Journal of Real-Time Image Processing
https://doi.org/10.1007/s11554-022-01200-8

ORIGINAL RESEARCH PAPER

A real-time high precision eye center localizer

Nikolaos Poulopoulos[1] · **Emmanouil Z. Psarakis**[1]

Received: 19 September 2021 / Accepted: 8 January 2022

Abstract
Precise eye center localization remains a very promising but challenging task, while its real-time performance constitutes a critical constraint in many human interaction applications. In this paper a new hybrid framework that combines the shape-based Modified Fast Radial Symmetry Transform (MFRST) and a Convolutional Neural Network (CNN), is introduced. The motivation of this work is to exploit the circularity of the iris to reduce the search space and consequently, the computational complexity of the fed CNN. Thus, the proposed hybrid scheme not only achieves real-time performance, but also increases substantially the localization accuracy by reducing the false detections of the MFRST. The experimental results that stemmed from the most challenging face databases demonstrated high accuracy, outperforming state of the art techniques even those that are based on end-to-end deep neural networks. To deal with unreliable data and provide valid evaluation, we manually annotated the FERET database, making the annotations publicly available. Moreover, the reduced computational time of the proposed scheme reveals that it can be incorporated in low-cost eye trackers, where the real-time performance is a basic prerequisite.

Keywords Real-time performance · Eye localization · Eye tracking · Gaze tracking · Deep neural networks · Convolutional neural networks

1 Introduction

Nowadays, human-computer interaction (HCI) is of growing interest due to the penetration of computer systems in every aspect of everyday life. The most characteristic features of the human face constitute the eyes as they provide significant information for the emotional and cognitive human state. Moreover, eye centers' location per se can be exploited in applications such as face alignment, face recognition, control devices for disabled people, user attention and gaze estimation (e.g., driving and marketing) [1, 2].

Despite the active research in this field, the accuracy of such eye center localization systems has room for improvement and usually downgraded by many limitations. The main challenges are related to the wide variety of human eye colors and shapes, the eye states (open or closed), the facial expressions and orientations etc. Moreover, the presence of occlusions from hair and glasses, reflections and shadows as well as poor lighting and low image resolution further degrades the localization accuracy. Accurate eye center localization becomes even more challenging where the low complexity is substantial for incorporating in real-time applications [3].

End-to-end deep neural network learning constitutes a state-of-the-art approach for solving several problems [4]. However, nowadays the pruning [5] of these Deep architectures has attracted the interest of scientific community, mainly due to their high computational complexity.

Powerful yet unexploited tools for reducing the size and thus simplifying the architecture of the deep network are techniques that can be used in a preprocessing stage for identifying and extracting appropriate features of the under identification objects. This in turn will achieve the desired reduction of the computational complexity of the whole system.

Adopting this approach, in this paper we introduce a novel high precision eye center localization scheme, based on the combination of the MFRST [6] with a shallow CNN. This transform emphasizes on the shape of the iris and combines the edge information that results from an

✉ Emmanouil Z. Psarakis
psarakis@ceid.upatras.gr

Nikolaos Poulopoulos
npoul@ceid.upatras.gr

[1] Department of Computer Engineering and Informatics, University of Patras, Patras, Greece

Published online: 03 February 2022

edge-preserving filtering scheme and the intensity information, to find shapes with high radial symmetry. Moreover, a two-hidden layers CNN was trained, exploiting the false detections of the MFRST, to correct the weaknesses of the localizer and thus increasing the overall performance. Generally speaking, CNNs face efficiency issues and require a large amount of computing resources to be trained properly. The proposed method overcomes this limitation by applying a CNN only to the candidate eye regions that result from the modified FRST. Thus, the input size of the CNN is reduced drastically, from the original size of the images (e.g., 640x480) to small windows (e.g., 28x28, 32x32). This makes the training process computationally efficient even in cases of a large amount of training data, while permits real-time performance.

The main contributions of this work are summarized as follows:

- A novel hybrid eye center localization scheme that combines the MFRST with a shallow CNN.
- A CNN-based approach for reducing the false detections of the MFRST.
- Superior accuracy over state-of-the-art techniques in three public databases.
- Reduced network size (0.15M) that permits real-time performance.
- Publicly available precise eye center annotations of FERET database.

The rest of this paper is organized as follows: Sect. 2 presents a literature review in the area of eye localization. Section 3 describes the proposed method. Section 4 provides the details of the experimental setup and the results, compared with the state-of-the-art methods. Finally, Sect. 5 concludes the paper.

2 Related work

This section provides a brief overview of relevant eye localization methods that have been proposed in the literature. These methods working under challenging conditions can be categorized into the following broad classes:

- feature based methods and
- appearance based methods.

Feature-based methods exploit the special form of the eye structure and detect the eye centers by applying appropriate filters based on shape, geometry, symmetry and color. The obtained features are robust to shape and scale variances and don't require any machine learning techniques. Methods that exploit the circularity of the iris using Hough transform have been employed [7]. Despite their simplicity, its use is limited only in frontal or near frontal images without photometric distortions. Valenti et al. in [8] introduced a voting process based on isophote curvatures for localizing the eye centers, enhanced by a SHIFT descriptor and a k-NN based classifier. Recently this method was improved by Xia et. al. in [9] by combining a regression method, called Supervised Descent Method, with the isophote curvature method. Radial symmetry operators, trying to highlight the circularity of the iris, have also attracted much popularity for the automatic eye center localization [3, 6, 10, 11]. Yang et al. in [12] tried to detect the eye regions by applying Gabor filters and then localize the eye centers using a radial symmetry transform. Skodras et al. in [13] combined the information of color and radial symmetry of the eyes to localize their center locations.

Appearance-based methods incorporate the holistic eye and its surrounding appearance to a prior model and perform eye center localization by fitting this trained model. For this purpose, many machine learning algorithms have been proposed, based on AdaBoost [14], Support Vector Machines (SVMs) [15] and Bayesian models [16].

Convolutional Neural Networks (CNNs) have recently attracted interest, while deep CNNs have achieved several improvements over the feature-based methods, especially in regression [17, 18] and classification [19, 20] problems. Such architectures have been adopted by many eye center localization techniques. Chinsatit and Saitoh [21] proposed a CNN-based pupil center localization method. In [22], Fuhl proposed a coarse to fine pupil localization scheme using two similar CNNs. Li et al. in [23] also proposed a two-stage CNN to determine the most likely eye regions and localize their centers. Xia et al. [24] proposed a heatmap based approach to localize the eye centers using a properly trained shallow FCN with a large kernel convolutional block. In [25] and [26] a deep FCN pipeline was introduced using heterogenous CNN models to detect the face, remove the eye glasses, extract the facial landmarks and finally localize the eye centers. Gou et al. [27] trained a cascade regression model on synthetic images to localize eye centers on real images. In [28] a deep CNN architecture was adopted to regress the eye center coordinates. The proposed technique differs from the aforementioned as it aims to reduce the computational complexity exploiting the shape-based MFRST, instead of training an end-to-end deep network.

