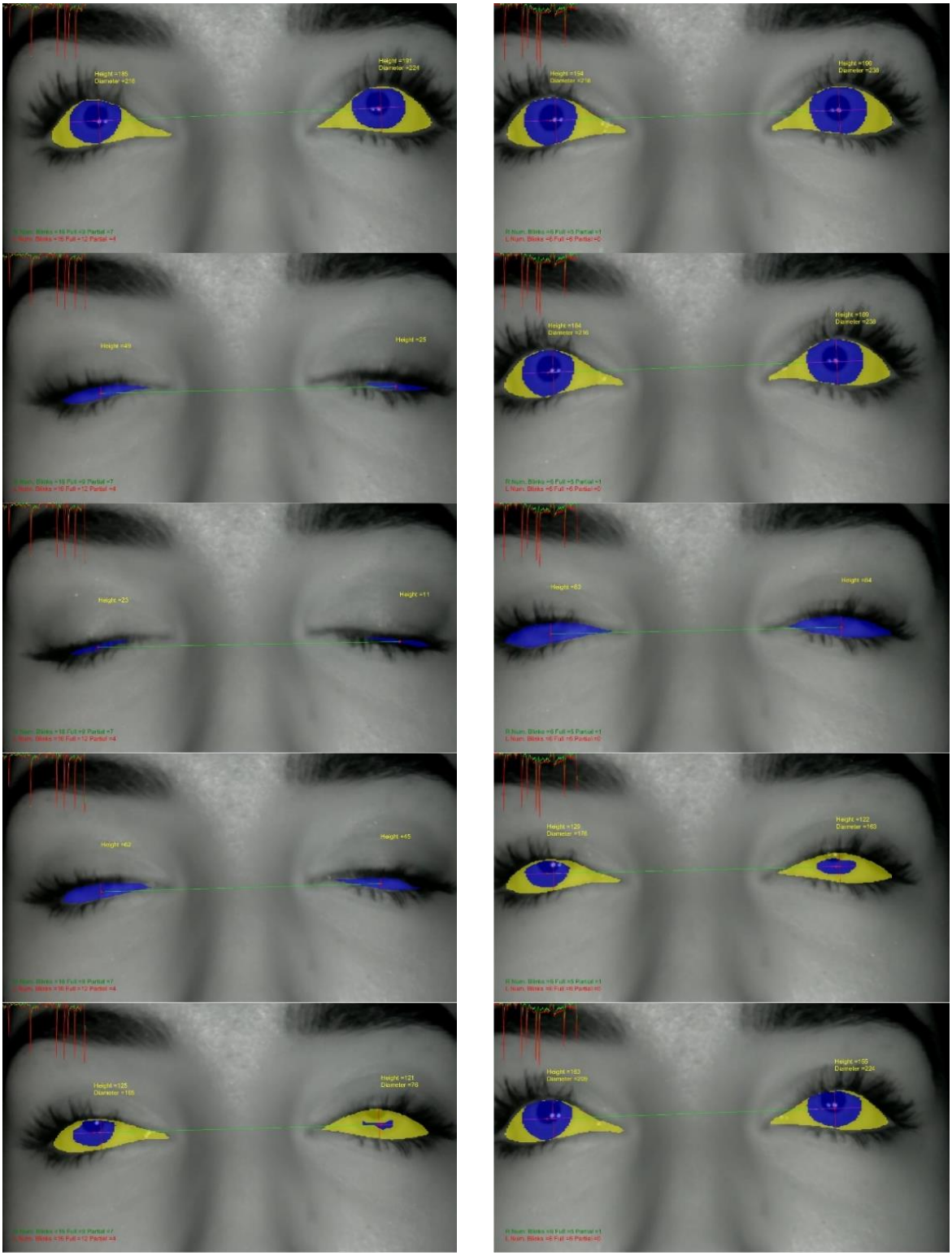


Figure 11: Example of complete (a) and incomplete (b) blink of Subject 1 (frames 1134–1138 and 460–464). DLED-segmented eyelids and superimposed iris shown in yellow and blue color, respectively.

The thick dashed lines indicate the current values of $T_2\bar{p}_R^k$ (red) and $T_2\bar{p}_L^k$ (green). The incomplete blink of Figure 11 corresponds to frames 460 – 464 and the full blink to frames 1134 – 1138.

Although the algorithm uses two global thresholds T_1, T_2, the inclusion of the moving median filtered horizontal iris diameter $\bar{d}_R^k, \bar{d}_L^k$ and the moving median filtered fractions $\bar{p}_R^k, \bar{p}_L^k$ are expected to provide robustness in handling different subjects, or motions / change in posture during video acquisition of the same subject. As it can be observed in Figure 14 the current thresholds of the fractions p_R^k, p_L^k for the incomplete blink ($T_1\bar{p}_R^k$ *and* $T_1\bar{p}_L^k$) and for the full blinks ($T_2\bar{p}_R^k$ *and* $T_2\bar{p}_L^k$) vary smoothly throughout the video of the same subject (red and green, thin and thick dashed lines).

After the application of the above algorithm to every frame of a given video, the detected blinks, with a duration less than three (3) frames, are discarded.

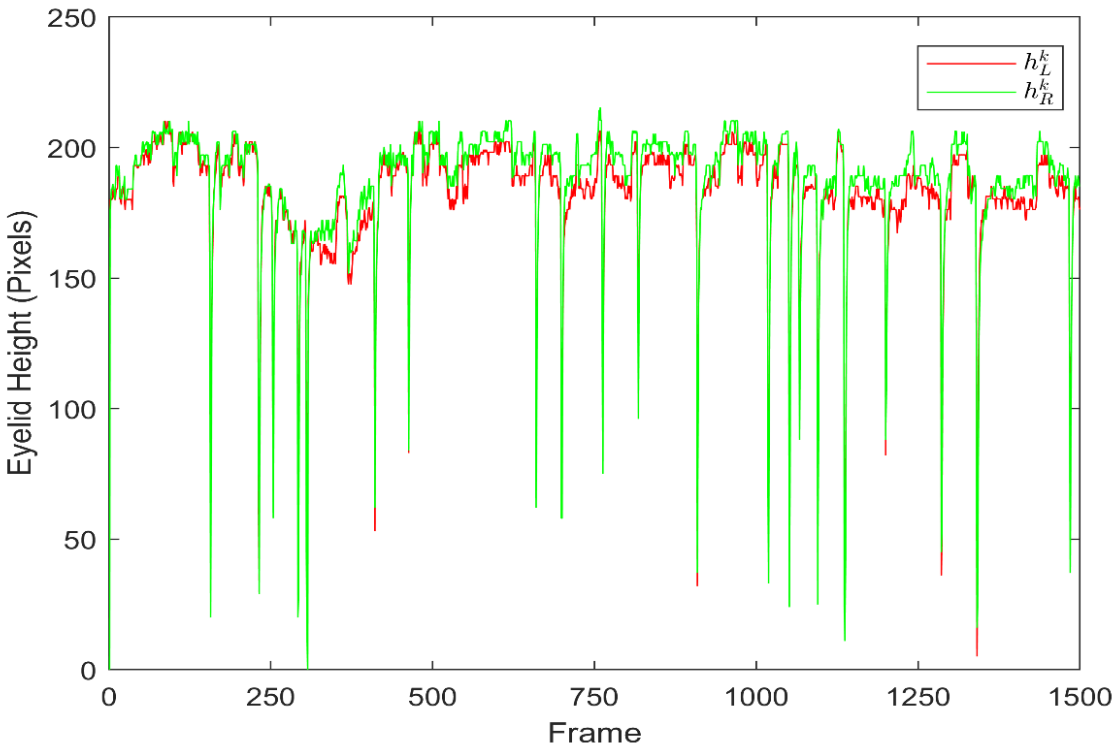


Figure 12: Palpebral fissure height (pixels).

