Khan, W, Hussain, A, Kuru, K and Al-askar, H

Pupil Localisation and Eye Centre Estimation using Machine Learning and Computer Vision

Type of the Paper (Article)

Pupil Localisation and Eye Centre Estimation using Machine Learning and Computer Vision

Wasiq Khan [1,*], Abir Hussain [2], Kaya Kuru [3], and Haya Al-askar [4,]

[1] Liverpool John Moores University, Liverpool, UK; W.khan@ljmu.ac.uk
[2] Liverpool John Moores University, Liverpool, UK; A.Hussain@ljmu.ac.uk
[3] University of Central Lancashire, Preston, UK; KKuru@uclan.ac.uk
[4] Prince Sattam Bin Abdulaziz University, Saudi Arabia; H.Alaskar@psau.edu.sa

* Correspondence: W.Khan@ljmu.ac.uk;

Abstract: Various methods have been used to estimate the pupil location within an image or a real-time video frame in many fields. However, these methods lack the performance specifically in low-resolution images and varying background conditions. We propose a coarse-to-fine pupil localisation method using a composite of machine learning and image processing algorithms. First, a pre-trained model is employed for the facial landmark identification to extract the desired eye-frames within the input image. We then use multi-stage convolution to find the optimal horizontal and vertical coordinates of the pupil within the identified eye-frames. For this purpose, we define an adaptive kernel to deal with the varying resolution and size of input images. Furthermore, a dynamic threshold is calculated for reliable identification of the best-matched candidate. We evaluated our method using various statistical and standard metrics along-with a standardized distance metric we introduce first time in this study. Proposed method outperforms previous works in terms of accuracy and reliability when benchmarked on multiple standard datasets. The work has diverse artificial intelligence and industrial applications including human computer interfaces, emotion recognition, psychological profiling, healthcare and automated deception detection.

Keywords: Pupil detection; Deep eye, Iris detection; Eye centre localisation; Eye gaze; Facial analysis, Image convolution; Machine intelligence, Pupil segmentation

1. Introduction

Detection and localization of the objects within images or real time video frames is considered an essential task in various computer vision algorithms [1]. Various studies have addressed the detection and tracking of facial landmarks including the iris and pupil which has various applications particularly, eye gaze estimation for human-machine interfaces. Control of assistive devices for disability [2], driver safety improvements [3-4], the design of diagnostic tools for brain diseases [5], cognitive research [6], automated deception detection system (ADDS) [7] and academic performance analysis [8] are some examples of such applications.

Research studies for the eye detection and eye tracking mostly focus on the iris and pupil localization. Once the coordinates of pupils are determined, it can be used for the eye tracking, gaze estimation and eye movements within the images and video frames [6]. Eye images can be characterized by the intensity distribution of iris, pupil and the cornea, in addition to their shapes. It should be noted that various aspects can influence the appearance of the eye including the viewing angle, ethnicity, head position, eye colour, light conditions as well as the texture, eye state (e.g. half closed, fully closed) and current wellbeing [6].

Overall, eye detection techniques can be classified as shape-based, feature-based, appearance-based and hybrid methods. In the shape-based methods, open eyes are described by their shapes including the pupil and iris contours as well as shape of the eyelids [9-11]. For the feature-based methods, objective is to identify the local features within the eye that are less sensitive to the varying illumination as well as viewpoint [12-15]. Appearance-based methods depend upon detecting and tracking of the eyes using the photometric look which is characterized by colour distribution and filter responses to eyes and their surroundings [16-18]. The hybrid methods aiming to combine various techniques to mitigate the particular disadvantages of these methods [19-20].

Standard methods in gaze estimation are based on corneal reflections that needs an accurate localization of the pupil centre as well as the glints [21]. Pupil and glints localization algorithms are usually based on image processing such as morphological operators for the detection of contour [22] and intensity threshold identification followed by the fitting using ray-based ellipse [23]. Topography based hybrid method is introduced in [24] which uses series of filters for the iris centre estimation. However, these techniques assume that the pupil exists in the darkest area of the input image and may susceptible to varying illumination conditions that might require manual tweaking to the threshold parameters [25].

Table 1. Eye movements and classification algorithms

Reference	Model	Aims and Feature Used
[26]	Hidden Markov model	Use of fixation count, fixation durations to distinguish between expert and novice participants
[27]	Multi-layer perceptron (MLP)	Use pupil size & point-of-gaze for predicting the users' behaviours (e.g., word searching, question answering, looking for the most interesting title in a list)
[28]	Naïve Bayes classifier	Use of fixation duration, mean and standard deviation to identify various visual activities (e.g., reading, scene search)
[29]	MLP	Use of Pupil dilation, gaze dispersion to classify various tasks on decision making
[30]	Decision tree, MLP , support vector machines (SVM), linear regression	Use of fixation rate, fixation duration, fixations per trial, saccade amplitude, relative saccade angles to identify eye movements to predict visualization tasks

There are four main eye movement behaviours which are likely to show different details related to cognitive efforts when responding to tasks including blinks, pupillary responses, fixations, and saccades [31]. Blinking represents the involuntary deed of opening and closing the eyelids. Pupillary responses are the changes in pupil size restrained by the involuntary nervous system. Fixation represents the collection of gaze points that are relatively stable and near in spatial and temporal vicinity. Saccade represents the rapid and small eye movements when moving from one object to another [31]. These four eye-movement behaviours reveal the details about cognitive efforts and therefore can be used as suitable inputs for designing the machine learning (ML) systems as illustrated in Table 1 which shows various supervised ML algorithms to predict categorical responses from the eye movements.

In addition to conventional methods, existing works also utilize the deep learning (DL) approaches for the pupil detection while using hierarchical image patterns to enhance and eliminate artefacts with Convolutional Neural Networks (CNNs). For instance, [21] proposed the use of fully connected CNNs for segmentation of the entire pupil area in which they trained the network on 3946 video oscillography images. These images were hand annotated and generated within a laboratory environment. The authors claim that the proposed network enables them to perform elliptical contour detection, pupil centre estimation and blink detection. More explicitly, pupil centre are predicted with a median accuracy of one pixel and gaze estimation accuracy within 0.5 degrees. However, varying image resolution might provide different accuracy measures. More specifically,

[32] indicated the eye tracking as an important tool that can have a range of applications from scientific research to commercial sector. The authors show that the use of tracking software based on commodity hardware including tablets and smartphones, allows these advanced technologies to be available for everyone. The system is called iTracker which uses a CNNs model indicating 2.53cm and 1.71cm prediction error without calibration on tablets and smartphones respectively which is reduced to 2.12cm and 1.34cm using calibration.

Research presented in [23] proposed a pipeline of two CNNs cascaded for pupil detection. Authors claim that their method outperforms state-of-the-art techniques with detection rate up to 25% while avoiding computational complexity. To benchmark their proposed technique, 79000 hand labelled images were used in which 41000 were complementary to existing images from the literature. A similar work is presented in Naqvi et al. [33] which indicate that automobile accident deaths could be minimized using drivers' gaze region to provide their point of attentions. In this respect, the authors suggest the use of DL for gaze detection with the use of near-infrared camera sensors. They incorporate driver head and eye movement into their study. Gaze estimation accuracy was benchmarked using loosely correct estimation rate and strictly correct estimation rate in which the study claim achieving good accuracy when benchmarked with the previous gaze classification techniques.

Recent work that uses the CNNs based deep learning model for the pupil estimation [34] indicate around 70% accurate estimations while error threshold is within the 5 pixels. However, this accuracy is limited to be used in real time specifically, the applications that consider micro-movements within the eyes such as ADDS [7]. Similar work that uses CNNs for the pupil detection [35] indicates varying detection rate (70-90%) with respect to the tolerance level as pixel error and dataset they employed for testing. The study outcomes clearly indicate the trade-off between the error tolerance level and accuracy measure. Furthermore, the performance metric used in these studies is not standard (i.e. the error as number of pixels) and might produces varying accuracy with respect to image size and resolution. In contrast to CNNs, [36] utilizes the wavelet transform to extract the distinguishing features while SVM is used for the pupil classification. This work indicates 88.79% of accurate pupil estimation on a benchmarked dataset while utilizing the standard validation metric.