3 The proposed method

The proposed technique consists of the following steps: firstly, the face is detected and the two eye Regions of Interest (RoIs) are selected. An edge-preserving filter operator is applied in sequel to enhance the circular shape of the eyes and separate them from the skin. Then, a two-stage MFRST

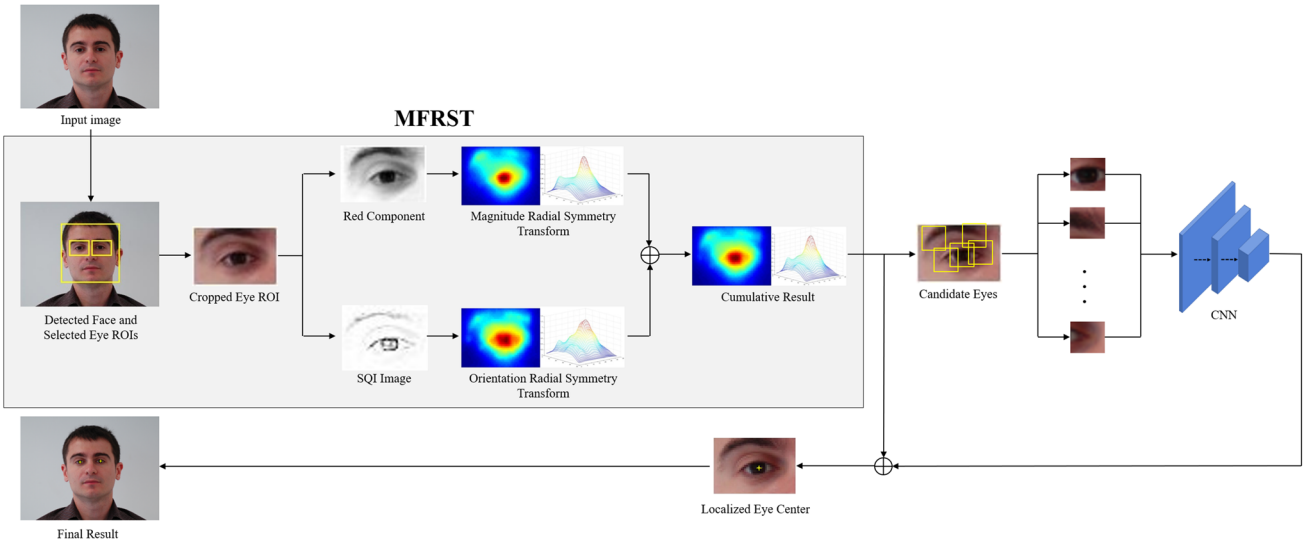

Fig. 1 Overview of the different stages of the proposed system

[6] is used to localize the eye centers. Specifically, The Magnitude-based FRST is applied to the Red component of the image and the Orientation-based FRST to a properly filtered version of the original one. Then, the superposition of their normalized counterparts is used to identify the most likely eyes centers. Finally, a properly trained convolutional neural network is used to classify these candidate eye regions and the results are combined with the MFRST to select the precise eye center positions. The block diagram of the proposed technique is shown in Fig. 1.

3.1 Face detection and eye RoI selection

In the first stage of the proposed pipeline, the face is detected for every input image using the real-time face detector proposed by Viola and Jones [29] as it was described in [6]. In the sequel, the two eye RoIs are selected based on the face geometry [13] (let us denote them by $\mathbb{A}_s,\ s \in \{L,\ R\}$ for the Left and Right eye respectively). The proposed method is then applied to the two cropped eye RoIs, shown in Fig. 1, to minimize the computation time and increase the accuracy of the detected eye centers.

3.2 Edge preserving filtering

3.2.1 Self quotient image

Illumination is considered as one of the main factors that strongly affect the eye localization accuracy. Poor and variable illumination may degrade the ability to distinguish the eyes from the skin. Nowadays this problem has received much attention and many algorithms have been proposed to eliminate this undesirable lighting effect in the image. Such a powerful technique is the well known Self Quotient Image (SQI) [30] technique. This method aims to remove effectively the shadows and compensate the illumination variations, thus, constructing a light invariant representation for every input image. To this end, let $I_\sigma(\boldsymbol{x})$ be a smoothed version of the image $I(\boldsymbol{x})$ that results from its convolution with the isotropic Gaussian kernel $G_\sigma(\boldsymbol{x})$ with subscript σ denoting its standard deviation, that is:

$$I_\sigma(\boldsymbol{x}) = I(\boldsymbol{x}) * G_\sigma(\boldsymbol{x}) \tag{1}$$

where "$*$" denotes the convolutional operator. Then, the Self Quotient Image is defined as follows:

$$Q(\boldsymbol{x}) = \frac{I(\boldsymbol{x})}{I_\sigma(\boldsymbol{x}) + \epsilon}, \tag{2}$$

with ϵ a small constant that is used to avoid undesirable singularities. Note that the standard deviation σ of the Gaussian kernel, controls the width of the edges in the shadows free image defined in (2).

In the next paragraph, a denoising scheme is proposed to eliminate the undesirable amplification of the noise due to the division operation of the SQI.

3.2.2 The proposed denoising scheme

To this end, let as define the following sigmoid function $S : \mathbb{R} \to \mathbb{R}$:

$$S(x) = \frac{1}{1 + e^{-\alpha x}} \tag{3}$$

with $x \in \mathbb{R}$ and α denoting the growth factor that determines its slope. In our experiments we used $\alpha = 15$, since that

Fig. 2 Face images with strong illumination distortions (1-st, 3-rd and 5-th column) and the resulting SQIs (2-nd, 4-th and 6-th column) after the application of the proposed denoising scheme with $\alpha = 15$, and $\sigma = \sigma' = 1$

value minimizes the mean square error of the localization error. Then, the following *two − step* procedure is proposed:

$\mathcal{S}_1$: *Sigma Correction.* Application of the non-linear function $S(.)$ on the SQ Image defined in (2), i.e.:

$$T(\boldsymbol{x}) = S(Q(\boldsymbol{x})) \tag{4}$$

with $Q(\boldsymbol{x})$ denoting the intensity of the SQI at pixel $\boldsymbol{x}$. This step suppresses the undesirable noise of the original SQI providing an enhanced version.