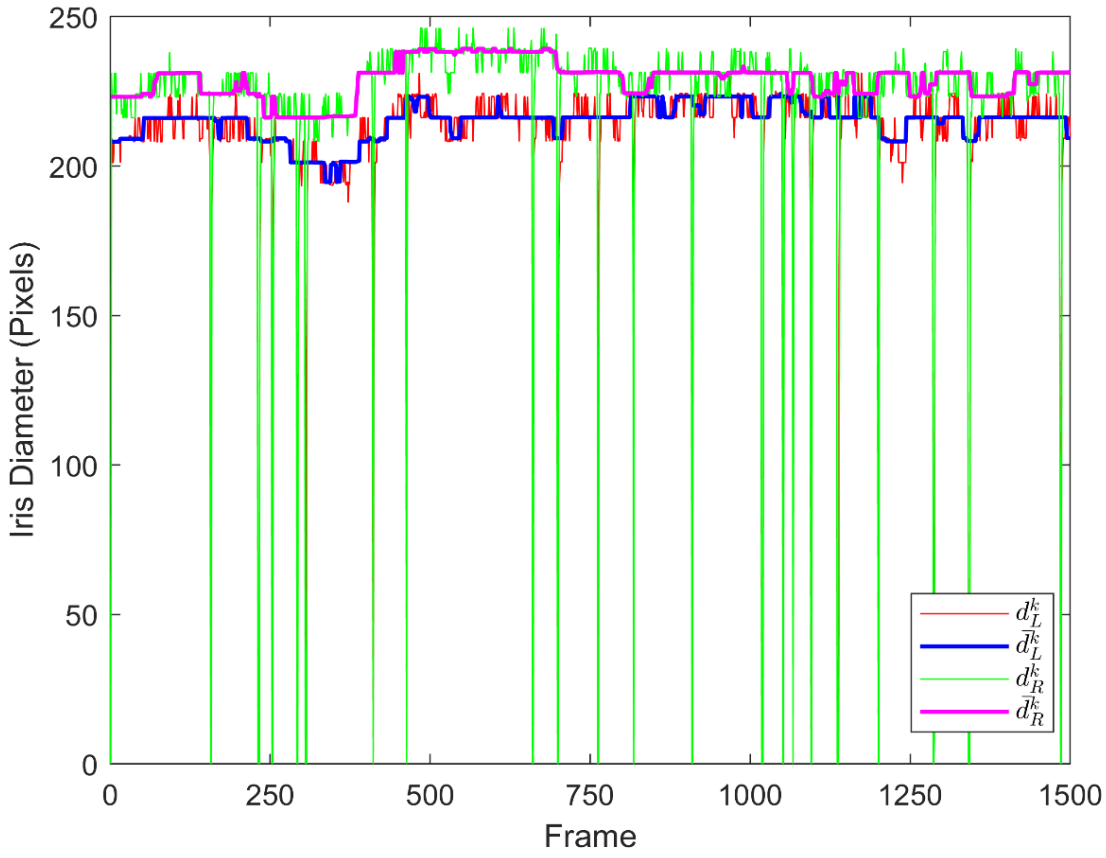


Figure 13: Iris diameter (pixels) of each eye. Also, the median value has been plotted. Obviously, when the blink is full (closed eye) the iris' diameter is zero, since the iris can't be detected.

2.5 Clinical setting

This was a prospective study. Protocol adhered to the tenets of the Declaration of Helsinki and written informed consent was provided by all participants. The institutional review board of Democritus University of Thrace approved the study protocol (protocol number/date of approval: ES2/Th15/25-2-2021). The clinical study was conducted between October 2020 and March 2021. Study official registration number is NCT04828187.

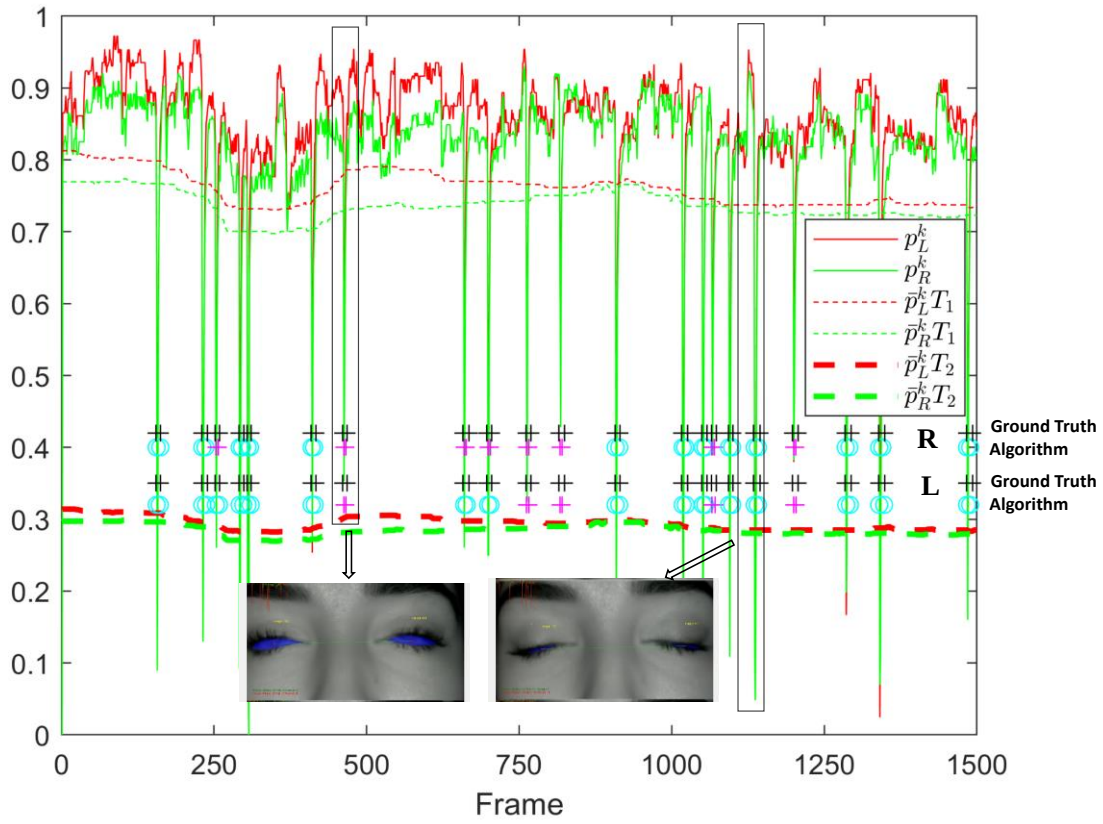


Figure 14: Current values of p_R^k, p_L^k (green and red continuous lines). "o" and "+": start and end of complete and incomplete blinks, determined by ground truth (black) and proposed system (blue), for the right (R) and left (L) eye. Thin dashed lines: the current values of $T_1\bar{p}_R^k$ (red) and $T_1\bar{p}_L^k$ (green). Thick dashed lines: current values of $T_2\bar{p}_R^k$ (red) and $T_2\bar{p}_L^k$ (green).

3 Results

3.1 Image and Video Datasets, blink ground truth annotation

The two DLEDs used for the segmentation of eyelids and iris were trained using 536 images with manual delineation of the object of interest (481 images of the iris dataset and 55 additional images of partially or fully closed eyes), resized to 288×288 pixels. All images were acquired using Raspberry Pi Camera Module IR-CUT v2 (5MP, 1080p) with 2 IR LED lights, connected to a Raspberry Pi 3 model v1.2, operating at the original resolution of 1080×1920. Camera setup is depicted in Figure 15. The participants were placed at a distance of 15 to 20 cm from the camera lens and were asked to watch a TV set about 2.5 meters away.