Despite the variety of existing methods for the pupil localisation, further improvements are required in terms of a precise estimation for the pupil location. For instance, DL-based pupil localisation and gaze estimation in [21] uses pixel distance to validate the performance which is not a standard representation of the error in case of varying resolutions. Furthermore, the validation is performed on a dataset containing artificially rendered images which in most cases, does not reflect the real time dynamics. Likewise, [37] presented gaze estimation that utilizes the DL-based facial landmarks detection following the image segmentation to identify the pupil within the input images. However, the 81% accuracy produced by the algorithm on a benchmark dataset indicates the lack of preciseness in pupil localisation that might lead to the incorrect gaze estimation. Furthermore, this study along with [23, 24] utilizes a static threshold while considering the pupil as the darkest area within the image that may susceptible to various illumination conditions [25] and low-resolution images. Likewise, the use of static size kernel for the template matching to find out the best-matched candidate (i.e. pupil in this case) within the image might causes local maxima. For instance, a smaller sized kernel may cause attention to noisy details (i.e. local maxima) whereas, larger size may lead to mismatches and incorrect estimation of pupil location [38].

In proposed work, we introduce an efficient algorithm for the pupil identification within low-resolution images (and video frames) using a composite of DL and image processing algorithms. To clarify the novelty of this paper, the contributions are outlined as follows. a) utilizing the pre-trained DL model to identify the facial landmarks and extraction of desired eye-frames within the input images; b) unidirectional cascades of two-dimensional (2D) convolution is used to determine the pupil coordinates within the eye-frames of varying characteristics; c) an adaptive size of kernel is used to deal with the varying size of input images (i.e. eye-frame) during the template matching; d) we used a dynamic threshold to identify the best matched candidate more reliably; e) for the first time, we introduce a relative error metric to measure the standardized distance (i.e. error) between the estimated and actual pupil centres; f) we validated the proposed methodology over multiple publicly available and benchmark datasets containing high diversity in gaze positions, participants background, lighting illuminations, image background, and comparatively smaller size of eye-frames.

The remainder of this paper is organized as follows. Section 2 entails the proposed methodology and algorithms. Section 3 presents the detailed experimental design and newly introduced evaluation metric. Statistical results and technical discussions are presented in Section 4 followed by a conclusion and future works in Section 5.

2. Proposed Method

The proposed pupil detection utilizes composite of techniques along with new algorithms while leveraging the DL-based facial landmark detection [39] to extract the eye information within an image/video frame. Existence of background noise and dark patches within the image frame and specifically prominent eyebrow parts, are normally detected as pits that might cause mismatch for computer vision-based iris and pupil detection [24, 38, 40]. However, this issue can be resolved readily by utilizing modern DL algorithms for a reliable face and eye-frame extraction from an ordinary quality images or video frames. In the first step, we utilize the facial landmark detection to extract the desired segments containing only the eye-frames (both left and right) from input image. We then convolve the extracted eye-frame with a pre-defined kernel in horizontal and vertical directions to identify the iris and pupil respectively within the eye-frame. We adapt the kernel size dynamically with respect to the varying eye-frame size to resolve the possible occurrences of local maxima being false representation of best matched patches. We further define a dynamic threshold for the identification of best-matched patch within the current eye-frame to reduce the impact of noisy matches. Figure 1 shows the sequential processing in our work to identify the pupil coordinates within an input image/video frame. The major components are: a) DL-based eye-frame extraction, b) image processing based iris localisation, and c) pupil detection, which are detailed in the following sub-sections.

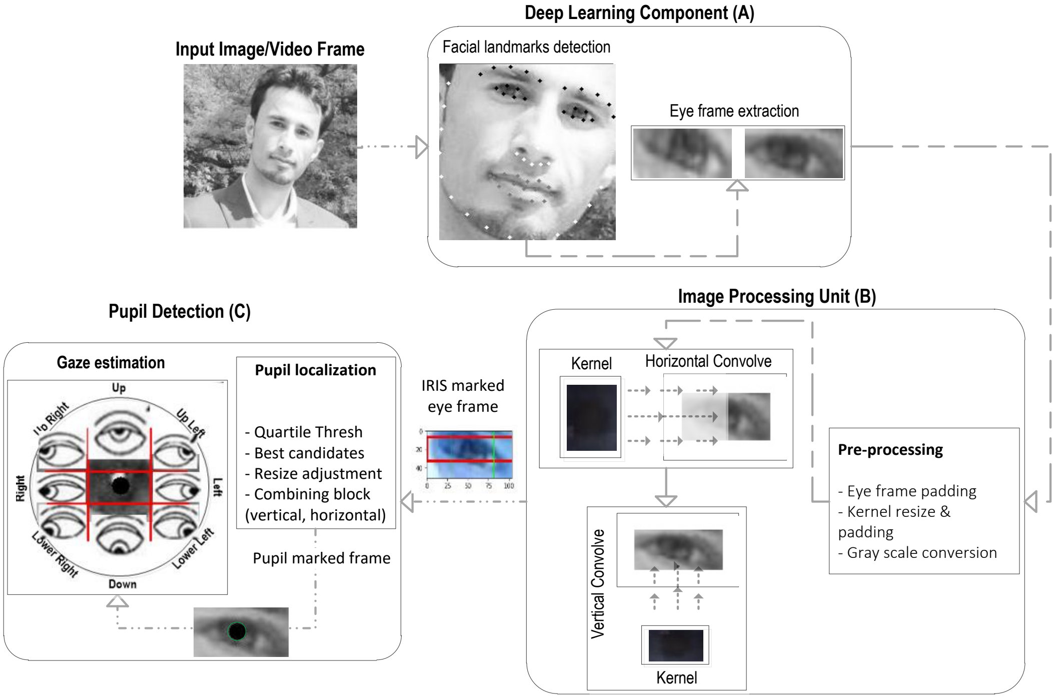


Figure 1. Sequential processing components of the proposed method comprising A) DL library (i.e. Dlib-ml) for the eye-frame extraction, B) computer vision algorithm for localizing the potential iris and pupil candidates within eye-frames, C) post-processing for the pupil coordinate measurement. In images, eyes view is reversed (e.g. the left eye in an image is the right eye in actual and vice versa)

2.1. *Eye frame extraction*

The DL component utilises a well-known toolkit (Dlib-ml) [39] which can reliably identify the facial landmarks while producing extensive fiducial points (68 in total) on the face including eye corners and eye lids as shown in figure 1(A). We first extract the face rectangle from an image using Dlib-ml that not only removes the unnecessary portion of input frame but also helps to eliminate the major noisy components that might exist in background region of the image frame. Within the face region, we then note the identified extreme points (left, right, top, bottom) for eye corners and eye lids which are used to crop the exact eye-frames within the identified face rectangle. This is one of the major advantages of using Dlib-ml which reliably eliminates the unnecessary portion of an image and extract the exact region of interest (i.e. eye-frames in this case) from the input frame. Only the input images (or video frames) with exactly one face rectangle and two eye-frames are considered as 'valid'. The output of this component in form of eye-frames (left, right) are processed further to identify the iris and pupil within the image.

2.2. *Iris segmentation and pupil localization*

Following the eye-frame extraction, a convolution function is applied for the template matching between a custom kernel and eye-frame to localise the best matching segment within the eye-frame. Firstly, we built a custom kernel representing 100 iris frames (cropped from eyes frames) randomly chosen from datasets described in Section 3. The advantage of custom kernel over an ordinary black colour kernel, is a more generalized representation of an iris for a diverse population and morphology characteristics (e.g. geometry, patterns within the iris, colour etc.). Another common factor that can affect the template matching performance, is size of the template (i.e. kernel). Smaller sized kernel may cause attention to noisy details (i.e. local maxima) whereas, larger size may lead mismatches and incorrect estimation [38].

To resolve this issue, the adaptive size kernel is employed using the interpolation and extrapolation techniques where the size ($w_k \times h_k$) varies with respect to the input frame size (i.e. eye-frame). Furthermore, eye-frame (*E*) is padded with a rim of white pixels (see figure 1 and figure 2) to enlarge it enough that the convolution kernel (*K*) fits inside the padded image to provide all possible best matches identification (i.e. between kernel *K* and overlapped eye-frame patches of a size similar to *K*). More specifically, when the desired patch (i.e. iris) is located at extreme positions (e.g. looking extreme left/right positions).

$$y[i,j] = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} K[m,n].\,E[i-m,j-n] \qquad (1)$$

Equation 1 represents a 2D convolution function where *E* is the current eye-frame (within the input image) to be convolved with the kernel matrix *K* resulting *y* as the output image. The indices *i*, *j* and *m*, *n* represent the indices within the *E* and *K* matrices (i.e. image pixels), respectively.