$\mathcal{S}_2$: *Gaussian Smoothing.* Further smoothing of the SQI is performed by applying convolution with a Gaussian kernel with its deviation σ' controlling the strongness of smoothing effect, i.e.:

$$Q_f(\boldsymbol{x}) = T(\boldsymbol{x}) * G_{\sigma'}(\boldsymbol{x}). \tag{5}$$

Following the aforementioned procedure, the SQ images derived from nine (9) face images are depicted in Fig. 2. As it is evident, the filtered images are illumination independent and the eyes can be easily distinguished from the skin.

3.3 Modified fast radial symmetry transform

Symmetry constitutes a key feature of the eyes and can be exploited to highlight their centers effectively. It can be exploited both in the red color component of the original image as well as in the SQ Image defined in (5) [6] by properly modifying the Radial Symmetry Transform [10] which is a low complexity voting procedure that highlights the circular shapes. To this end, let us define the following set of radii:

$$\mathbb{N} = \{r \in \mathbb{Z} : r_{\min} \leq r \leq r_{\max}\}, \tag{6}$$

where $r_{\min}$, $r_{\max}$ are estimated from the under consideration face size as follows:

$$r_{\min} = \max\left\{\frac{\text{FaceWidth}}{60}, 3\right\}, \quad r_{\max} = \frac{\text{FaceWidth}}{6}.$$

Assuming that the gradient $\nabla I(\boldsymbol{x})$ of a given image $I(\boldsymbol{x})$ is known, then for a fixed radius $r \in \mathbb{N}$ and each $\boldsymbol{x}_m \in \mathbb{A}$, we can define the following sets:

$$\mathbb{S}^r_{\boldsymbol{x}_m} = \left\{\boldsymbol{x}_n \in \mathbb{A} : \boldsymbol{x}_n + \text{round}\left(\frac{\nabla I(\boldsymbol{x})}{||\nabla I(\boldsymbol{x})||_2}\right) r = \boldsymbol{x}_m\right\}, \tag{7}$$

where $||\boldsymbol{x}||_2$ denotes the l_2 norm of the vector $\boldsymbol{x}$. Note that some of the above defined sets might be empty.

Let us now define a "Magnitude" $M_r(\boldsymbol{x})$ and an "Orientation" $O_r(\boldsymbol{x})$ Projection Image with support $\mathbb{A}$ each, and with their values in the pixel $\boldsymbol{x}_m$, defined as follows:

$$M_r(\boldsymbol{x}_m) = \sum_{n=1}^{|\mathbb{S}^r_{\boldsymbol{x}_m}|} ||\nabla I(\boldsymbol{x}_n)||_2 \tag{8}$$

$$O_r(\boldsymbol{x}_m) = |\mathbb{S}^r_{\boldsymbol{x}_m}|, \tag{9}$$

with $|\mathbb{X}|$ denoting the cardinality of set $\mathbb{X}$. The above defined projection images are in general different as radius r takes values in the set $\mathbb{N}$ defined in Eq. 6. Finally, the projection images are convolved with a Gaussian kernel $G_{\sigma_r}(\boldsymbol{x})$ and summed over r to form the final result:

$$S_M(\boldsymbol{x}) = \sum_{r=r_{min}}^{r_{max}} M_r(\boldsymbol{x}) * G_{\sigma_r}(\boldsymbol{x}) \tag{10}$$

$$S_O(\boldsymbol{x}) = \sum_{r=r_{min}}^{r_{max}} O_r(\boldsymbol{x}) * G_{\sigma_r}(\boldsymbol{x}), \tag{11}$$

with the standard deviation σ_r of the Gaussian kernel being controlling the smoothing of the magnitude and orientation

component at radius $r \in \mathbb{N}$, and their normalized counterpart are computed as follow:

$$\bar{S}_M(\boldsymbol{x}) = \frac{S_M(\boldsymbol{x})}{\max\limits_{\boldsymbol{x}\in\mathbb{A}}\{S_M(\boldsymbol{x})\}} \tag{12}$$

$$\bar{S}_O(\boldsymbol{x}) = \frac{S_O(\boldsymbol{x})}{\max\limits_{\boldsymbol{x}\in\mathbb{A}}\{S_O(\boldsymbol{x})\}}. \tag{13}$$

Note that in the proposed Modified FRST, the normalization step takes place after the summation of the convolved projection images defined in Eqs (10) and (11) over the set $\mathbb{N}$. In addition, in our approach, the separate use of the magnitude and orientation component of the above defined transform is proposed. In this way, the performance of MFRST was improved as it was pointed out in [3, 6].

Specifically, the contrast existing between the eyes and the skin, was exploited by computing the Magnitude Projection Image of the Red component of the original image. This specific color component was selected due to the usual color of the skin, whose pixel values are higher in this component than in the other two. Thus, a greater contrast between the eyes and the skin can be achieved. Moreover, for distinguishing the eye shape, the Orientation Projection Image was applied to the SQ Image, defined in (2).This component of the MFRST emphasizes on the pixels that form the circular shape of the eye, counting only on their gradient orientation instead of their actual pixel values.

Having computed the above mentioned quantities, in [6] the location of the eye center was proposed to result from the solution of the following optimization problems:

$$\boldsymbol{x}_s^{\star} = \arg\max_{\boldsymbol{x}\in\mathbb{A}_s}\{C_s(\boldsymbol{x})\}, \text{ where} \tag{14}$$

$$C(\boldsymbol{x}) = \bar{S}_{M_s}(\boldsymbol{x}) + \bar{S}_{O_s}(\boldsymbol{x}) \tag{15}$$

and the subscript $s \in \{L, R\}$. Note that the aforementioned proposition was based on the assumption that the certainty of the correct eye center detection, in some sense, is proportional to the MFRST value, that is, the point where the MFRS Transform attains its global maximum is the most likely eye center.

3.4 Candidate eye centers

However, we observed that in extreme cases of photometric distortions or/and occlusions the true eye center position may differs from the global maximum of the MFRST. To overcome this problem and increase the accuracy of the proposed method, instead of using the global maximum (i.e., the solution of the optimization problem (14)) as the final eye center position, a set of candidate eyes centers defined by:

$$X_s(T) = \left\{\boldsymbol{x} \in \mathbb{R}_s : C(\boldsymbol{x}) > T\right\}, \tag{16}$$

with the subscript $s \in \{L, R\}$, is proposed to feed the input of a CNN for further investigation. It is clear, that the cardinality N_s of this set will be strongly depended on the value of threshold T which in turn should be depended on the strongness of the aforementioned distortions. Higher values of this threshold limit the candidate eye centers and thus lead to lower maximum accuracy. On the other hand, lower values increase the maximum possible accuracy but also increase the computational time. The dependency between the threshold and the maximum possible accuracy for the MUCT database is depicted in Fig. 3. This leads us to the selection of the value $T = 0.9$, as it satisfies the requirement for higher accuracy with the higher threshold.