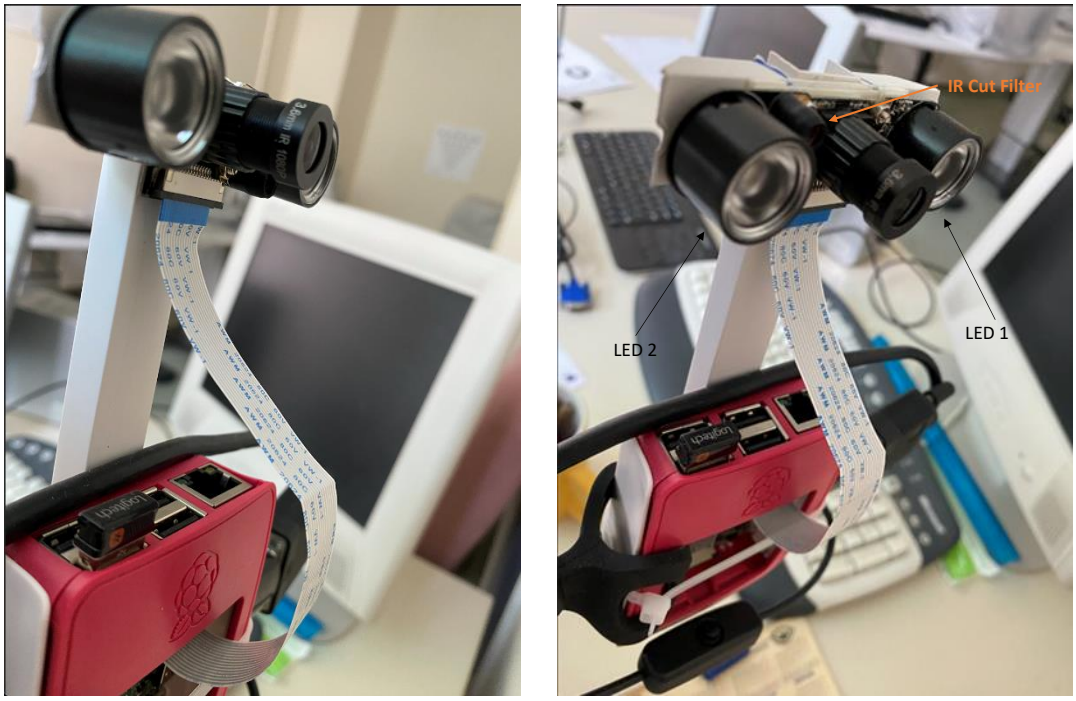


Figure 15: The IR camera setup for the clinical extraction of videos. The two IR LED lights are visible and annotated with arrows at the right image.

The training of DLEDs took place using MATLAB on a Microsoft Windows 10 computer with the following specifications: i7-2700K @3.5GHz, 16GB Ram, GPU: Nvidia GeForce GTX1650 super. The training time for each DLED was approximately 48 hours. The forward pass of a single frame through each DLEDs was equal to 0.06 seconds on average, including image resizing. After DLED training, the proposed system was tested on videos from eight different patients. Each video duration was between 4 and 10 minutes.

Two independent blink identifications are assumed to agree, if and only if there is sufficient temporal overlapping and the type of blink is the same. Given two identifications of a blink, defined by the starting-ending frame, $[a_1, a_2]$ and $[b_1, b_2]$, the fraction of (temporal) intersection over union (*IOU*) between the two blink identifications is calculated as [33]:

$$IOU = \frac{A \cap B}{A \cup B} = \frac{\min(a_2, b_2) - \max(a_1, b_1)}{\max(a_2, b_2) - \min(a_1, b_1)} \quad (6).$$

If the *IOU* is greater than or equal to 0.2, as proposed in the study of Choi et al. [24], then the two blink identifications agree. Generic examples of the application of *IOU*

are depicted in Figure 16. Please note that in Blink 3 the intersection will obtain negative value, thus, by definition, the two blinks are not in agreement.

This definition for blink identification agreement is used to compare expert – ground truth and proposed algorithm – ground truth identifications.

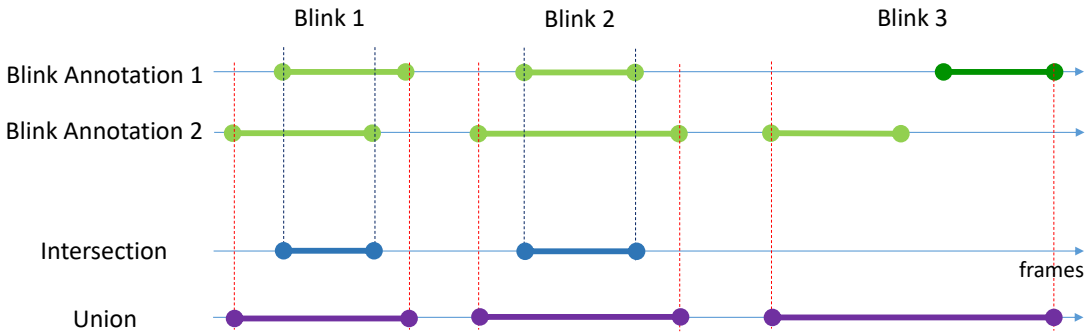


Figure 16 Examples of IOU calculation between two blink identifications.

The starting and ending frame and the type of each blink was annotated in every video by three (3) independent experts, according to their clinical experience. The annotations were reviewed by a senior expert, who provided the final blink annotation in case of disagreement between any two experts: non-unanimous type of blink, or $IOU<0.2$, as depicted in Figure 17, Blink 2 and 3, respectively. If all three experts agree for a specific blink, then the ground truth for this blink is generated using the average starting and average ending frame indicated by the experts (as depicted in Figure 17, Blink 1).

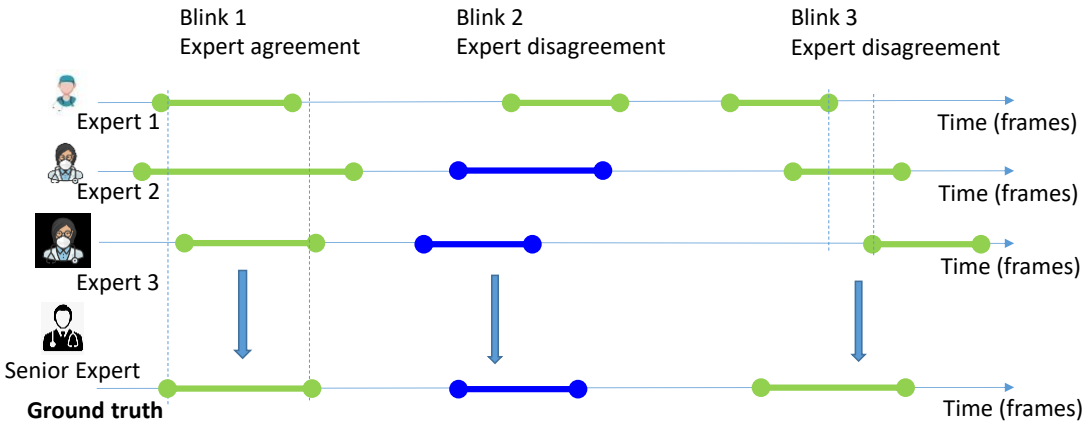


Figure 17 Annotation by the 3 independent experts and the final ground truth, generated by resolving expert disagreement by a senior expert. The green and blue lines indicate detection of complete and incomplete blink respectively.

Blinks were annotated in all videos of the 8 participants using this method and the details are shown in Table I. The results of the proposed system on these videos will be compared with the ground truth in the next subsection.

3.2 Quantitative segmentation results

In Figure 19 two short sequences of frames are presented, depicting a complete and an incomplete blink with the segmented iris and eyelid visualized in different color. Since the type of blink in Figure 19(a) is "complete", the palpebral fissure has not been detected. In the "incomplete" blink of Figure 19(b), the palpebral fissure is detected in each frame of the sequence, with smaller height, as expected.