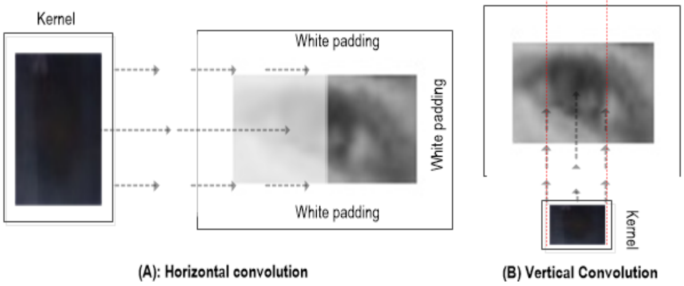


Figure 2. Horizontal convolution (A) and vertical convolution (B) between adaptive size kernel *K* and white outlined eye-frame *E*.

In contrast to the ordinary way of 2D-convolution where kernel *K* slides along *E* with a fixed overlapping window (usually 1 pixel) in both horizontal and vertical directions, we perform a comparatively simple and efficient convolutional steps (only one slide per horizontal and vertical directions) as shown in figure 2. The reason behind an adaptive kernel selection is the geometric features of iris and pupil which are considered approximately circular and black compared to the rest of the eye with pupil as the most dark segment. First, kernel height h_k is resized to eye-frame height (i.e. $h_e = h_k$) and width w_k is set to 0.4 of the eye-frame width. The convolution function then slides through *E* in the horizontal direction to determine the x-coordinate of iris centre within the *E*. It compares the overlapped patches of *E* ($w_k \times h_k$) against *K* to calculate the matching scores at each horizontal stride (i.e. 1 pixel). The normalised correlation coefficient calculates a total matching score for the current patch in *E* using equation 2.

$$S(x,y) = \frac{\sum_{x'y'}(K'(x',y')\ .\ E'(x+x',y+y'))}{\sqrt{\sum_{x'y'}K'(x',y')^2\ .\ \sum_{x'y'}E'(x+x',y+y')^2}} \quad (2)$$

Where $S(x,y)$ is the matching score of current overlap (*x*, *y*) between *K* and *E* patch of size equal to *K* ($w_k \times h_k$). The summation in equation 2 is performed over the *K* and *E* patch where $x'= 0 \ldots w_{k-1}$, $y'= 0 \ldots h_{k-1}$. As the kernel height h_k is aligned with height of the eye-frame (i.e. $h_e = h_k$), there are no vertical overlapping (i.e. no vertical overlapping/strides) which means, the kernel will only be able to move along *E* in the horizontal direction while computing the matching scores for overlapped patches in *E*.

Once all the horizontal matching scores are calculated, the next step is to find the coordinates of the best matching segment. There have been several approaches to select the optimal match but the candidate with maximum match have been commonly used in similar works [12, 38, 41]. However, it can easily cause local maxima specifically in low-resolution images [38]. Likewise, using a predefined matching threshold can provide varying matching scores regarding the environment and can also mislead because of varying dynamics such as illuminations. We utilised quantile measure to select all candidates (*M*) in the horizontal direction) that crosses the adaptive threshold of 90th percentile of the matching scores sorted in ascending order. i.e. $M \in SC_h$ such that $\forall\ SC_h > 90^{th}$ percentile of sorted SC_h.

The mean of horizontal (*x-axis*) coordinate of *M* selected patches is calculated using (3) which represents the x-coordinate of top-left corner ($R_{x,y}$) of the final best matched patch (i.e. estimated iris rectangle).

$$R_x = \frac{\sum_{i=1}^{m} M_x(i)}{m} \quad (3)$$

Where *m*, are the total number of elements (i.e. best-matched candidates) in *M*, M_x is the horizontal coordinate of corresponding best-matched candidates *M*.

The iris rectangle *I* is identified using R_x and kernel width w_k which is then used for the vertical convolution to identify the y-coordinate of iris centre. Similar to horizontal convolution-based matching, kernel height w_h is resized to 0.4 of the height of *I* for overlapped stride matchings while keeping the width same. Vertical convolution steps are then performed to compute the matching score for *K* and overlapped patches of *I* along the vertical direction only. The output matrices SC_v contains all the corresponding matching scores for vertical convolutions between the *K* and *I* overlapped patches. The quantile measure is used in a similar way to select all candidates (*N*) in the vertical direction) that crosses the adaptive threshold of 90th percentile of the matching scores sorted in ascending order where; $N \in SC_v$ such that $\forall\ SC_v > 90^{th}$ percentile of sorted SC_v. The mean of vertical (*y-axis*) coordinate of *N* selected patches is then calculated using equation 4 which represents the y-coordinate of top-left corner ($R_{x,y}$) of the final best-matched patch (i.e. estimated pupil rectangle).

$$R_y = \frac{\sum_{i=1}^{n} N_y(i)}{n} \quad (4)$$

where *n,* is the total number of elements (i.e. best-matched candidates) in *N*, N_y is the vertical coordinate of corresponding best-matched candidates *N*.

Finally, the centre coordinates of the best-matched patches within *E* in horizontal (C_x) and vertical directions (C_y) represent the pupil location along the *x-axis* and *y-axis* respectively and are calculated as:

$$C_x = R_x + w_k/2, \; C_y = R_y + h_k/2 \qquad (5)$$

where w_k, h_k are the width and height of kernel *K*, respectively. Algorithm 1 summarizes all the sequential steps involved in the proposed methodology to determine the pupil coordinate within an image frame.

Algorithm 1: Proposed algorithm for iris detection and pupil localization in an image/video frame

```
Inputs: image/video frame F, a custom-defined kernel frame K
Output: Pupil coordinates (Cx, Cy), iris rectangle (top-left; bottom-right)
STEP1:
-   Initialise validation Score = 0 for current F
-   Use Dlib-ml for the facial landmark detection within input frame F
-   Crop the face rectangle (Face) using the detected landmarks
-   IF count (Face) ==1 (i.e. exactly one face in image is found)
        - Score ++
        - Extract the eye-Frames ($E_L$, $E_R$) for left and right eye
        - IF count ($E_L$, $E_R$) ==2. i.e. exactly 2 eyes within Face rectangle
                ▪ Score ++
                ▪ Goto STEP 2
        - ELSE
                ▪ Mark it as invalid frame
                ▪ Goto STEP 1 for the next F
-   ELSE
        - Mark it as invalid frame
        - Goto STEP 1 for the next F
STEP2:
    -   Foreach eye-frame E in $E_L$, $E_R$
            ▪ Convert E into grayscale
            ▪ Outline E with white paddings
            ▪ Adapt the kernel K height to height of E and width to 0.4*width(E)
            ▪ Convolve K with E by sliding Horizontally with 1-pixel stride/sliding window
            ▪ Store the matching scores for overlapped E patches in a vector $SC_h$
            ▪ Store the horizontal elements with high matching scores in lists M for
              $M \in SC_h$ such that $\forall SC_h > 90^{th}$ percentile of sorted $SC_h$.
            ▪ Find the top-left of best-identified iris rectangle by taking mean ($\mu$) of x-coordinates
              for M (i.e. $R_x$) using equation 3
            ▪ Find the iris rectangle I, using $R_x$ and $w_k$
            ▪ Goto STEP3
    -   End Loop
STEP3:
    -   Adapt the kernel K height to 0.4*height(I) for vertical convolution
    -   Convolve K with I by sliding Vertically with 1-pixel stride/window
    -   Store the matching scores for overlapped I patches, in a vector $SC_v$
    -   Find the elements with high matching scores (call them N) where
        $N \in SC_v$ such that $\forall SC_v > 90^{th}$ percentile of sorted $SC_v$.
    -   Find the top-left coordinate of best-identified rectangle by taking mean ($\mu$) of y-coordinates of
        N (i.e. $R_y$) using equation 4
    -   Find the pupil centre $C_x$, $C_y$ by adding width and height of K into $R_x$ and $R_y$ respectively
        using equation 5.
```

3. Experimental Design

We conducted detailed experiments to validate the proposed methodology while using various datasets and validation metrics. We also performed a critical analysis based on various conditions and validated the proposed algorithm while considering the diversity in validation datasets as well as validation metrics. Following sections explain the validation datasets and metrics along-with detailed experimental design.