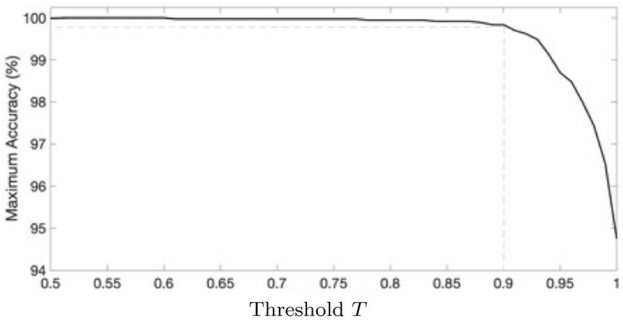

Fig. 3 The maximum possible accuracy drops as the threshold increases. The value $T = 0.9$ satisfies the requirement for higher accuracy with the higher threshold

3.5 Convolutional neural network

It is well known that CNNs achieve high performance in classification problems [4]. To exploit that point we reformulate the eye center localization problem into a classification one. In addition, the use of MFRST reduces dramatically the search space of the candidate eye centers from the total number of pixels in the input image to few positions (e.g., 2,3), reducing in this way the computational cost of the CNN. Moreover, the CNN is able to correct false localizations of MFRST and thus improving the accuracy of the proposed method vs state of the art localization methods.

3.5.1 Training set

Convolutional Neural Networks working with high resolution face images require a large amount of computing resources during the training process. The proposed architecture exploits the MFRST in order to drastically reduce

Fig. 4 Examples of eye region (left) and non-eye region (right) images with size 32x32 pixels, that form the two training datasets of the CNN

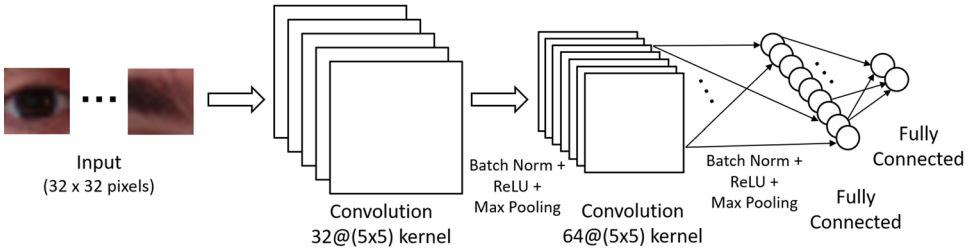

Fig. 5 Structure of the CNN

the dimensionality of the input images and thus reducing the computational burden of the training phase. In particular, the CNN has to classify every candidate image as an eye or non-eye region. The selection of the training set affects the performance of the classifier. By taking into account that the input images of the CNN are the local maxima of the MFRST (Fig. 1), we exploit the inaccurate localizations of the MFRST to select the non-eye region images of the training set. In this way, we succeed to train a network that corrects the localization errors of the MFRST and thus increases the accuracy of the proposed method.

To build up the training database, the following normalized error [31], representing the worst eye center estimation of the two eyes, is adopted:

$$e = \frac{\max\{||\hat{C}_L - C_L||_2,\ ||\hat{C}_R - C_R||_2\}}{||C_L - C_R||_2} \tag{17}$$

where, $\hat{C}_L$, $\hat{C}_R$ are the estimated left and right eye center coordinates respectively and C_L, C_R are the corresponding ground-truth. The term $||C_L - C_R||_2$ constitutes a normalization factor representing the actual distance between the eye centers. The resulting accuracy is expressed as the percentage of the eye center localizations that fall below the assigned error threshold. Points with $e \leq 0.25$ belong to a disk area that extends from the eye center to the eye corners (lacrimal caruncle), points with $e \leq 0.1$ belong to the iris area while points with $e \leq 0.05$ belong to the pupil area. We are going to use the same error as a performance measure in the next section. But for the moment let us build up the training database as follows:

– $\mathcal{S}_1$: The eye-region set (or category) is created using the ground-truth of each face database. To contain the entire eye inside the image, a square of size $d \times d$, with $d = \text{FaceHeight}/15$, is cropped with each eye coordinates in its center (Fig. 4).
– $\mathcal{S}_2$: The non-eye region set is created using the local maxima of the MFRST that differ significantly from the ground-truth ($e \geq 0.15$). Then, the candidate regions are cropped as previously. It is evident from Fig. 4, that the majority of these images includes parts from eyebrows, eyelids, eye corners and glass frames.

3.5.2 Network architecture

The proposed architecture aims to classify an input image by estimating its probability to be eye region. The core architecture of the convolutional neural network is summarized in Fig. 5. It consists of two convolutional layers with 32 and 64 kernels, followed by batch normalization and rectified linear layers. Finally, two consecutive fully connected (FC) layers with 32 and 2 neurons are used to estimate the probabilities of each class. By taking into account the special form (i.e., binary) of the classification problem we have to solve, for the training of the aforementioned CNN we adopted the binary cross-entropy loss function [4] between the ground truth and the above mentioned probabilities. The proposed network was trained using the ADAM optimizer [32] for 25 epochs, with initial learning rate of 0.001 and batch size of 100 images. Moreover, to prevent the network from overfitting during training we employ L_2 regularization of weights with coefficient of 0.005.

Although we could use the trained CNN, fed by the MFRST scheme, as a standalone eye localizer system, in the next subsection we are going to fuse its output with that of the MFRST scheme to maximize the accuracy of the proposed localizer.

3.6 Optimization problem

To this end, let us consider the following N_s $(n = 1, \cdots, N_s)$ events:

$$\mathcal{E}_n = \left\{ \boldsymbol{x}_s(n) \in X_s(T) : \ \boldsymbol{x}_s(n) \text{ is a candidate eye center} \right\}, \tag{18}$$

and let us denote as $\mathbb{P}_{\text{CNN}}(\mathcal{E}_n)$, $n = 1, \cdots, N_s$ the probability assigned by the CNN to each one of the above defined events. In addition, let us define the corresponding probabilities assigned by the MFRST using the softmax function as follows:

$$\mathbb{P}_{\text{MFRST}}(\mathcal{E}_n) = \frac{C(\boldsymbol{x}_s(n))}{\sum_{k=1}^{N_s} C(\boldsymbol{x}_s(k))}, \ n = 1, \cdots, N_s \tag{19}$$

where $C(\boldsymbol{x})$ is defined in Eq. (15).