Table I THE DETAILS OF BLINK-ANNOTATED VIDEOS

Patient	Duration (min)	Frames, fps	Complete blinks	Incomplete blinks
1	6	9.000, 25	183	67
2	10	15.000, 25	59	27
3	10	15.000, 25	448	0
4	5.3	8.000, 25	52	42
5	4	5.800, 25	182	16
6	4	6.000, 25	158	4
7	4	6.000, 25	97	51
8	5	7.000, 25	130	85

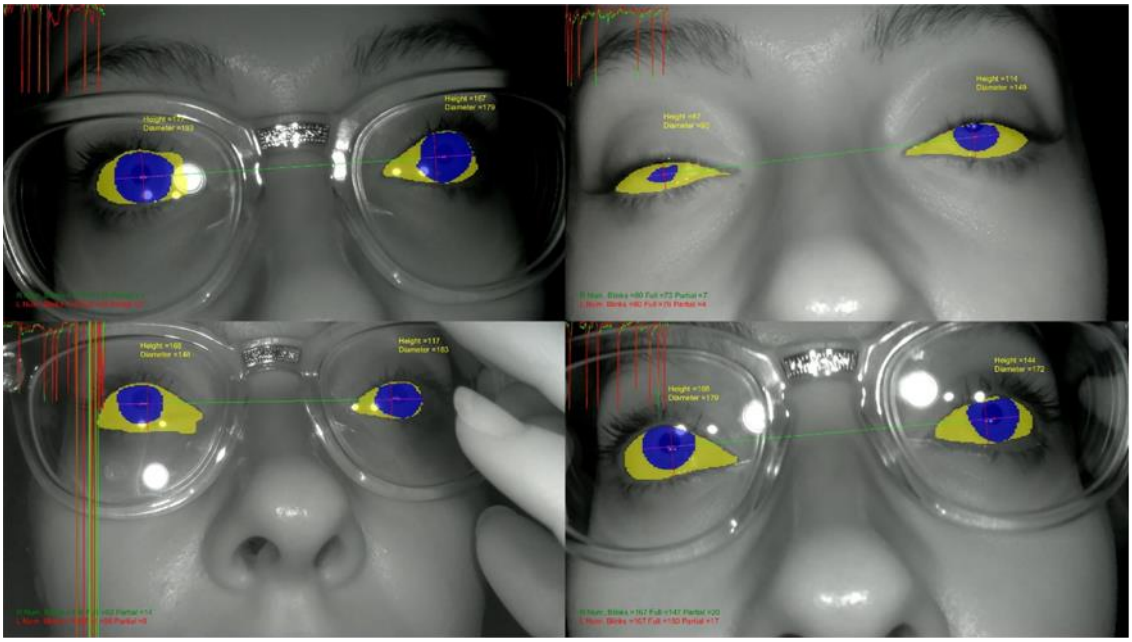
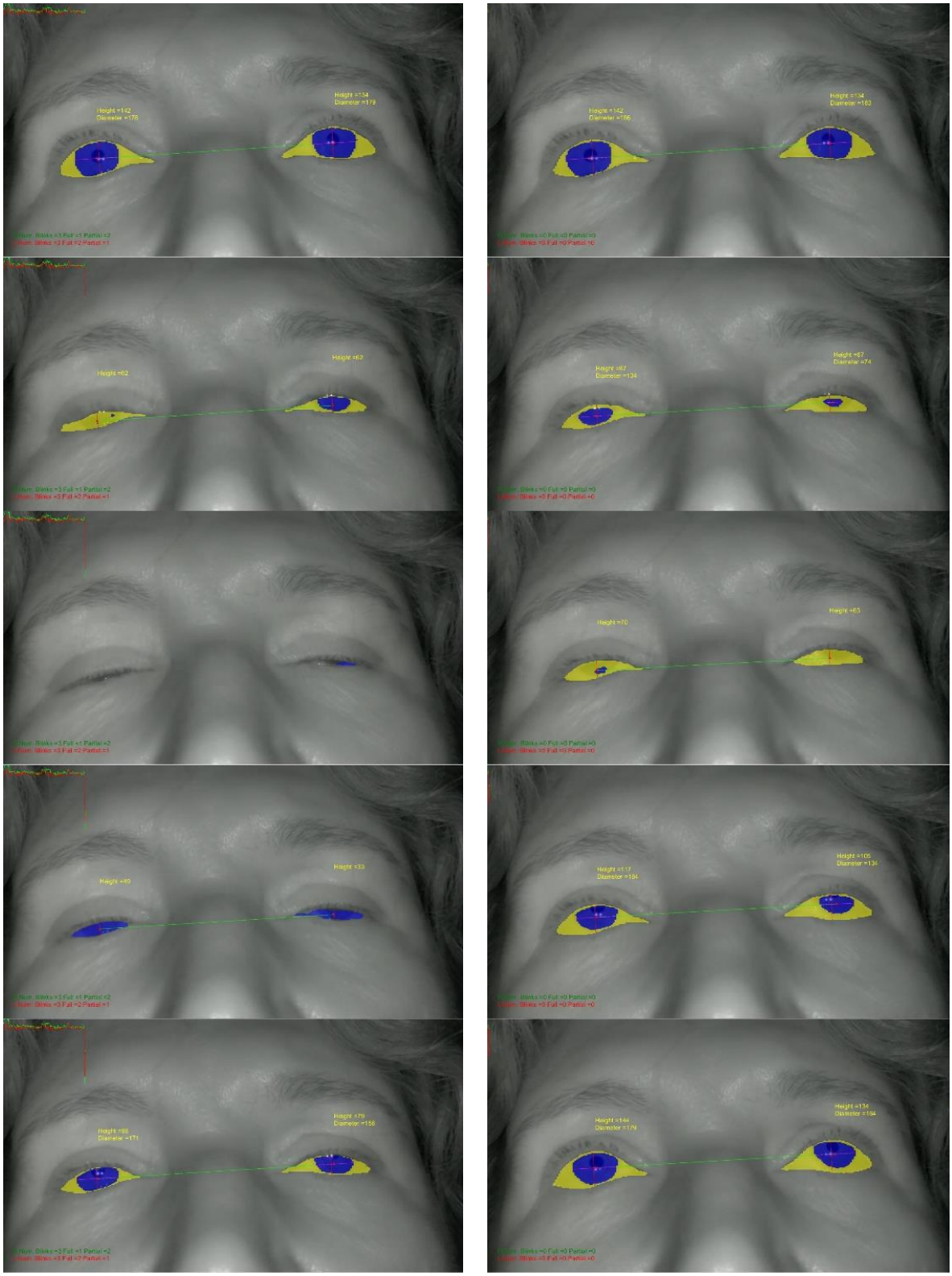


Figure 18: Iris and eyelids (sclera) are detected despite patient's movement.



(a) (b)

Figure 19 Sequence of frames for (a) “complete” and (b) “incomplete” blink from one of the participants, with segmented iris and eyelids. The palpebral fissure height and iris diameter have also been drawn.

The quality of eyelid and iris segmentation, achieved by the two DLEDs is further assessed qualitatively in Figure 18 that illustrates a few frames with successful segmentation, despite the patient's action to put the spectacle on and off.

3.3 Parameterization and quantitative results

In order to determine the type of blink, the values of thresholds T_1 and T_2 that are used in subsection 2.2 need be optimally set. To this end, the following algorithm is applied. A four-minute video of a different subject is input into the algorithm, using all combinations of T_1 in range [0.6, 0.9] and T_2 in range [0.1, 0.4] with step equal to 0.02. For these combinations, the system's overall accuracy is calculated and plotted in Figure 20. Based on this graph, the values of T_1 and T_2 that maximize optimally the overall blink detection and classification accuracy were detected as $T_1 = 0.88$ and $T_2 = 0.34$. These two optimal values of thresholds were used for all the results of this work.