3.1. Datasets

To validate the proposed methodology and reliable performance measure, we used three different publicly available datasets. The first dataset is known as Talking-Face [42] and have been used in previous works [37]. This dataset contains 5000 video frames captured during the engaged conversation from a person for 200 seconds. The original objective of this dataset was to model the facial behaviour during a natural conversation. Data is captured with a static positioned camera with a frame size of 720x576 pixel. Every frame is annotated semi-automated manner containing 68 facial points including the pupil coordinates. Following our validation check in Algorithm I (i.e. frames with exactly 2 eyes/frame) and removing the fully closed eyes (manually, found 280 images) images, we are left with 4720 frames for the validation purpose. The dataset contains varying gaze positions, facial and body movements, diverse natural expressions and variations in eye-state (e.g. closed, open, half closed) . However, because it is captured from individual person , the diversity within the eye characteristics is very limited. In other words, there are no variations in terms of eye characteristics (e.g. Iris or pupil colour, intensity, iris pattern etc.) and hence, not very challenging for the algorithm validation.

In contrast to Talking-Face, we used the BIO-ID dataset [43] which is comparatively more challenging and has been used as a benchmark in various relevant studies such as [37, 41]. The data was acquired from 23 different subjects during multiple sessions and has 1521 images in total containing varying gaze positions, illuminations, background scene, eye features (e.g. eye colour, gender, ethnicity, iris size), camera focus and hence eye-frame (and face rectangle) size. The interesting aspect of this dataset is a comparatively lower resolution (grayscale 384x288 pixel) that makes the validation of pupil localisation algorithm more challenging but reliable. Besides, the dataset contains natural expressions such as images with half-closed eyes that further help to measure the validity of the proposed algorithm. Our algorithm detects only seven frames as invalid (i.e. not containing exactly two eyes) whereas we found 45 images (manually) with fully closed eyes that were excluded, resulting 1469 remaining dataset for validation purpose.

Furthermore, we evaluated our method on comparatively larger dataset known as GI4E [44] containing more diversity involving various morphology types (e.g. eye size, eye/iris features, gender, ethnicity, varying background and illuminations). It should be noted that despite higher resolution images (800×600 pixels), size of the eye-frame rectangles is comparatively small. This is because of the larger distance of the capturing device from the subject resulting lower ratio of eye-frame to entire image. In other words, the whole frame covers more background pixels as compared to the actual face within the image which makes the eye-frame and hence iris/pupil localization more challenging. The dataset is much diverse containing 103 subjects (each with 12 images) with 1236 total images involving 12 different gaze position. Also, there is no open eyes or invalid frame in this dataset.

3.2. Validation Metrics

One of the important factors in validation of the pupil detection and proposed work is the metric we chose for the performance measure. This is because of the nature of pupil localization problem. For instance, the absolute error in the estimated pupil/eye centre and actual eye centre might vary with respect to image size/resolution. Hence the standard distance measure such as Euclidean distance (ED) and/or R^2 coefficient will not give a true representation of the accuracy measure. The authors in [43] introduced a relative error measure (d_{eye}) to deal with this issue which has been utilized in various related works [37, 38, 43, 45]. It uses the maximum of the estimated pupil coordinates distances from left and right eyes (d_l) and (d_r) respectively, between the actual eye centres (C_l, C_r) and the estimated ones ($\tilde{C}_l$, $\tilde{C}_r$) using equation 6.

$$d_{eye} = wec = \frac{max\ (\|\tilde{C}_l - C_l\|,\ \|\tilde{C}_r - C_r\|)}{\|C_l - C_r\|} \quad (6)$$

For the normalisation, the calculated distance is divided by the distance between two actual eye centres $\|C_l - C_r\|$ as shown in equation 6. The normalisation factor makes the error measure independent of the image scale and hence eye-frame size. Furthermore, [37] used best eye centre (bec) which utilizes the minimum of the error between estimated and actual centres as:

$$d_{eye} = bec = \frac{min\ (\|\tilde{C}_l - C_l\|,\ \|\tilde{C}_r - C_r\|)}{\|C_l - C_r\|} \quad (7)$$

Although the *wec* (i.e. worst eye centre) metric provides a relative error estimate, it is based on some assumptions such as '*on average population, the distance between the inner eye corners is equal to width of a single eye of the corresponding subject*'. Likewise, a relative error of d_{eye}= 0.25 is considered as half of an eye width which may not be valid in every case. Interested readers can get further details in [43] study.

To further deal with the metric generalisation issue, we first time introduce a standardized error measure (*SED*) as a function of distance between the estimated and actual coordinates within an eye-frame. It calculates the relative distance as percentage of the total possible ED (i.e. error) between the actual and estimated pupil coordinates. The *SED* measure interprets the error within the single eye-frame without depending on the second eye or interpupillary distance used in other related works. Besides, the *SED* metric can measure the relative error regardless of image/face or eye-frame size and hence the image resolution. Mathematically, the proposed *SED* is defined as:

$$S_{ED} = \frac{\sqrt{(Cx_e - Cx_a)^2 + (Cy_e - Cy_a)^2}}{\sqrt{(x_{min} - x_{max})^2 + (y_{min} - y_{max})^2}} \times 100 \quad (8)$$

Where Cx_e, Cx_a represent the estimated and actual pupil horizontal coordinates respectively and Cy_e, Cy_a represent the estimated and actual pupil vertical coordinates respectively. The x_{min}, y_{min} are coordinates of the nearest corner of eye-frame (usually top left corner) whereas, x_{max}, y_{max} are coordinates of the farthest corner of eye-frame (usually bottom right). The numerator in equation 8 represents the error (in terms of pixels) between the actual and estimated positions whereas the denominator is the total possible error and is used as a normalisation factor. The resulting *SED* gives the percentage error representing a standardised distance between actual and estimated pupil positions in pixels which is not affected by the image size and resolution. In addition to evaluate the pupil detection techniques, the proposed standardised distance measure can also be useful for other related works such as object localisation, image segmentation and object tracking etc.

In summary, a comprehensive comparative analysis is performed to evaluate the proposed methodology using aforementioned metrics including *wec*, *bec*, and *SED* along with other standard accuracy measures including the ED, absolute mean difference, and R^2 (coefficient of determination).

4. Results and Discussions

Following the experimental design, performance of the proposed pupil detection approach is evaluated using various gold standards, validation metrics and benchmarked datasets. As discussed in the experimental design, it is important to use appropriate evaluation methods due to nature of the problem. To maintain the reliability in our performance measure, we utilised different metrics as well as the newly introduced SED in this work.

Table 2. Performance analysis of the proposed model using *wec*, *bec* with varying error threshold

	wec(%)		*bec*(%)	
Dataset	**Error ≤ 0.05**	**Error ≤ 0.1**	**Error ≤ 0.05**	**Error ≤ 0.1**
BIO-ID	94.5	100	98.34	100
Talking face	97.10	100	99.7	100
GI4E	95.05	100	98.71	100

Table 2 summarizes the results achieved from the proposed approach using *wec* and *bec* metrics that have been used in recent similar works [36-38, 43-45]. We are specifically interested in *wec* measure when *error≤0.05* which indicates the model estimation within the *pupil* diameter (i.e. more restricted). Best accuracy achieved by the proposed method is 97.1% while tested over the *Talking-Face* dataset which outperforms the 89.59% presented in recent work[37] that uses the same dataset. The high accuracy is expected because of the comparatively less challenging nature of dataset (see Section 3.1). Firstly, the dataset contains high resolution images. Secondly, the data is captured from only one person hence, a generalization of *iris* and eye pattern is easily detected. It is important to note that despite the dataset is collected from single person, it contains high variations in terms of gaze, head movements, facial expressions and sufficient quantity (i.e. 5000 images) with annotated pupil coordinates. On the other hand, the proposed method achieves 100% *wec* accuracy while tested for error threshold≤0.1 indicating the robustness of the proposed methodology. This means that model estimation about *pupil* coordinates are within the *iris* in all cases (i.e. 5000 images). Overall, proposed method outperforms the most recent works related to pupil localization [37] while evaluated on the *Talking-Face* dataset.

