


## A comprehensive assessment of deep learning techniques for eye gaze estimation: A comparative performance analysis



 Hao Xu<sup>1+</sup>

 Masitah Ghazali<sup>2</sup>

 Nur Zuraifah Syazrah Othman<sup>3</sup>

<sup>1,2,3</sup>Faculty of Computing, Universiti Teknologi Malaysia, Malaysia.

<sup>1</sup>Email: [xuhao@graduate.utm.my](mailto:xuhao@graduate.utm.my)

<sup>2</sup>Email: [masitah@utm.my](mailto:masitah@utm.my)

<sup>3</sup>Email: [zurairah@utm.my](mailto:zurairah@utm.my)



(+ Corresponding author)

### ABSTRACT

#### Article History

Received: 22 May 2025

Revised: 5 August 2025

Accepted: 28 August 2025

Published: 17 September 2025

#### Keywords

AI

Computer vision

Eye gaze estimation

Eye tracking

Human recognition

Machine learning

Simulation.

The study aims to transform Convolutional Neural Networks (CNNs) for eye gaze estimation and prediction, providing relevant data on the limitations of traditional gaze tracking systems, which are often constrained by limited environments and expensive equipment. The authors propose a dual-task approach, where gaze estimation and gaze prediction are separated to enable more granular improvements in each process. Using the MPII Gaze dataset, collected under real-life conditions, various CNN architectures such as YOLOv3, SSD, and Mask R-CNN are evaluated and compared based on accuracy, precision, recall, and F1-measure. Each unique spatiotemporal sequence of eye images is utilized to enhance the predictive power of individual frames, allowing the model to identify temporal patterns and improve estimation through gaze continuity. Additional measures to increase model robustness and responsiveness include image normalization, region-of-interest extraction during preprocessing, and a geometric features-based blink detection mechanism. The results demonstrate that deep learning models can effectively improve gaze estimation accuracy under varying lighting conditions, head movements, and user diversity. This makes the technology applicable in fields such as education, medicine, automotive safety, adaptation, assistive technologies, and human-computer interaction. Overall, this research contributes to the development of scalable, adaptable, and precise gaze-tracking algorithms utilizing state-of-the-art automated learning methods, offering valuable insights for researchers in the field.

**Contribution/ Originality:** This work advances gaze estimation using CNNs by enabling multi-user, unconstrained environments, introducing a dual-task (estimation and prediction) framework, surveying CNN architectures, using eye-image sequences for temporal prediction, and highlighting real-world applications in fields such as security, education, gaming, and healthcare.

## 1. INTRODUCTION

Eye tracking has the potential to reach transformational levels in the lives of users with motor constraints by providing an easy and direct method of interaction or by delivering precise information on the attention levels of these users. Traditionally, individuals with impairments relied solely on costly, specialized hardware to access the expanding opportunities offered by eye-tracking technology. Often, these systems were used in isolation and required significant funding or a laboratory environment for psychological experimentation. The most significant finding is the ability to determine what a person is looking at based on a single snapshot, without the need for additional details or depth perception [1]. Knowing how to gaze is essential in interpersonal communication, and it is important because it fulfills different roles, which include: analysis of communication, diagnosis of illnesses, health appraisal, and disease diagnosis.

The past decades have been characterized by the impressive and steady growth of computer vision into various application domains, which is why the issue of eye gaze estimation and prediction now comes to the forefront. The development of deep neural networks over the last ten years has changed machine learning and gaze tracking. Appearance-based systems make direct judgments of gaze direction based on deep convolutional networks (CNNs) against the camera frame of reference. The study in question explores and analyzes different designs of CNNs in relation to gaze estimation and prediction, to which two tasks were applied, each concerning eye assessment [2, 3].

The potential of human gaze estimation as a method in the setting of Human-Computer Interaction (HCI) and computer vision applications that identify what are the points of interest among users cannot be underestimated. The influence of deep learning is especially noted in the gaze estimation literature. Gaze estimating systems have transformed single-user, constrained settings to multi-user, unconstrained ones by deploying deep learning to complex tasks with high variability since they can be unconstrained [4]. The study extensively covers both single-user and multi-user gaze estimation in deep learning in terms of state-of-the-art approaches, available datasets, coordinate systems, environmental limitations, architecture of deep learning models, and their performance evaluation metrics.