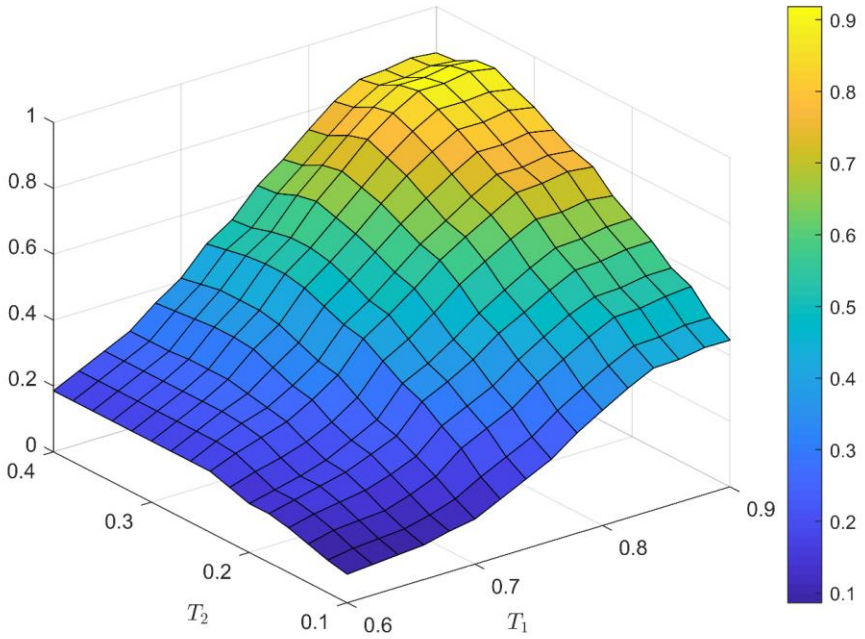


Figure 20 The proposed system's accuracy for different values of T_1, T_2. The surface is clearly unimodal, indicating the existence of optimal thresholds

After the determination of the two optimal thresholds, the proposed system was applied to the available videos and the resulting confusion matrices for each participant (collectively for both eyes) are shown in Table II. In the next eight sequenced tables,

Table III through Table X, the confusion matrices are presented for each patient and each expert.

Table II Confusion matrices of the proposed system summed for all participants

Ground Truth	Expert 1			Expert 2			Expert 3			System		
	C	**I**	**N**	**C**	**I**	**N**	**C**	**I**	**N**	**C**	**I**	**N**
C	1199	75	37	1169	124	19	1273	37	7	1285	18	6
I	79	200	17	18	271	7	40	255	1	75	213	4
N	31	27	0	8	2	0	5	18	0	4	15	0

Table III Confusion matrices for Patient 1

Ground Truth	Expert 1			Expert 2			Expert 3			System		
	C	**I**	**N**	**C**	**I**	**N**	**C**	**I**	**N**	**C**	**I**	**N**
C	140	23	20	179	4	0	183	0	0	179	3	1
I	26	39	2	7	60	0	0	67	0	13	53	1
N	10	5	0	0	2	0	0	2	0	0	1	0

Table IV Confusion matrices for Patient 2

Ground Truth	Expert 1			Expert 2			Expert 3			System		
	C	**I**	**N**	**C**	**I**	**N**	**C**	**I**	**N**	**C**	**I**	**N**
C	47	8	4	47	4	8	50	10	0	58	1	0
I	4	9	14	1	22	4	0	27	0	12	15	0
N	7	5	0	0	0	0	1	6	0	0	7	0

Table V Confusion matrices for Patient 3

Ground Truth	Expert 1			Expert 2			Expert 3			System		
	C	**I**	**N**	**C**	**I**	**N**	**C**	**I**	**N**	**C**	**I**	**N**
C	400	38	10	444	0	6	435	19	0	445	0	3
I	0	0	0	0	0	0	0	0	0	0	0	0
N	6	3	0	6	0	0	2	6	0	2	1	0

Table VI Confusion matrices for Patient 4

Ground Truth	Expert 1			Expert 2			Expert 3			System		
	C	**I**	**N**	**C**	**I**	**N**	**C**	**I**	**N**	**C**	**I**	**N**
C	51	1	0	47	4	1	44	5	3	52	0	0
I	3	39	0	9	32	1	2	39	1	6	36	0
N	5	7	0	0	0	0	0	0	0	0	0	0

Table VII Confusion matrices for Patient 5

Ground Truth	Expert 1			Expert 2			Expert 3			System		
	C	**I**	**N**	**C**	**I**	**N**	**C**	**I**	**N**	**C**	**I**	**N**
C	177	2	3	134	44	4	178	0	4	178	2	2
I	4	11	1	0	16	0	6	10	0	8	8	0
N	3	3	0	2	0	0	2	2	0	2	5	0

Table VIII Confusion matrices for Patient 6

Ground Truth	Expert 1			Expert 2			Expert 3			System		
	C	**I**	**N**	**C**	**I**	**N**	**C**	**I**	**N**	**C**	**I**	**N**
C	156	0	0	149	7	0	156	0	0	158	0	0
I	4	4	0	0	8	0	4	4	0	2	2	0
N	0	0	0	0	0	0	0	0	0	0	0	0

Table IX Confusion matrices for Patient 7

Ground Truth	Expert 1			Expert 2			Expert 3			System		
	C	I	N	C	I	N	C	I	N	C	I	N
C	97	0	0	88	9	0	97	0	0	97	0	0
I	33	18	0	1	50	0	23	28	0	20	31	0
N	0	4	0	0	0	0	0	2	0	0	1	0

Table X Confusion matrices for Patient 8

Ground Truth	Expert 1			Expert 2			Expert 3			System		
	C	I	N	C	I	N	C	I	N	C	I	N
C	131	3	0	81	52	0	130	3	0	118	12	0
I	5	80	0	0	83	2	5	80	0	14	68	3
N	0	0	0	0	0	0	0	0	0	0	0	0

In order to calculate the classification metrics (accuracy, sensitivity, specificity, precision, negative predictive value, false positive rate, false discovery rate) separately for the two blink classes, the following quantities are defined:

- TP (true positive): the number of blinks correctly characterized as the current class,
- TN (true negative): the number of blinks correctly characterized as the other class,
- FP (false positive): the number of wrong blink classifications as the current class,
- FN (false negative): the number of blinks wrongly classified as “no-blink”.

Based on Table II, the classification metrics are calculated separately for each patient and type of blink in Table XI. The metric’s definitions are presented in the below Equations (10)-(11).

$$Accuracy\ or\ Acc = \frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})} \quad (12)$$

$$Sensitivity\ or\ Sens = \frac{(\mathrm{TP})}{(\mathrm{TP}+\mathrm{FN})} \quad (13)$$

$$Specificity\ or\ Spec = \frac{(\mathrm{TN})}{(\mathrm{TN}+\mathrm{FP})} \quad (14)$$

$$Precision\ or\ Prec = \frac{(\mathrm{TP})}{(\mathrm{TP}+\mathrm{FP})} \quad (15)$$

$$Negative\ Predictive\ Value\ or\ NPV = \frac{(\mathrm{TN})}{(\mathrm{TN}+\mathrm{FN})} \quad (16)$$

$$False\ Positive\ Rate\ or\ FPR = \frac{(FP)}{(\mathrm{FP}+\mathrm{TN})} \quad (17)$$

$$False\ Negative\ Rate\ or\ FNR = \frac{(FN)}{(\mathrm{FN}+\mathrm{TP})} \quad (18)$$

$$False\ Discovery\ Rate\ or\ FDR = \frac{(FP)}{(\mathrm{FP}+\mathrm{TP})} \quad (19)$$

If none of the blinks have been characterized by the experts as one of the two classes and the system has also classified no blinks in this class, then specificity, precision, FPR and FDR cannot be calculated and are characterized as Not Applicable (N/A) in Table XI.

Finally, the annotations of each expert are also compared with the ground truth for each available video. The overall classification accuracy of the three experts, as well as the proposed system, is calculated as the fraction of blinks correctly classified and shown in Table XII. The proposed system outperformed at least 1 of the 3 experts in all participants. More specifically, the proposed system appears to be more accurate in blink detection surpassing all 3 experts in 3 out of 8 participants and 2 experts in 2 of the remaining participants.