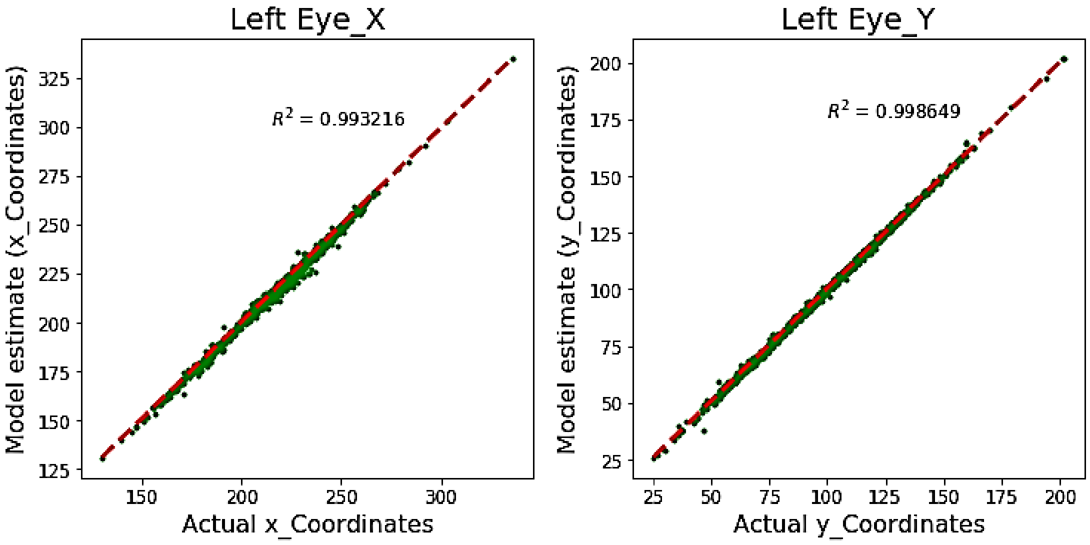


Figure 3. Comparison of estimated pupil coordinates using proposed model, with actual annotated coordinates (BIO-ID dataset) using R-squared error

To further evaluate the model performance, the BIO-ID dataset is used which contains various subjects, high variations in gaze, head pose and body movements. Furthermore, the image quality (i.e. resolution) is comparatively lower (i.e. 286x384) which makes it more challenging when focusing the identified eye-frame and/or iris/pupil within the image. Also, a large proportion of the entire image contains background rather than the face itself which makes the dataset further challenging as addressed by the previous works [38]. Despite the associated challenges, proposed approach shows robust pupil estimations as shown in Table 2. The model indicated significant improvements with 94.5% *wec* measure with error threshold≤0.05 when benchmarked with the works of [37] and [40] of 81% and 84%, respectively. Furthermore, the model indicated 100% accuracy when evaluated for error threshold≤0.1 which means that pupil localization is within the iris in all cases (i.e. 1521 cases in total). Despite the 100% of *wec and bec* accuracy for error threshold≤0.1, the main focus is to maximize the *wec* accuracy (which is the most challenging) with minimum error threshold (i.e. ≤0.05) to restrict the model estimation within the pupil diameter.

Figure 3 shows the R^2 coefficient for the proposed model tested on BIO-ID dataset. It can be observed that x-axis and y-axis estimated coordinates are almost overlapping to actual annotations with R^2 value of 0.993 and 0.998 for x-axis and y-axis respectively. Although, R^2 is a well-known statistical measure to determine the goodness of model fit, it might not be effective for validating the model estimation in pupil detection or similar problems because of the varying error rate with respect to the image size (and resolution).

Table 3. Performance comparison between previous works based on *wec* measure using BIO-ID dataset

wec % accuracy with varying error (e) threshold

Methods	***e<0.05***	***e<0.1***	***e<0.15***	***e<0.2***
[24]	81.1	94.2	96.5	98.5
[36]	88.7	95.2	96.9	97.8
[37]	80.9	91.4	93.5	96.1
[38]	82.5	93.4	95.2	96.4
[40]	84.1	90.9	93.8	97.0
[41]	57.2	96.0	98.1	98.2
[43]	38.0	78.8	84.7	87.2
[45]	47.0	86.0	89.0	93.0
[46]	85.8	94.3	96.6	98.1
Proposed Model	**94.5**	**100**	**100**	**100**

Table 3 summarizes the comparative results from various previous works while weighted over the challenging BIO-ID dataset using *wec* metric with varying thresholds. It can be noticed that the proposed model outperforms (94.5%) all previous works specifically with the most restricted error threshold≤0.05. Recent works that uses similar approach [37] achieved an accuracy of 80.9% and 82.5% [38] with e≤0.05 whereas best accuracy of 88.79% is indicated by [36] that are significantly lower than the proposed method. Research study in [21] presented a robust technique for the pupil localization and gaze estimation, however, the measured performance is not standard (i.e. uses the mode of pixel distance which is not the true representation of error with varying resolutions). Furthermore, the validation is performed on different dataset containing artificially rendered images which in most cases, does not reflect the real time dynamics.

Besides the Talking-Face and BIO-ID datasets, we evaluated the performance of proposed approach on another challenging dataset GI4E. It can be noted from Table 2 that our model produces 95.05% *wec* and 98.71% *bec* accuracy respectively with critical threshold≤0.05. While most of the existing works used BIO-ID as benchmark dataset, some of them also used GI4E to evaluate their techniques. For instance, recently study on eye centre localisation [24] reported 93.9% *wec* accuracy on GI4E dataset which is slightly lower than our approach (i.e. 95.05%) however, their accuracy was decreased to 881.2% when tested on BIO-ID dataset. This indicates the robustness of proposed approach for pupil detection in varying datasets containing diversity in terms of eye colour, gaze position, facial emotions and real movements. Similarly, [46] indicated 89.28% *wec* on GI4E dataset which are significantly lower than the proposed approach. A clustering-based approach [47] produced mean pixel error of 2.73 pixels as compared to proposed model with 1.7 pixels while validated on GI4E. However, it is important to be noted that this metric does not represent a standard accuracy measure as described in Section 3.2.

In addition to *wec,* [24, 41] used average point-to-point error (m_{e17}) with the inter-ocular distance between the left and right eye pupil. Recent works [21] that uses the DL to localize the pupil and estimate the gaze position also employed the median of absolute difference in x-axis and y-axis.

However, variations in image size, zoom-in/out due to body/head movements and/or camera positions might affect the mean difference in corresponding error estimate resulting variations in accuracy measure. The *wec* metric which has been used extensively in related works such as [24, 36-38, 40, 45, 46], gives a comparatively better indication of the performance measure. However, these metrics measures the performance in terms of coordinate estimation within the pupil/iris diameter with a varying error threshold as shown in Table 2. Also, it is based on relative error assumption (d_{eye}= 0.25) as half an eye width, which may not be true in every case. Therefore, model estimations and performance (specifically in pupil localization task) is needed to be evaluated using more standard metric representing the distance between estimated and actual pupil coordinates.

Table 4. Comparing model estimations using newly introduced SED, Euclidean distance (ED), R^2, and absolute error metrics

Dataset	**$\mu\|x_a$-$x_e\|$**	**$\mu\|y_a$-$y_e\|$**	**R^2_x**	**R^2_y**	**ED(c_a, c_e)**	**%ED(c_a, c_e)**
BIO-ID	1.04	0.57	0.993	0.998	1.43	3.98
Talking face	1.23	0.97	0.990	0.956	1.96	2.49
GI4E	1.32	0.71	0.996	0.999	1.70	3.87

To overcome this issue, we first time introduce a standardized Euclidean distance (SED) which represents the percentage distance error as ED using equation 7 (see Section 3.2). The error represents the displacement between the actual and estimated pupil coordinates as a percentage of the whole image size (i.e. eye-frame) in terms of number of pixels. The major advantage of SED is a standard representation of the error which can be used to measure the accuracy regardless of image size and resolution which is not the case in *wec*, m_{e17} and other metrics used in most of the existing studies. Table 4 presents the comparative analysis of proposed model estimations in terms of mean pixel difference in each axis, for both eyes (left and right), R^2 coefficient, ED between centre of estimated and actual pupil coordinates and the newly introduced SED. The proposed method indicates 1.04 and 0.57 absolute pixel error on x-axis and y-axis respectively (i.e. 0.8 on average for both) as compared to 2.91 in [47] on BIO-ID dataset. Similarly, a DL-based model in [35] indicated their optimal performance with pixel error>10. However, they used different datasets which in case of high resolution, is not comparable with proposed method and clearly indicates the need of standard metric similar to SED.