The other important result of this survey is the determination of limitations, challenges, along with the possible uses of multiuser gaze estimating techniques. In addition, this publication can be considered a useful resource that includes all the suggested models. It proceeds further in discussing machine learning methods of eye gaze evaluation and discusses the common pitfalls of the task of human-computer interactions and other behavioral evaluations [5]. The purpose is to communicate about the different types of models in estimating gaze eyes and presentation of the predicted outcomes. Eye fixations are driven by visual landmarks in non-constrained environments, and new methods based on appearances have been successful compared to methods based on features and models in non-constrained real-world settings. These methods are versatile to visual artifacts and imperfect light conditions.

The main research gap that this study addressed is that traditional gaze estimation techniques lack flexibility in multi-user, unconstrained situations. Although earlier models have shown promise in controlled environments, their performance declines when applied to real-world scenarios due to variations in lighting, head pose, or user behavior.

The objective of this study is to implement a dual-task approach as well as spatiotemporal data in evaluating CNN-based architectures for predicting and estimating gaze under natural conditions. Through this, the study aims to address the following two primary research questions:

1. What are the performances of CNN-based models compared to traditional algorithms in unconstrained gaze estimation environments?
2. Does the inclusion of spatiotemporal features have an effect on gaze prediction accuracy?

The same as it is depicted in Figure 1, it presents a framework of an eye gaze estimation model.

## 2. LITERATURE REVIEW

In recent years, a significant amount of research has been dedicated to the design of eye gaze estimation systems utilizing deep learning strategies. These systems are based on advanced neural networks, where, in the case of intelligent parking systems, image analysis plays a crucial role with cloud layer perception. Various algorithms, including path algorithms, hash algorithms to represent paths, and Zigbee detection in manual parking systems, have been proposed. The highest accuracy achieved is 25%, with a quotient of 6.4. A second area of investigation involves the effects of different statistical algorithms using the Florence Mobility Data as a comparison tool. It is important to note that Bayesian neural network algorithms demonstrated excellent performance, achieving 94.9% accuracy on various historical datasets [6].

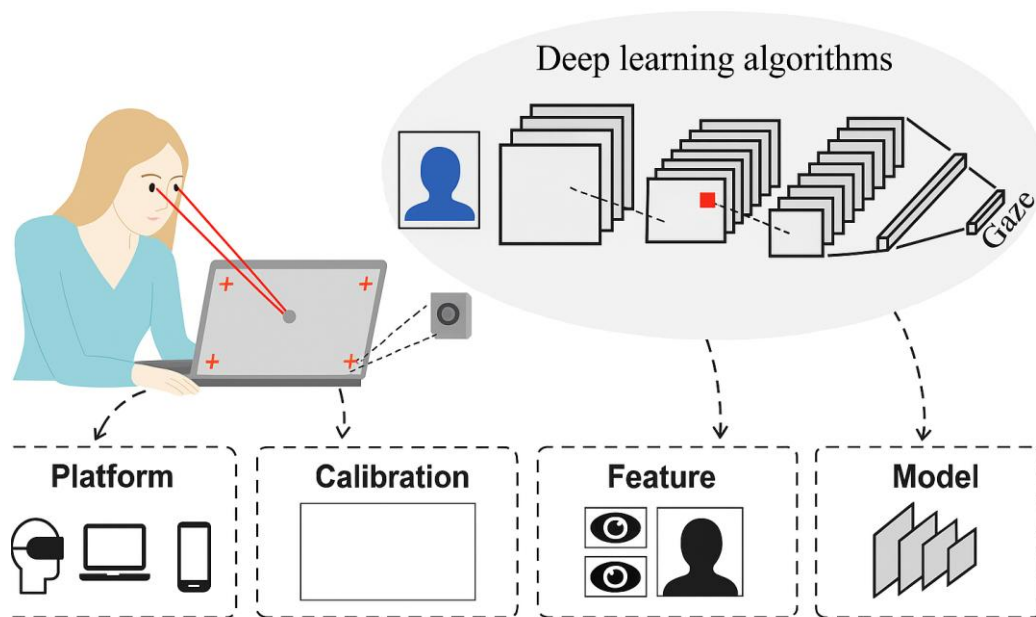


Figure 1. Framework of eye gaze estimation model.