Table XI CLASSIFICATION METRICS (%) FOR EACH SUBJECT, FOR TWO BLINK CLASSES (C: COMPLETE, I: INCOMPLETE)

Pat.	Class	Acc.	Sens.	Specif.	Prec.	NPV	FPR	FNR	FDR
1	C	93.15	97.81	80.00	93.23	92.86	20.00	2.19	6.77
	I	92.40	77.61	97.81	92.86	92.27	2.19	22.39	7.14
2	C	84.88	98.31	55.56	82.86	93.75	44.44	1.69	17.14
	I	78.49	55.56	87.88	65.22	82.86	12.12	44.44	34.78
3	C	98.89	99.33	0	99.55	0	100	0.67	0.45
	I	99.78	N/A	99.78	0	100	0.22	N/A	100
4	C	93.62	100	85.71	89.66	100	14.29	0	10.34
	I	93.62	85.71	100	100	89.66	0	14.29	0
5	C	93.00	97.80	44.44	94.68	66.67	55.56	2.20	5.32
	I	92.54	50.00	96.22	53.33	95.70	3.78	50.00	46.67
6	C	98.77	100	50.00	98.75	100	50.00	0	1.25
	I	98.77	50.00	100	100	98.75	0	50.00	0
7	C	86.49	100	60.78	82.91	100	39.22	0	17.09
	I	85.91	60.78	98.98	96.88	82.91	1.02	39.22	3.13
8	C	87.74	90.77	82.93	89.39	85.00	17.07	9.23	10.61
	I	86.51	80.00	90.77	85.00	87.41	9.23	20.00	15.00
All	C	93.56	98.17	72.85	94.21	89.83	27.15	1.83	5.79
	I	92.98	72.60	97.50	86.53	94.14	2.50	27.40	13.47

(Acc: accuracy, Sens: sensitivity, Spec: specificity, Prec: precision, NPV: negative predictive value, FPR: false positive rate, FDR: false discovery rate)

The proposed system was tested on the Talking Face dataset [34] and it was compared with state-of-the art methods of Soukopova et al. [31] and Fogelton et al. [28], in terms of F1 score. Talking Face dataset is a 200-second (5000 frames) video of a person engaged in conversation. Although it was not intended for blink classification, ground truth is provided with indication for complete blinks. In Table XIII, F1 score is presented for each method, for both types of blink, as well as for the simple blink detection.

TABLE XII OVERALL BLINK CLASSIFICATION ACCURACY FOR EACH PATIENT, ACHIEVED BY THE PROPOSED SYSTEM, AS WELL AS THE THREE MEDICAL EXPERTS (THE BEST PERFORMER IS SHOWN IN BOLD).

Patient	Expert 1	Expert 2	Expert 3	Proposed system
1	0.8283	0.9127	0.9094	**0.9243**
2	0.6980	0.8023	**0.8191**	0.7849
3	0.7912	0.9737	0.9416	**0.9867**
4	0.9158	0.8830	0.8085	**0.9362**
5	**0.95**	0.7677	**0.95**	0.9073
6	**0.9756**	0.9573	**0.9756**	0.9691
7	0.7566	**0.9324**	0.8333	0.8591
8	0.7727	0.6750	**0.8750**	0.8651

It can be observed that the proposed system performs better than the other two methods in blink classification (complete and incomplete), but marginally worse in blink detection.

Table XIII THE F1 SCORE ACHIEVED BY THE PROPOSED METHOD COMPARED WITH TWO EXISTING METHODS FOR THE TALKING FACE DATASET [34].

Blinks	Method 1 [31]	Method 2 [29]	Our Method
Complete	-	0.939	0.970
Incomplete	-	0.25	0.66
Total	0.948	0.971	0.928

3.4 Testing regular light-condition videos

The two DLEDs were tested also on videos that captured using a simple web camera. Despite the fact that the training dataset of DLEDs did not contain any images that were captured during normal conditions (like physical light or light from a table lamp), the two neural networks are able to detect and segment sclera and iris, respectively.

The steps of the algorithm that was followed are:

1. Application of an eye tracking algorithm
2. Creation of a .txt file that contains the bounding box of eye region
3. Crop of the original video using the txt file
4. Apply the two (2) DLEDs for blink detection and classification
5. Replace the cropped-segmented video region with the corresponding region of the original video

The above steps are illustrated in Figure 22, while an example of a segmented blink sequence is depicted in Figure 21. Also, at left lower corner of each video frame are displayed the total number of right and left blinks, as also and its class ("complete" or "incomplete"). At left upper corner, a continuous line is generated indicating the fractions p_R^k, p_L^k for each eye with different color, throughout the video. Details for each eye, like height (palpebral fissure) and iris diameter are also denoted on each frame of the output video.

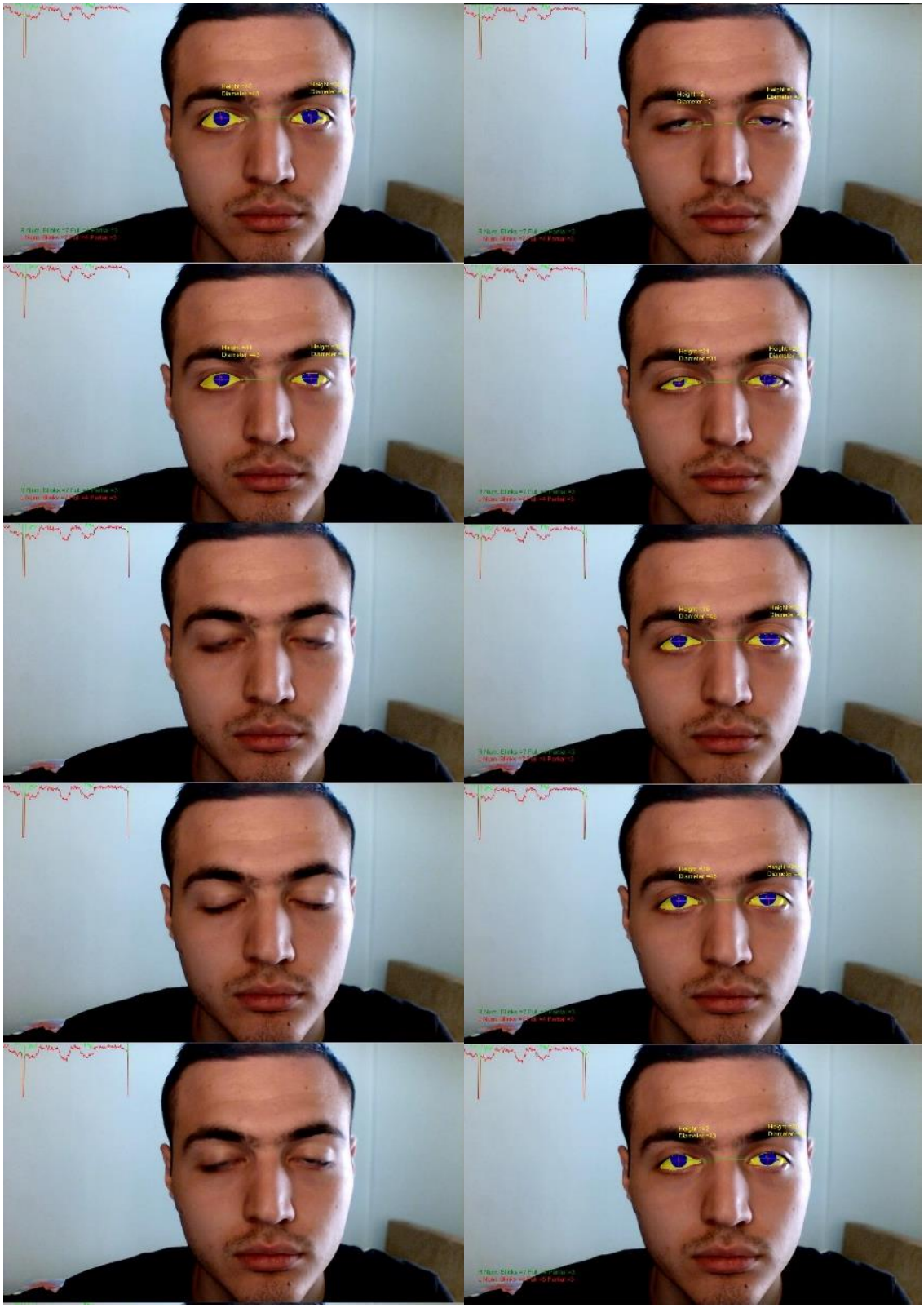


Figure 21: The application of the DLEDs for a blink sequence of a “daylight” video