It can be analyzed that the model performs comparatively better for Talking-Face and BIO-ID datasets as compared to GI4E dataset based on the corresponding properties (as discussed in Section 3). However, there are several crucial aspects to be noted in each case. First, in contrast to *wec* measures in Table 2, the ED(c_a, c_e) error in Table 4 for Talking-Face is 1.96 which is higher than the other two datasets (1.43 and 1.70 for BIO-ID and GI4E respectively) despite the high quality and fewer variations in the former case. This is because the size of images in Talking-Face dataset is comparatively larger than other datasets and consequently, the ED(c_a, c_e) error as well as absolute error ($\mu|x_a$-$x_e|$, $\mu|y_a$-$y_e|$) in each axis, are also high. However, results from these metrics (i.e. ED, $\mu|x_a$-$x_e|$, $\mu|y_a$-$y_e|$) does not align with results in Table 2 (*wec measure*) and therefore, does not reflect the true measure of the standardized difference between estimated and actual pupil coordinates. In contrast, SED provides more generic and standard representation of error between the actual and estimated coordinates as a percentage of the eye rectangle size. The SED error for Talking-Face dataset is 2.49% which is less than 3.98% and 3.87% of BIO-ID and GI4E datasets respectively, and also aligns with the *wec* outcomes in Table 2. As mentioned earlier, SED represents a standardized distance (i.e. pixels) using current eye-frame without depending upon the second eye or interpupillary distance which is not the case in *wec* measurement. Furthermore, SED interprets the error in term of pixel distance without using any thresholds (as in case of *wec*) and can be utilized as a standard metric to evaluate the true performance of such models in similar problems.

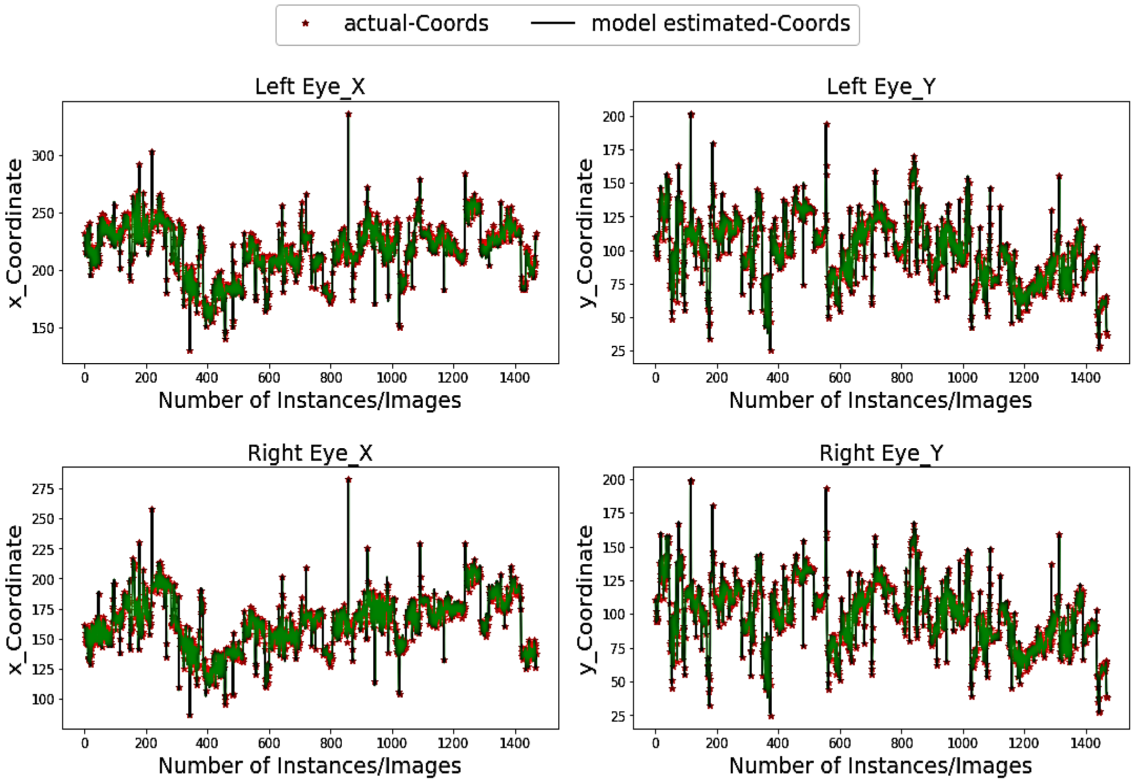


Figure 4. Pupil coordinates estimations (green color) vs actual (red) coordinates within BIO-ID dataset

Figure 4 demonstrates the pupil estimation performance of the proposed model for both left and right eye (*x-axis* and *y-axis*) on BIO-ID dataset. The model indicates a perfect overlapping for both axis and more specifically, at the peak positions which represent the extreme pupil and/or iris positions looking extreme left or right, and top or bottom positions. One of the reasons of such robust overlapping is the use of white paddings in our model that helps the adaptive kernel to achieve maximum overlaps at extreme positions resulting in appropriate matching candidates during horizontal and vertical cascades.

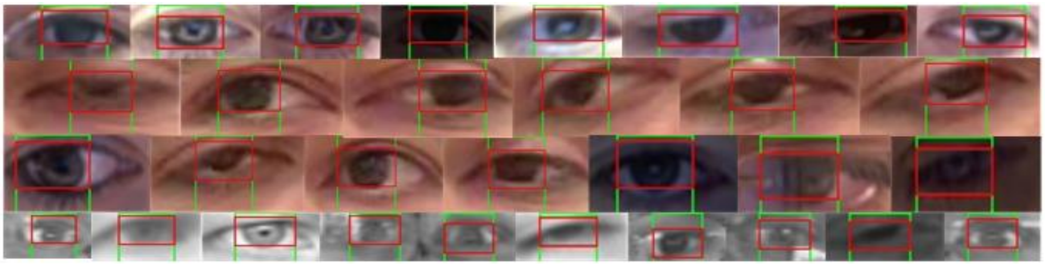


Figure 5. Horizontal and vertical convolution-based pupil coordinates localization (in randomly selected images from BIO-ID, GI4E and Talking-Face dataset) for dynamic conditions such as gaze position, eye color, intensity, noise interference, eye size and image resolution

As discussed earlier, a custom kernel might help for optimal representation of iris diversity. Additionally, adaptation of kernel size regarding the eye-frame and dynamic threshold for best candidate selection further improves the reliability of our method specifically in dynamic conditions. Figure 5 demonstrates various test cases of iris/pupil detection using proposed methodology for diverse eye properties and varying environmental conditions (e.g. patterns, gaze direction, varying background, half/full closed eyes, colour, intensity, illuminations, resolution, pupil/iris size, gender, ethnicity etc.). It indicates the robustness of model estimations in both horizontal and vertical convolutions specifically at extreme positions (such as left/right corners, top right, half-closed etc.)

Primarily, the proposed method is leveraging the pre-trained Dlib-ml that can locate the facial landmarks efficiently and reliably. It helps to filter out the unnecessary background segments within the input image as well as irrelevant facial components excluding the desired regions that contain exact eye-frames. Secondly, the proposed method uses efficient algorithm to adapt the kernel size in accordance with the eye-frame and padding the eye-frame with white surrounding pixels which further reduce the probability of selecting noisy matched candidates as mentioned by [37, 38]. The use of quantile based dynamic threshold to identify the best matching patch further enhances the reliability in proposed algorithm (e.g. outcomes in Figure 4-5).