Then, the probability assigned by the CCN to every candidate eye center is combined with the result from the MFRST in an convex way, through the hyperparameter β, taking values in the interval [0, 1]. Finally, the dominant eye center positions result from the solution of the following optimization problem:

$$\mathcal{E}_{n^\star} = \arg\max_{n \in N_s} \left\{ (1-\beta)\mathbb{P}_{\text{MFRST}}(\mathcal{E}_n) + \beta \mathbb{P}_{\text{CNN}}(\mathcal{E}_n) \right\} \tag{20}$$

with the subscript $s \in \{L, R\}$. The optimum value of the hyperparameter β is examined in the next section.

4 Experiments

4.1 Experimental Setup

Experiments were carried out on three publicly available face databases for fair comparison and the performance of the proposed scheme was extensively evaluated and compared with the state-of-the-art. Specifically, the selected MUCT [33], BioID [31] and FERET [34] databases were widely used by well-known in the bibliography eye center localization techniques while they were regarded as extremely challenging in terms of degradations.

The MUCT database comprises 3755 color images of low resolution (640×480 pixels) with frontal or near frontal faces. These images include a wide variety of degradations related to pose and lighting variations as well as occlusions from hair, glasses and reflections.

The BioID database consists of 1521 grayscale images of low resolution (384×288 pixels) and is regarded as one of the most challenging databases as it contains wide scale and pose variations while many subjects are wearing glasses or their eyes were closed or hidden by strong reflections on glasses. To explore the eye center localization task, 29 images that contain totally closed eyes were manually removed.

The FERET database contains 11338 images including 994 subjects with image resolution of 512×768 pixels. In our experiments, in order to compare the proposed method to the state-of-the-art techniques, a partition of this database containing 2636 frontal face (fa) and alternate frontal face (fb) images was considered. The subjects contain a wide variety of ages and ethnicities with many degradations from glasses and occlusions.Although the color FERET database contains high resolution images, the ground-truth annotation is sometimes unreliable, as it was also reported in [8, 13]. It must be stressed at this point, that on several images the eyes are annotated anywhere within the iris area and not in their true center. To overcome this limitation and provide valid evaluation of the proposed method, we manually annotated the eye center positions. These annotations are publicly available.[1]

4.2 Experimental results

The evaluation of the proposed method demonstrates that it is highly accurate and robust under many challenging situations including shadows, pose and scale variations as well as occlusions by hair, glasses or strong reflections (Fig. 6). The localization accuracy degrades only in cases when the eyes are totally closed and in extreme cases of irregular illuminations, shadows and occlusions. Moreover, the degraded accuracy of the face detector to non-frontal images constitutes a significant limitation to the overall performance. To investigate how the CNN behaves on totally new images, a 5-fold cross validation was adopted.

4.2.1 Comparisons against different architectures

Three different CNN architectures with a different number of convolutional layers were evaluated to investigate their impact on the performance of the proposed method. All CNNs were trained and tested in BioID [31] database and the obtained results are presented in Table 1. The CNN with two convolutional layers outperformed, in terms of accuracy, the other two architectures thus justifying our choice to select it as the proposed CNN architecture. Note that, due to the small number and size of the training images, as the number of convolutional layers increases, the probability of overfitting the network also increases. The training and validation curves depicted in Fig. 7 reveal that the selected CNN is properly fitted on the data.

[1] http://xanthippi.ceid.upatras.gr/people/psarakis/publications/feret-ground-truth.zip.

Fig. 6 Precise eye center localization results in MUCT (first row), FERET (second row) and BioID (third row) databases

Table 1 Accuracy vs. normalized error in BioID for three different CNN architectures

CNN architecture	Accuracy (%)		
	$e \leq 0.05$	$e \leq 0.1$	$e \leq 0.25$
1-layer	95.93	99.79	**100**
2-layers	**96.65**	**100**	**100**
3-layers	95.15	99.86	**100**

The best performances are indicated in bold

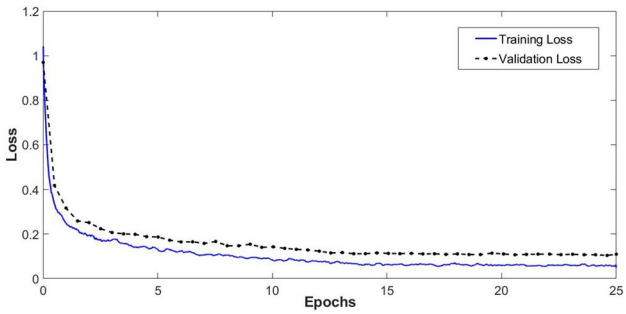

Fig. 7 Learning curves of the used CNN during its training (blue) and validation (dotted black) respectively

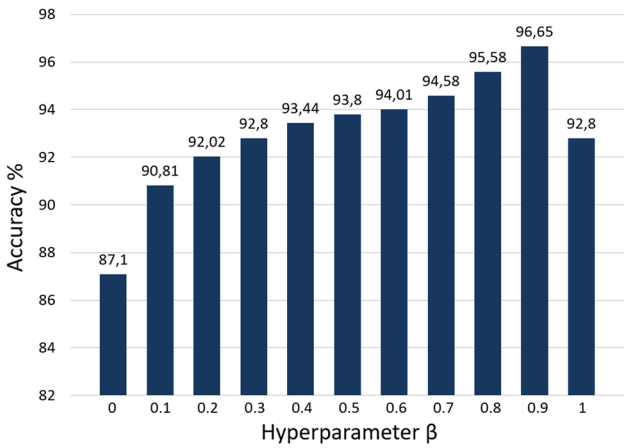

Fig. 8 Accuracy vs. hyperparameter β in the BioID database

4.2.2 Comparisons against different architectures

We also investigate the significance of the convex combination between the MFRST and CNN (please see Eq. (20) in the manuscript). Figure 8 presents the localization accuracy in respect to hyperparameter β for the BioID database. The optimum performance is achieved for $\beta = 0.9$ and thus this value is adopted in the experiments we have conducted. Note that both the achieved accuracies from the use of the MFRST scheme ($\beta = 0$) and the CNN alone ($\beta = 1$), are inferior to the optimum one. Specifically, as it is evident from Fig. 8, the proposed hybrid scheme substantially improves the accuracy by 9.55% as it is compared with that achieved by the MFRST scheme.