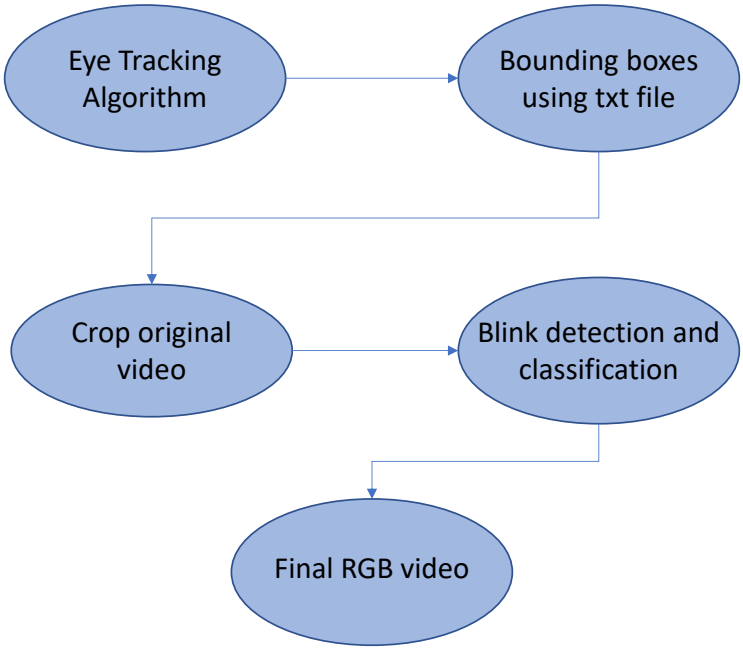


Figure 22: Steps of RGB video creation algorithm

The usage of bounding box allows capturing video from a longer distance than the one of clinical examination. The segmentation result is better, when focusing only on the region of eyes. A typical example of video segmentation is depicted in Figure 21. The bounding box is represented in Figure 23 for a testing frame of the video.

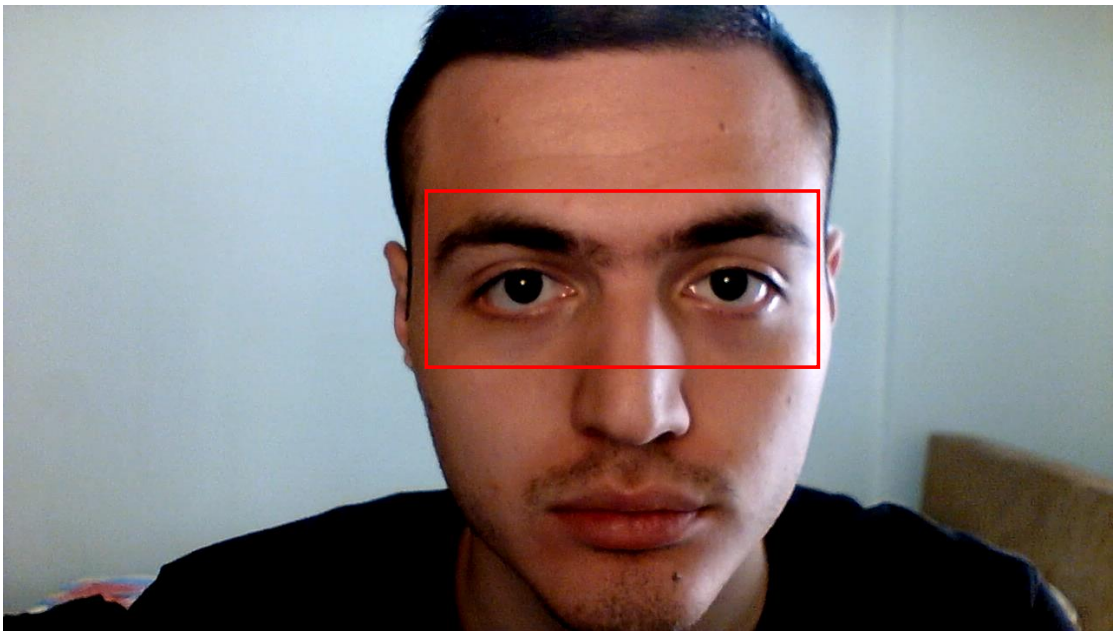


Figure 23: Bounding box is created after the eye tracking algorithm

4 Discussion

The task of eye-blink detection has been tackled, mainly with rather invasive approaches, such as using EOG (electro-oculogram), such as [Divjak, M., & Bischof, H.] and [Jammes, B., Sharabty, H.]. Video analysis is another popular approach that requires no physical contact with the participant and can be immediately applied using a simple camera. Numerous computer vision techniques have been proposed for video-based blink detection [19-28] with a very high accuracy, on several proprietary and public datasets.

The classification of blinks into complete and incomplete has been proposed for assessment of diseases. Among them are the benign eyelid spasm [1], and the Ocular Surface Disease, a disease possibly caused by derangements of the eyelid structure or its secretions [2], leading to extreme dry eyes and hyperosmoticity of tears [3]. Pathological blink patterns have also been linked to Schizophrenia [4], Tourette syndrome [5], or Parkinson's disease [6], which show abnormal pause between blinking, and progressing hypernuclear paralysis, where abnormalities are observed in all types of blinking [7].

However, the classification of eye-blink is a far less researched subject. The only work that deals with complete blink detection is the one proposed by Fogelton and Benesova [29]. The proposed work utilizes video analysis techniques to classify blinks into "complete" and "incomplete". As already described, it uses deep learning (DL) to segment the eyelids and the iris in each frame and a series of post-processing steps to identify and classify blinks. The images of the dataset used to train the DL neural networks for the segmentation were acquired during the clinical examination. Due to the examination setup, the participant' faces were illuminated only by the TV set and the two camera IR LEDs, acquiring high resolution images of the upper half of the face. Thus, our datasets differ significantly from the typical face-datasets that are available for face and blink detection.

The comparison with state-of the-art methods (Table XIII) was performed on a publicly accessed dataset [34], for which the performance of the method of Soukupova [31] and [29] (the only blink classification method) had been reported. Despite the substantially different training set, the proposed method outperformed [29] at blink classification, when each class is considered separately and was only marginally

outperformed by [29] and [31] in terms of blink detection (ignoring blink classification). In addition, in [29] the positions of the eyes are not extracted but manually annotated. Similarly, in [31] six landmarks are utilized around each eye, as provided using [37]. This method requires the whole of the face to be visible and it cannot be applied to our datasets.

The technique employed for the generation of ground truth (three independent medical experts, whose conflict were resolved by a senior expert), enabled the comparison of the proposed system with the human experts in detecting and classifying blinks. As it can be observed in Table XII, the proposed system outperformed at least 1 of the 3 experts in all 8 participants and surpassed all 3 experts in 3 out of 8 participants, in terms of blink detection accuracy. We believe this can be mainly attributed to the usage of thresholds that vary with time and are automatically defined relative to each participant's non-blink eyelid geometry.

Furthermore, results have shown that borderline complete and incomplete blinks may be misclassified both by the proposed system and the human experts. The proposed quantification of the fissure height as a fraction of moving median of the corresponding iris diameter enables the system to handle robustly different eye-types and subject movements and may alleviate the problem of border-line blink classifications. Thus, it could be investigated in the future as an alternative to blink classification. Future work will also include the utilization and validation of the proposed system for non-invasive extraction of blink-related biomarkers for specific ophthalmological and neurological diseases, or even investigate the possibility of utilizing it as a self-assessment tool, in a pervasive computing environment.