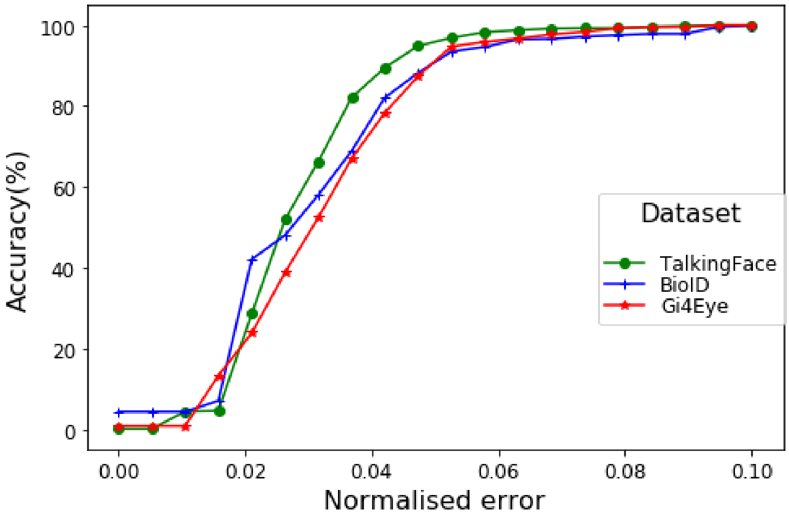
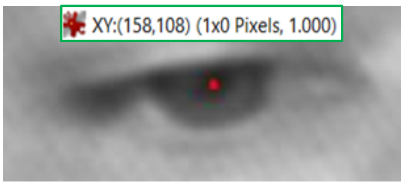


Figure 6. The wec measure for different datasets using proposed method

Figure 6 shows the performance of the proposed method for pupil coordinate estimation using BIO-ID dataset while varying error threshold, to measure the mean *wec* for both eyes. The visualization indicates accuracy over 90% in all cases (i.e. dataset) while considering the strict constraint of $e \leq 0.05$. More explicitly, the model indicates that in over 97% of cases with high-resolution images/videos (which are ordinary for current technological advancement), the error in estimated pupil position is less than the diameter of pupil itself. Even in the worst-case scenario (i.e. small-size eye-frames in GI4E dataset), the model achieves above 95% accuracy.



Ground truth file	#LX	LY	RX	RY
For BioID_0000.eye	232	110	161	110

Figure 7. Example of annotation error in BIO-ID dataset

It is also imperative to mention that some annotation errors may slightly influence the performance measure even though, this is observed in very few cases. For instance, Figure 7 indicates the eye centre coordinates annotations in BIO-ID dataset (R_x:161, R_y:110) provided by [43] for the right-eye of subject *BioID_0000.eye*. However, the correct values are R_x:158, R_y:108 (refer to Figure 7) which indicate approximately 2 pixels difference in each axis. This is significant for micro-movements estimation and would affect the model performance substantially (e.g. *wec*, SED).

Finally, it can be noted that the proposed model performs initial checks on the current frame quality to assure the existence of exactly two eyes (Algorithm 1) within the identified face rectangle. However, additional constraints can further improve the accuracy specifically, in real-time scenarios and video stream data. For instance, [37] used the DL model to identify the blinking eyes which can further improve the accuracy of proposed model while filtering out the images/frames without distinctive iris/pupil (i.e. separating the closed eyes not to be analyzed for pupil localization). Additional post-processing constraints such as symmetry constraints over the estimated pupils' coordinates in both eyes might improve the gaze estimation accuracy. This might be useful to improve the eye-state information extraction approaches such as [7] for the deception detection through facial micro-gestures.

5. Conclusions and Future Works

We proposed a novel pupil estimation method utilising the deep learning based facial landmark detection and an image processing algorithm to determine the eye centre within an image frame. Reliable extraction of the eye-frames within the input image is one of the major advantages of using Dlib-ml. This eliminates most of the background and irrelevant segment of the image which helps to identify the target segment using intelligent image processing. We developed a customized iris kernel using multiple images from various datasets, for its generalized representation. The iris kernel is then convolved with eye-frame in two stages (horizontal and vertical) such that no nested strides are performed by convolution function. White paddings surrounding the kernel as well as eye-frame, proved very helpful for template matching between the kernel and overlapped eye-patches, specifically for the extreme eye positions (e.g. left/right corners). Also, utilising a dynamic threshold for identifying the best-matched patch further contributed to reliability in our method.

Compared to several state-of-the-art pupil detection methods, the proposed approach indicated significant improvements in pupil estimation accuracy specifically, with lower-resolution images and minimum error thresholds. We also introduced a standardized distance metric to measure the relative error in model estimation. This metric can be used regardless of image size and resolution which is not the case with most of the existing validation metrics used in similar works. In future, proposed method will be utilised along with eye-blink detection models, to determine eye gaze, in particular for infraduction iris positions. Our method can be useful in various computer vision applications specifically the one requiring precise pupil and eye centre estimation. For instance, the eye related feature extraction in [7] can be replaced with our method to extract the more reliable and micro-level movements within the eyes to distinguish the truthful and deceptive behaviour. More explicitly, this work is expected to direct several application areas such as human-computer interfaces, gaze estimation, emotion recognition, psychological profiling, fatigue detection, healthcare, visual aid and automated deception detection.

Funding: This work is funded by Prince Sattam bin Abdulaziz University, KSA under grant number: 2020/01/1174.

Author Contributions: Conceptualization, Wasiq Khan; methodology, Wasiq Khan and Abir Hussain; software, Wasiq Khan, Kaya Kuru and Haya Al-Askar; validation, Wasiq Khan, Abir Hussain, and Kaya Kuru; investigation, Wasiq Khan; writing—original draft preparation, Wasiq Khan; writing—review and editing, Abir Hussain and Kaya Kuru; visualization, Haya Al-Askar; supervision, Wasiq Khan and Abir Hussain; funding acquisition, Haya Al-Askar, Wasiq Khan and Abir Hussain.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Monforte, P.H.B.; Araujo, G.M.; De Lima, A.A. Evaluation of a New Kernel-Based Classifier in Eye Pupil Detection. Proceeding of 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, 2018, pp. 380-385.
2. Al-Rahayfeh, A.; Faezipour, M. Eye tracking and head movement detection: A state-of-art survey. *IEEE J. Transl. Eng. Health Med* 2013, vol. 01, pp. 1-12, doi: 10.1109/JTEHM.2013.2289879.
3. Guan, X.F.X.; Peli, E.; Liu, H.; Luo, G. Automatic calibration method for driver's head orientation in natural driving environment. *IEEE Trans. Intell. Transp. Syst.*, 2013, vol. 14, no. 01, pp. 303–312.
4. Horak, K. Fatigue features based on eye tracking for driver inattention system.In Proceeding of 34th Int. Conf. Telecommun. Signal Process. (TSP), Aug. 2011, pp. 593–597.