4.2.3 Comparisons against the state of the art methods

The proposed method was compared with the state-of-the-art techniques and the results are presented in this subsection. Except from the proposed method, which takes into account the CNN to enhance the localization accuracy, the scenario where the eye centers result directly from the global maximum of the MFRST is also examined. This scenario,

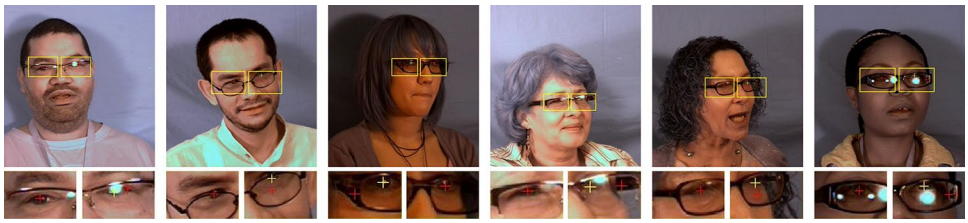

Fig. 9 MFRST localizations (red cross) and the corrected ones using CNN (yellow cross)

Table 2 Accuracy vs. normalized error in the MUCT database

Method	Accuracy (%)		
	$e \leq 0.05$	$e \leq 0.1$	$e \leq 0.25$
Proposed	**96.71**	**99.70**	**99.95**
Yolo v3 (2018) [18]	94.75	98.28	99.34
MFRST (2017) [6]	94.41	98.67	99.76
Faster R-CNN (2015) [17]	94.15	97.48	99.34
Skodras (2015) [13]	92.9	97.2	99.0
Valenti (2012) [8]	63.1	76.7	94.1
Timm (2011) [36]	78.6	94.9	98.6

The best performances are indicated in bold

Table 3 Accuracy vs. normalized error in the FERET database

Method	Accuracy (%)		
	$e \leq 0.05$	$e \leq 0.1$	$e \leq 0.25$
Proposed	**99.47**	**99.81**	**100**
Yolo v3 (2018) [18]	98.22	99.01	99.59
MFRST (2017) [6]	98.32	99.12	99.77
Faster R-CNN (2015) [17]	97.89	98.43	97.54
Skodras (2015) [13]	98.43	99.08	99.81
Timm (2011) [36]	93.92	97.40	99.16

The best performances are indicated in bold

Table 4 Accuracy vs. normalized error in the BioID database

Method	Accuracy (%)		
	$e \leq 0.05$	$e \leq 0.1$	$e \leq 0.25$
Proposed	96.65	**100**	**100**
Lee (2020) [26]	**96.71**	98.95	**100**
Choi (2020) [25]	93.30	96.91	**100**
Ahmed N.Y. (2020) [37]	91.68	97.85	98.66
Xia (2020) [9]	88.10	98.80	**100**
Ahmed M (2019)[38]	86.98	95.20	99
Gou (2019)[27]	92.30	99.10	–
Xia (2019) [24]	94.40	99.90	**100**
Xiao (2018) [39]	94.35	98.75	99.80
Li (2018) [23]	85.60	95.90	99.50
Wang (2018)[40]	82.15	98.70	**100**
Yolo v3 (2018) [18]	93.28	98.63	99.93
Araujo (2017) [41]	88.30	92.70	98.90
MFRST (2017) [6]	87.10	98.15	**100**
Anjith (2016) [42]	85.00	94.30	–
Cai (2015) [43]	84.10	95.60	99.80
Faster R-CNN (2015) [17]	90.51	96.56	99.03
Markus (2014) [44]	89.90	97.10	99.70

The best performances are indicated in bold

denoted hereafter as MFRST method, is examined as an alternative solution where the use of the training dataset is not feasible. This method generally achieves slightly lower accuracy rates; however, it doesn't require training (Fig. 9). We also trained the state-of-the-art deep convolutional networks Faster R-CNN [17] and Yolo v3 [18] using the convolutional layers of the Alexnet [35] as their feature extraction networks. We fed their inputs with the RoI of the images as they were defined in Eqs. (1) and (2) and the eye centers derived from the centers of the detected eye boundaries.

All the following tables confirm the enhanced accuracy of the proposed method over the state-of-the-art. Table 2 contains the results obtained for the application of the proposed method as well as its state of the art rivals in the MUCT database. The proposed technique achieved a significant improvement of 2.3% in performance over the best method for the fine accuracy level ($e \leq 0.05$). It failed to localize the eye centers precisely only in extreme cases where the eyes were partially occluded by hair or reflections on the glasses.

In Table 3 the accuracy achieved by the proposed method in FERET database is shown. In this case, the proposed method achieved to localize precisely the eye centers in almost every image mainly due to the absence of occlusions and irregular illuminations. To evaluate the accuracy of the compared methods in relation to our ground-truth data, we used their algorithm implementations.

Finally, the performance of the proposed method in the low resolution images of BioID database is presented in Table 4. The proposed technique achieved almost equal performance to the best method [25] for the fine accuracy level ($e \leq 0.05$), while outperformed its rivals in all other cases. It is worth mentioned that despite the presence of challenging images in terms of degradations, the proposed hybrid scheme was the only one that achieved 100% accuracy for the case of $e \leq 0.1$.

Table 5 Illumination invariance in MUCT database for different accuracies

Illumination groups	Accuracy (%)		
	$e \leq 0.05$	$e \leq 0.1$	$e \leq 0.25$
A (1857 images)	96.72	99.62	99.95
B (1821 images)	95.55	99.56	99.95

The above mentioned results lead us to the conclusion of a significant improvement of the proposed method against other state-of-the-art methods. Moreover, the use of the CNN improves the accuracy of the localization up to 9.55% for the BioID database. We explored also the impact of training the proposed method in MUCT database and testing in FERET. In this case, the accuracy decrease was less than 1%, achieving 98.71% and 99.50% for the cases of $e < 0.05$ and $e < 0.1$ respectively, demonstrating robustness of the proposed technique to unseen images.

Finally, the Yolo v3 demonstrates enhanced accuracy over the Faster R-CNN network, especially in lower resolution images. However, despite their high detection rates, the proposed method outperforms both of them in terms of localization accuracy, revealing the advantage of the proposed hybrid scheme compared to deep end-to-end regression approaches.

4.2.4 Pose and illumination invariance

The quantitative evaluation of the robustness of the proposed method to pose and lighting variations is performed using the MUCT database, which provides separation of the images into groups of different lighting and pose. The lighting subsets were divided into two groups A and B, containing the most and least well illuminated images respectively. The results of the accuracy of each illumination group that reported in Table 5 confirm the independency of the proposed method to lighting variations. Specifically, the maximum deviation is 1.17 for the error distance $e \leq 0.05$, while for the other two error distances the results differ imperceptibly.

The arrangement of 5 cameras, as described in [33], allowed the categorization of all subjects in 5 pose categories, that is neutral pose (0^o), $+20^o$ and $+38^o$ yaw angles, $+21^o$ and -22^o pitch angles, each containing 751 images. Table 6 contains the performance of the proposed method in terms of accuracy for the five pose categories in the MUCT database.