5 Conclusion

In this work, we propose an automatic system that detects and classifies blinks from a video sequence acquired using an embedded camera within a close distance to the subject's face. The system utilizes two (2) DLEDs that are trained to segment and detect the iris and eyelids of the eyes. Each segmented frame is post-processed to calculate the iris diameter and the palpebral fissure height of each eye, whose fraction is the main indication for blink detection and classification. The usage of temporal median filtering of the iris diameter and the applied thresholds, with the contribution of the moving median value of the aforementioned fraction of each eye, provides robustness in various

scenarios, such as motion of the patient that changing the orientation of the head, or patient-camera distance during the examination. Results also showed that subjects with different characteristics, such as spectacles, can be handled robustly by the algorithm, even in cases of strong light reflections or actions such as putting spectacles on or taking them off. Also, the proposed system seems to be quite accurate for different testing videos, like normal day-light conditions, in terms of blink detection.

The system was tested on eight (8) participants and the overall blink detection accuracy is compared with the results achieved from three (3) experts, in terms of overall accuracy. The proposed system constantly outperformed at least one, and in certain participants even all of the three experts. Finally, the proposed system was proven competitive, against state-of-the-art methods in blink classification, on a public dataset, quite different than the type of videos that it had been trained and tested upon. Further research is been conducted, where the proposed system is used in order to connect and in long-term help to be taken medical decisions that are related with pathological and neurological diseases of the eye.

6 References

[1] S. A. Hasan et al., "The role of blink adaptation in the pathophysiology of benign essential blepharospasm," Arch. Ophthalmol., vol. 115, no. 5, pp. 631–636, 1997.

[2] A. A. Cruz, D. M. Garcia, C. T. Pinto, and S. P. Cechetti, "Spontaneous eyeblink activity," Ocul. Surf.,. 9(1), pp. 29–41, 2011.

[3] J. P. Craig et al., "TFOS DEWS II definition and classification report," Ocul. Surf., vol. 15, no. 3, pp. 276–283, 2017.

[4] J. R. Stevens, "Eye blink and schizophrenia: psychosis or tardive dyskinesia?," Am. J. Psychiatry, 1978.

[5] D. J. Cohen, J. Detlor, J. G. Young, and B. A. Shaywitz, "Clonidine ameliorates Gilles de la Tourette syndrome," Arch. Gen. Psychiatry, vol. 37, no. 12, pp. 1350–1357, 1980.

[6] N. Kimura et al., "Measurement of spontaneous blinks in patients with Parkinson's disease using a new high-speed blink analysis system," J. Neurol. Sci., vol. 380, pp. 200–204, 2017.

[7] M. Bologna et al., "Voluntary, spontaneous and reflex blinking in patients with clinically probable progressive supranuclear palsy," Brain, vol. 132, no. 2, pp. 502–510, 2009.

[8] W. P. Blount, "Studies of the movements of the eyelids of animals: blinking," Q. J. Exp. Physiol. Transl. Integr., vol. 18, no. 2, pp. 111–125, 1927.

[9] Jie Y, Sella R, Feng J, Gomez ML, Afshari NA. Evaluation of incomplete blinking as a measurement of dry eye disease. Ocul Surf. 2019 Jul;17(3):440-446.

[10] Portello JK, Rosenfield M, Chu CA. Blink rate, incomplete blinks and computer vision syndrome. Optom Vis Sci. 2013 May;90(5):482-7.

[11] Collins MJ, Iskander DR, Saunders A, Hook S, Anthony E, Gillon R. Blinking patterns and corneal staining. Eye Contact Lens. 2006 Dec;32(6):287-93.

[12] Hirosawa K, Inomata T, Sung J, Nakamura M, Okumura Y, Midorikawa-Inomata A, Miura M, Fujio K, Akasaki Y, Fujimoto K, Zhu J, Eguchi A, Nagino K, Kuwahara M, Shokirova H, Yanagawa A, Murakami A. Diagnostic ability of maximum blink interval together with Japanese version of Ocular Surface Disease Index score for dry eye disease. Sci Rep. 2020 Oct 22;10(1):18106..

[13] Ronneberger O., Fischer P., Brox T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N., Hornegger J., Wells W., Frangi A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science, vol 9351. Springer, Cham.

[14] V. Badrinarayanan, A. Kendall and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 2481-2495, 1 Dec. 2017.

[15] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A.L., 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, *40*(4), pp.834-848

[16] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv:1606.00915*, 2016

[17] Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587 (2017)

[18] Chen, Liang-Chieh, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. "Encoder-decoder with atrous separable convolution for semantic image segmentation." In *Proceedings of the European conference on computer vision (ECCV)*, pp. 801-818. 2018.

[19] T. Drutarovsky and A. Fogelton, “Eye Blink Detection Using Variance of Motion Vectors,” in Computer Vision - ECCV 2014 Workshops, Cham, 2015, pp. 436–448.

[20] A. Fogelton and W. Benesova, “Eye blink detection based on motion vectors analysis,” Comput. Vis. Image Underst., vol. 148, pp. 23–33, 2016.

[21] J.-W. Li, “Eye blink detection based on multiple Gabor response waves,” in 2008 International Conference on Machine Learning and Cybernetics, 2008, vol. 5, pp. 2852–2856.

[22] L. Pauly and D. Sankar, “A novel method for eye tracking and blink detection in video frames,” in 2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS), 2015, pp. 252–257.

[23] A. Królak and P. Strumi\l\lo, “Eye-blink detection system for human–computer interaction,” Univers. Access Inf. Soc., vol. 11, no. 4, pp. 409–419, 2012.

[24] I. Choi, S. Han, and D. Kim, “Eye detection and eye blink detection using adaboost learning and grouping,” in 2011 Proceedings of 20th International Conference on Computer Communications and Networks (ICCCN), 2011, pp. 1–4.

[25] S. Al-gawwam and M. Benaissa, “Robust eye blink detection based on eye landmarks and Savitzky–Golay filtering,” Information, vol. 9, no. 4, p. 93, 2018.

[26] I. Bacivarov, M. Ionita, and P. Corcoran, “Statistical models of appearance for eye tracking and eye-blink detection and measurement,” IEEE Trans. Consum. Electron., vol. 54, no. 3, pp. 1312–1320, 2008.

[27] M. Lalonde, D. Byrns, L. Gagnon, N. Teasdale, and D. Laurendeau, “Real-time eye blink detection with GPU-based SIFT tracking,” in Fourth Canadian Conference on Computer and Robot Vision (CRV’07), 2007, pp. 481–487.

[28] A. Panning, A. Al-Hamadi, and B. Michaelis, “A color based approach for eye blink detection in image sequences,” in 2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), 2011, pp. 40–45

[29] A. Fogelton and W. Benesova, “Eye blink completeness detection,” Comput. Vis. Image Underst., vol. 176, pp. 78–85, 2018.

[30] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4510-4520).

[31] Soukupová, Tereza and J. Čech. "Real-Time Eye Blink Detection using Facial Landmarks." (2016).

[32] Bresenham, J.E., 1965. Algorithm for computer control of a digital plotter. *IBM Systems journal*, *4*(1), pp.25-30.

[33] Su, Y., Liang, Q., Su, G., Wang, N., Baudouin, C., & Labbé, A. (2018). Spontaneous eye blink patterns in dry eye: clinical correlations. Investigative ophthalmology & visual science, 59(12), 5149-5156. Blink rate

[34] Talking face video, University of Manchester, available at: https://personalpages.manchester.ac.uk/staff/timothy.f.cootes/data/talking_face/talking_face.html, accessed in 31/3/2021

[35] Divjak, M., & Bischof, H. (2009). Eye blink based fatigue detection for prevention of Computer Vision Syndrome. In *MVA: Proceedings of the 2009 IAPR Conference on Machine Vision Applications* (pp. 350–353). Tokyo: MVA

[36] Jammes, B., Sharabty, H., & Esteve, D. (2008). Automatic EOG analysis: A first step toward automatic drowsiness scoring during wake-sleep transitions. *Somnologie: Schlafforschung und Schlafmedizin, 12,* 227–232.

[37] X. Xiong and F. De la Torre, "Supervised descent methods and its applications to face alignment," inProc. CVPR, 2013.