5. Harischandra, J.; Perera, M.U.S. Intelligent emotion recognition system using brain signals (EEG).Proceeding of IEEE EMBS Conf. Biomed. Eng. Sci. (IECBES), Dec. 2012, pp. 454–459.
6. Hansen, D.W.; Ji, Q. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Trans. Pattern Anal. Mach. Intell* 2010, vol. 32, no. 03, pp. 478–500.
7. O'Shea, J.; Crockett, K.; Wasiq, K.; Kindynis, P.; Antoniades, A.; Boultadakis, G. Intelligent Deception Detection through Machine Based Interviewing.IEEE International Joint conference on Artificial Neural Networks (IJCNN), Rio de Janeiro, 2018, pp. 1-8, doi: 10.1109/IJCNN.2018.8489392.
8. Waheed, H.; Hassan, S.; Aljohani, N.R.; Hardman, J.; Alelyani, S.; Nawaz, R. Predicting academic performance of students with VLE big data using deep learning models. *Computer in human behavior* 2020, vol. 104, doi: https://doi.org/10.1016/j.chb.2019.106189
9. Fasel, I.R.; Fortenberry, B.; Movellan, J.R. A Generative Framework for Real Time Object Detection and Classification. *Computer Vision and Image Understanding* 2005. vol. 98, no. 01, pp. 182210.
10. Feng, G.C.; Yuen, P.C. Variance Projection Function and Its Application to Eye Detection for Human Face Recognition. *Int'l J. Computer Vision* 1998, vol. 19, pp. 899-906.
11. Feng, G.C.; Yuen, P.C. Multi-Cues Eye Detection on Gray Intensity Image. *Pattern Recognition* 2001, vol. 34, pp. 1033-1046.
12. Kawato, S.; Ohya, J.Two-Step Approach for Real-Time Eye Tracking with a New Filtering Technique. Proceeding of Int'l Conf. System, Man and Cybernetics, 2000, pp. 1366-1371
13. Kawato, S.; Ohya, J. Real-Time Detection of Nodding and Head-Shaking by Directly Detecting and Tracking the BetweenEyes, Proceeding of IEEE Fourth Int'l Conf. Automatic Face and Gesture Recognition, 2000, pp. 40-45.
14. Kawato, S.; Tetsutani, N. Detection and Tracking of Eyes for Gaze-Camera Control. Proceeding of . 15th Int'l Conf. Vision Interface, 2002.
15. Kawato, S.; Tetsutani, N.; Real-Time Detection of Between-theEyes with a Circle Frequency Filter. Proceeding of Asian Conf. Computer Vision '02, 2002, vol. 02, pp. 442-447, 2002.
16. Huang, W.M.; Mariani, R. Face Detection and Precise Eyes Location. Proceeding of Int'l Conf. Pattern Recognition, 2000.
17. Pentland, A.; Moghaddam, B.; Starner, T. View-Based and Modular Eigenspaces for Face Recognition. Proceeding of IEEE Conf. Computer Vision and Pattern Recognition, June 1994.
18. Huang, J.; Wechsler, H. Eye Detection Using Optimal Wavelet Packets and Radial Basis Functions (RBFs). *Int'l J. Pattern Recognition and Artificial Intelligence* 1999, vol. 13, no. 07.
19. Hansen, D.W.; Hansen, J.P.; Nielsen, M.; Johansen, A.S.; Stegmann, M.B. Eye Typing Using Markov and Active Appearance Models. Proceeding of IEEE Workshop Applications on Computer Vision, 2003, pp. 132-136.
20. Hansen, D.W.; Hansen, J.P. Robustifying Eye Interaction. Proceeding of Conference on Vision for Human Computer Interaction, 2006, pp. 152-158.
21. Yiu, Y.H.; Aboulatta, M.; Raiser, T.; Ophey, L.; Flanagin, V.L.; Eulenburg, P.Z.; Ahmadi, S.A. DeepVOG: Open-source pupil segmentation and gaze estimation in neuroscience using deep learning. *Journal of Neuroscience Methods* 2019, vol. 324, 108307.
22. Li, D.; Winfield, D.; Parkhurst, D.J. Starburst: a hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005, p. 79.
23. Fuhl, W.; Santini, T.; Kasneci, G.; Kasneci, E. Pupilnet: Convolutional Neural Networks for Robust Pupil Detection. 2016,, arXiv:1601.04902 .
24. Villanueva, A.; Ponz, V.; Sanchez, S.; Ariz, M.; Porta, S.; Cabeza, R. Hybrid method based on topography for robust detection of iris centre and eye corners., *ACM Trans. Multimedia Comput. Commun. Appl*. 2013, vol. 04, no. 25, pp. 25:1-20, doi: http://doi.acm.org/10.1145/2501643.2501647.
25. Fuhl, W.; Kübler, T,; Sippel, K.; Rosenstiel, W.; KasneciExcuse, E. Robust pupil detection in real-world scenarios. Azzopardi, G.; Petkov. N (Eds.), *Computer Analysis of Images and Patterns, Springer International Publishing*, 2015, pp. 39-51.
26. Liu, Y.; Hsueh, P.Y.; Lai, J.; Sangin, M.; Nussli, M.A.; Dillenbourg, P. Who is the expert? Analyzing gaze data to predict expertise level in collaborative applications. IEEE International Conference on Multimedia and Expo, 2019, pp. 898-90.
27. Marshall, J.S.P. Identifying cognitive state from eye metrics Aviation", *Space, and Environmental Medicine* 2007, vol. 78, no. 05, pp. 165-175
28. Henderson, J.M.; Shinkareva, S.V.; Wang, J.; Luke, S.G.; Olejarczyk, J. Predicting cognitive state from eye movements. *PLoS One* 2013, vol. 08.

29. Król, M.; KrólA, M.E. Novel approach to studying strategic decisions with eye-tracking and machine learning. *Judgment and Decision Making* 2017, vol. 12, no. 06, pp. 596-609
30. Steichen, B.; Conati, C.; Carenini, G.; Inferring visualization task properties, user performance, and user cognitive abilities from eye gaze data. *ACM Transaction on Interactive Intelligent Systems (TiiS)* 2014, vol. 04, no. 02.
31. Shojaeizadeh, M.; Djamasbi, S.; Paffenroth, R.C.; Trapp, A.C. Detecting task demand via an eye tracking machine learning system. *Decision Support Systems* 2019, vol. 116, pp. 91-101.
32. Krafka, K.; Khosla, A.; Kelnhofer, P.; Kannan, H.; Bhandarkar, S.; Matusik, W.; Torralba, A. Eye tracking for everyone", IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016, doi: 10.1109/CVPR.2016.239.
33. Naqvi, R.A.; Arsalan, M.; Batchuluun, G.; Yoon, H.S.; Park, K.R. Deep Learning-based gaze detection system for automobile drivers using a NIR camera sensor. *Sensors* 2018, vol. 18, no. 02.
34. Vera-Olmos, F.J., Melero, H.; Malpica, N. DeepEye: Deep convolutional network for pupil detection in real environments. *Integrated Computer-Aided Engineering, IOS Press* 2019, vol.26, no. 01, pp. 85-95, doi: 10.3233/ICA-180584.
35. Fuhl, W.; Santini, T.; Kasneci, G.; Kasneci, E. PupilNet: Convolutional Neural Networks for Robust Pupil Detection. ArXiv, abs/1601.04902. 2016, doi: https://arxiv.org/pdf/1601.04902.pdf.
36. Chen, S.; Liu, C. Eye detection using discriminatory Haar features and a new efficient SVM. *Image and Vision Computing* 2015, vol. 33, pp. 68-77, doi: https://doi.org/10.1016/j.imavis.2014.10.007.
37. Borza, D.; Itu, R.; Danescu, R. In the Eye of the Deceiver: Analyzing Eye Movements as a Cue to Deception. *Journal of Imaging, MDPI* 2018, vol. 04, no. 10, pp. 1-20, doi: https://doi.org/10.3390/jimaging4100120.
38. Timm, F.; Barth, E. Accurate eye centre localisation by means of gradients. *Visapp*, 2011, vol. 11, pp. 125–130.
39. King, D.E. Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.* 2009, vol. 10, pp. 1755–1758.
40. Valenti, R.; Gevers, T. Accurate eye centre location and tracking using isophote curvature. Proceeding of IEEE, CVPR, Alaska, 2008, pp. 1–8.
41. Cristinacce, D.; Cootes, T.; Scott, I. A Multi-Stage Approach to Facial Feature Detection. Proceeding of British Machine Vision Conf. BMVA Press, 2004, pp. 231-240, doi:10.5244/C.18.30.
42. Cootes, T. Talking Face Video. Available online: http://www-prima.inrialpes.fr/FGnet/data/01-TalkingFace/talking_face.html (accessed on 16 October 2019).
43. Jesorsky, O.; Kirchberg, K.J.; Frischholz, R.W. Robust face detection using the hausdorff distance. Proceeding of International Conference on Audio-and Video-Based Biometric Person Authentication, Halmstad, Sweden, 2001, pp. 90–95.
44. Ponz, V.; Villanueva, A.; Cabeza, R. Dataset for the evaluation of eye detector for gaze estimation. ACM Conf. on Ubiquitous Computing 2012, pp. 681–684.
45. Asadifard, M.; Shanbezadeh, J. Automatic adaptive centre of pupil detection using face detection and cdf analysis. Proceeding of IMECS, Hong Kong, 2010, vol. 01, pp. 130–133.
46. George, A.; Routray, A. Fast and Accurate Algorithm for Eye Localization for Gaze Tracking in Low Resolution Images. In IET Computer Vision, 2016, vol. 10, no. 07, pp.660-669.
47. Fusek R.; Dobeš P. Pupil Localization Using Self-organizing Migrating Algorithm. 2020. In: Zelinka I.; Brandstetter P.; Trong Dao T.; Hoang D.V.; Kim. S. (eds) AETA 2018 - Recent Advances in Electrical Engineering and Related Sciences: Theory and Application. AETA 2018. Lecture Notes in Electrical Engineering, Springer, Cham, vol. 554, pp. 207-216, doi: https://doi.org/10.1007/978-3-030-14907-9_21.