It is evident that the accuracy is not affected for 0, + 22 yaw, −22 pitch angles, as the differences are negligible (maximum difference is 0.53 for $e \leq 0.05$). The slightly reduced accuracy for the cases of $+38^o$ yaw and $+21^o$ pitch can be explained since for such great rotation angles, the person's eyes get distorted and the nose, glasses and hair can cover a portion of the eye.

Table 6 Pose invariance in MUCT database for different accuracies

Pose	Accuracy (%)		
	$e \leq 0.05$	$e \leq 0.1$	$e \leq 0.25$
0°	97.05	99.87	100
+22° yaw	97.58	99.73	100
+38° yaw	96.70	99.00	99.28
−22° pitch	97.31	99.87	100
+21° pitch	95.06	99.07	100

The reduced accuracy of $+38^o$ yaw and $+21^o$ pitch is resulted from external factors

Table 7 Processing time in BioID database

Method	Proposed	Choi (2020) [25]	Ahmed N.Y. (2020) [37]	Gou (2019) [27]	Araujo (2017) [41]
Time (ms)	**25**	28	30	62.5	83.4

The best performance is indicated in bold

4.2.5 Real time performance

Real-time performance constitutes a critical limitation of many human-computer interaction applications such as the eye localization and tracking. To investigate the performance of the proposed method to such practical applications, a low-level C_{++} implementation using OpenCV library was evaluated on a single core intel i7 system.

The computation time of the proposed technique as well as other state of the art works (that were tested by their authors themselves) in BioID database is presented in Table 7. Although the face detector consumes large computational time, it can be significantly reduced to 15ms by following the procedure proposed in [37]. The proposed method requires approximately 10ms to process every cropped face image, where 3ms on average correspond to the CNN and the rest to the MFRST. Thus, the total processing time is approximately 25ms, which is significantly reduced compared to the rest of methods. We have also tested our technique in MUCT and FERET databases that contain images of higher resolution and the required computation time was approximately 41.8ms and 80ms respectively. Although the computation time of the MFRST depends on the RoI size, the accuracy improvement saturates when the resolution of the face exceeds the 150x150 pixels. This, permits us to rescale the higher resolution images to this

Table 8 The effect of each component of the proposed method to the localization accuracy

Components	Accuracy (%)		
	$e \leq 0.05$	$e \leq 0.1$	$e \leq 0.25$
Magnitude	82.75	95.44	99.93
Magnitude + orientation	87.10	98.15	**100**
Magnitude + Orientation +CNN	**96.65**	**98.95**	**100**

The best performances are indicated in bold

size, preserving a constant computation time of 29ms with a negligible accuracy decrease.

Moreover, the proposed selection of the training set reduces dramatically the training time. Specifically, the training process takes approximately 30 seconds to train 500 images, in contrast to Li's method [23] which takes 4 hours.

In terms of computational complexity, the MFRST presents a relatively low complexity $\mathcal{O}(KN)$, depending linearly on the size of the RoI K and the set of the radii N. Moreover, the proposed CNN contains only $0.15M$ parameters and is significantly reduced in comparison with other deep networks. Specifically, the architectures proposed in [26] and [25] contain $13.6M$ and $4.9M$ respectively only for the face detection and glasses removal networks, without considering the eye localization network.

4.2.6 Ablation Study

In this section, to further highlight the contributions of the proposed hybrid scheme, we analyze the effects of its components to the localization accuracy in BioID database. Specifically, Table 8 presents the localization accuracy under three scenarios. In the first case, the eye centers derive directly from the Magnitude component of the MFRST. In the second case, the Magnitude and Orientation components of the MFRST are combined to localize the eye centers, while in the third one, the MFRST is combined with the CNN to form the proposed hybrid scheme. Results demonstrate that each module provides a significant accuracy improvement, highlighting its importance to the overall performance.

5 Conclusion

In this paper, a new, high precision, real-time eye center localization technique was introduced. This hybrid method exploits the circular shape of the iris using a modified version of the FRST and reduces the false negatives using a properly trained CNN. An extensive evaluation of the proposed scheme was performed on three publicly available databases with varying resolution images, containing many different cases under challenging conditions. From a series of experiments we conducted, the proposed technique outperformed its state of the art rivals, demonstrating a significant improvement in terms of accuracy. As a future work, the proposed method could be adopted to enhance the gaze estimation performance and incorporated into low-cost gaze trackers.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Kar, A., Corcoran, P.: A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms. IEEE Access **5**, 16495–16519 (2017)
2. Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., Torralba, A.: Eye tracking for everyone. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2176–2184 (2016)
3. Poulopoulos, N., Psarakis, E. Z.: Real time eye localization and tracking. In: Proceedings of the 27th International Conference on Robotics in Alpe-Adria Danube Region, RAAD, Patras (2018)
4. Ian, G., Yoshua, B., Aaron, C.: Deep learning. MIT Press, London (2016)
5. Ghosh, S., Srinivasa, S. K. K., Amon, P., Hutter, A., Kaup, A.: Deep network pruning for object detection. In: IEEE International Conference on Image Processing (ICIP), pp. 3915–3919 (2019)
6. Poulopoulos, N., Psarakis, E. Z.: A new high precision eye center localization technique. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 2806–2810 (2017)
7. Dobes, M., Martinek, J., Skoupil, D., Dobesova, Z., Pospisil, J.: Human eye localization using the modified Hough transform. Opt. Int. J. Light Electron Opt. **117**(10), 468–473 (2006)
8. Valenti, R., Gevers, T.: Accurate eye center location through invariant isocentric patterns. IEEE Trans. Pattern Anal. Mach. Intell. **34**(9), 1785–1798 (2012)
9. Xia, Y., et al.: Hybrid regression and isophote curvature for accurate eye center localization. Multimed. Tools Appl. **79**, 805–824 (2020)
10. Loy, G., Zelinsky, A.: Fast radial symmetry for detecting points of interest. IEEE Trans. Pattern Anal. Mach. Intell. **25**, 959–973 (2003)
11. Bai, L., Shen, L., Wang, Y.: A novel eye location algorithm based on radial symmetry transform. IEEE Int. Conf. Pattern Recognit. **3**, 511–514 (2006)
12. Yang, P., Du, B., Shan, S., Gao, W.: A novel pupil localization method based on GaborEye model and radial symmetry operator. IEEE Int. Conf. Image Process. **1**, 67–70 (2004)
13. Skodras, E., Fakotakis, N.: Precise localization of eye centers in low resolution color images. J. Image Vis. Comput. **31**, 51–60 (2015)
14. Niu, Z., Shan, S., Yan, S., Chen, X., Gao, W.: 2D cascaded adaboost for eye localization. In: IEEE International Conference on Pattern Recognition (ICPR'06) (2006)

15. Campadelli, P., Lanzarotti, R., Lipori, G.: Precise eye localization through a general-to- specific model definition. BMVC **1**, 187–196 (2006)
16. Everingham, M., Zisserman, A.: Regression and classification approaches to eye localization in face images. In: IEEE 7th International Conference on Automatic Face and Gesture Recognition, pp. 441–446 (2006)
17. Shaoqing, R., Kaiming, H., Ross, G., Jian, S.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp. 91–99. Springer, Berlin (2015)
18. Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement, arXiv:1804.02767 (2018)
19. Hong, D., Gao, L., Yokoya, N., Yao, J., Chanussot, J., Du, Q., Zhang, B.: More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. IEEE Trans. Geosci. Remote Sens. **59**, 4340–4354 (2021)
20. Hong, D., Gao, L., Yao, J., Zhang, B., Plaza, A., Chanussot, J.: Graph convolutional networks for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. **59**, 5966–5978 (2021)
21. Chinsatit, W., Saitoh, T.: CNN-based pupil center detection for wearable gaze estimation system, applied computational intelligence and soft computing, pp. 1–10. Springer, Berlin (2017)
22. Fuhl, W., Santini, T., Kasneci, G., Kasneci, E.: PupilNet: convolutional neural networks for robust pupil detection, arXiv:1601.04902 (2016)
23. Li, B., Fu, H.: Real time eye detector with cascaded convolutional neural networks. Appl. Comput. Intell. Soft Comput. (2018). https://doi.org/10.1155/2018/1439312
24. Xia, Y., Yu, H., Wang, F.: Accurate and robust eye center localization via fully convolutional networks. IEEE/CAA J. Autom. Sin. **6**, 1127–1138 (2019)
25. Choi, J.H., Lee, K.I., Song, B.C.: Eye pupil localization algorithm using convolutional neural networks. J. Multimed. Tools Appl. **79**, 32563–32574 (2020)
26. Lee, K.I., Jeon, J.H., Song B.C.: Deep learning based pupil center detection for fast and accurate eye tracking system. In: Springer European Conference on Computer Vision (ECCV), pp. 1127–1138 (2020)
27. Gou, C., Zhang, K., Wang, K., Wang, F., Ji, Q.: Cascade learning from adversarial synthetic images for accurate pupil detection. J. Pattern Recognit. **88**, 584–594 (2019)
28. Larumbe-Bergera, A., Garde, G., Porta, S., Cabeza, R., Villanueva, A.: Accurate pupil center detection in off-the-shelf eye tracking systems using convolutional neural networks. Sensors **21**, 6847 (2021)
29. Viola, P., Jones, M.: Robust real-time face detection. Int. J. Comput. Vis. **57**, 137–154 (2004)
30. Wang, H., Li, S.Z., Wang, Y., Zhang, J.: Self quotient image for face recognition. ICIP **2**, 1397–1400 (2004)
31. Jesorsky, O., Kirchbergand, K. J., Frischholz, R.: Robust face detection using the hausdorff distance. In: Audio and Video Biom. Pers. Authentication, pp. 90–95 (2001)
32. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: IEEE International Conference on Learning Representations (2015)
33. Milborrow, S., Morkel, J., Nicolls, F.: The MUCT landmarked face database. Pattern Recognit. Assoc. S. Afr. **201**, 20 (2010)
34. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The FERET evaluation methodology for face-recognition algorithms. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **22**(10), 1090–1104 (2000)
35. Krizhevsky, A., Sutskever, I., Hinton, G.: imagenet classification with deep convolutional neural networks. In: International Conference on Neural Information Processing Systems (NIPS), pp. 1097–1105 (2012)
36. Timm, F.E.: Barth, accurate eye center localization by means of gradients, VISAPP, pp. 125–130 (2011)
37. Ahmed, N.Y.: Real-time accurate eye center localization for low-resolution grayscale images. J. Real-Time Image Process. **18**(1), 193–220 (2021)
38. Ahmed, M., Laskar, R.H.: Eye center localization in a facial image based on geometric shapes of iris and eyelid under natural variability. Elsevier J. Image Vis. Comput. **88**, 52–66 (2019)
39. Xiao, F., Huang, K., Qiu, Y.H.: Shen, Accurate iris center localization method using facial landmark, snakuscule, circle fitting and binary connected component. J. Multimed. Tools Appl. **77**, 25333–25353 (2018)
40. Wang, Z., Cai, H., Liu, H.: Based, robust eye center localization on an improved SVR method. In: Neural information processing. ICONIP, lecture notes in computer science, 11307th edn. Springer, Berlin (2018)
41. Araujo, G., Ribeiro, F., Junior, W., da Silva, E., Goldenstein, S.: Weak classifier for density estimation in eye localization and tracking. IEEE Trans. Image Process. **26**, 3410–3424 (2017)
42. George, A., Routray, A.: Fast and accurate algorithm for eye localization for gaze tracking in low-resolution images. IET Comput. Vis. **10**(7), 660–669 (2016)
43. Cai, H.-Bi, et al.: Convolution-based means of gradient for fast eye center localization. IEEE International Conference on Machine Learning and Cybernetics (ICMLC) (2015)
44. Markus, N., Frljak, M., Pandzic, I.S., Ahlberg, J., Forchheimer, R.: Eye pupil localization with an ensemble of randomized trees. Pattern Recognit. **47**(2), 578–587 (2014)

Nikolaos Poulopoulos was born in Patras, Greece in 1991. He received his B.Sc. degree from the Hellenic Airforce Academy in 2013 and his M.Sc. from the Department of Computer Engineering and Informatics, University of Patras, in 2017, where he is currently pursuing his Ph.D. in computer vision. His current research interests include image processing, computer vision and convolutional neural networks.

Emmanouil Z. Psarakis received the Ph.D. degree form the Department of Computer Engineering and Informatics, University of Patras, Rio-Patras, Greece, in 1991. From 1994 to 1996, he held a research position with the Computer Technology Institute of Patras, Greece. Since 2017, he has been an Associate Professor with the Department of Computer Engineering and Informatics, University of Patras, and a member of the Signal Processing and Communications Laboratory. His interests include biomedical signal processing, medical image analysis, seismological signal processing, computer vision and neural networks. Dr. Psarakis has served as a reviewer of many international technical journals such as IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Signal Processing, IEEE Transactions on Circuits and Systems, Elsevier Signal Processing, Elsevier Signal Processing Fast Communication, and EURASIP Journal on Applied Signal Processing.