Combination of machine learning and the Internet of Things (IoT) has been investigated to predict eye tracking [7] used advanced edge computing with the IoT to predict AI analysis. The random forest and decision tree algorithms have been utilized, demonstrating high performance levels even when compared to other state-of-the-art algorithms. Also, there are works that involve neural networks like Vicnet and Alexa Net to predict parking slot occupancy; there, the stochastic gradient descent algorithm has shown a high accuracy of 99 percent on alien data [8, 9].

Research on the definitions of optimal parking space identification and utilization within a public area was performed with the help of different algorithms. Khan and Lee [10] used the alphanumeric click algorithms that are connected to Euclidean distance and Ward method, where they employed a density-based clustering algorithm to detect parking spaces along horizontal lines of pedestrians in the streets [11, 12]. A gaze estimation approach based on deep learning was proposed, focusing on Convolutional Neural Networks (CNNs). The research introduces a new neural network model capable of modifying visual features and incorporates a data fusion approach for merging two or more gaze datasets. The trained weights, architectures, and datasets are available for building CNN-like models in the fields of eye-gaze tracking and classification [13].

Lim et al. [7] proposed a CNN-based end-to-end gaze estimation mechanism of near-eye display to address the problem of precise gaze estimation. Based on a collection of photos of people with their eyes focused on the calibration points of tablets, they built one of the basic CNNs. It is a modeling of the LeNet architecture. Results generated by the model receive the user images as input, and the direction of gaze is determined based on the x and y coordinate positions on the screen, with the problem viewed as a multi-class classification issue. However, this method showed good results, even though the error in the calculation of angles was 6.7 degrees when the recorded dataset was used. The next method is the one used by Khan and Lee [10] and Kumar and Ponnusamy [14], taking into consideration the face-representation method controlled by shape and intensity data, which demonstrates the possibility of correct identification of faces under poor-quality picture scenarios.

Deep learning methods for gaze estimation have also been optimized in recent research. These two studies include Xia et al. [12] on the analysis of hybrid neural network and logistic regression models, and Troya et al. [15] on the influence of eye gaze on clinical usage through computer-aided detection systems. Kumar and Ponnusamy [14]

emphasized the power of CNN-based models in medical scenarios, supporting the use of CNN models in the medical field of biomarkers. Recent studies demonstrate the increasingly popular trend of using CNNs in various gaze-tracking tasks and highlight the relevance of our comparative work.

To sum up, the studies indicate that the further development of eye gaze estimation systems is driven by various methodologies, algorithms, and technologies aimed at enhancing their precision and ability to operate effectively in real environments.

### 3. STATE OF ART ANALYSIS

Eye trackers are critical in various disciplines, including studies on cognitive neuroscience, psycholinguistics, visual perception, and product design. The evaluation of eye movements can be obtained through different methods. One popular approach involves recording videos that capture specific eye positions, as [16] explains.

#### 3.1. Focuses on Eye Can Be Appointed Under

Eye disclosure entails the process of identifying eye areas under conflicting face images, whereby the main objective is to establish the position of the eyes. In commonplace situations of eye identification, the areas where both eyes are located are either visible or highly separated. Such areas of the eyes are therefore traditionally depicted as rectangles, as in the studies conducted by Kottwani and Kumar [8].

Detailed feature extraction is intended to yield detailed information, such as the shape of the visible eyeball area, the outline of the iris and pupil, the position of the pupil in the visible area of the eye, and the condition of the eye (e.g., closed or open). Such operations in the computer vision area are very difficult because of the variables in environmental conditions, and the results of such activity can be vulnerable to failure. The eye region detection has a large amount of literature devoted to it some of which involves eye pupil movement recognition, eye feature extraction, eye state classification, and eye gaze recognition. A different set of methods has been studied in both still images and video sequences regarding real-time implementation as illustrated by Pathirana, et al. [17].

One possible technology that may be used to measure the retina resting potential is electrooculography (EOG). Despite a large potential difference between the cornea and the fundus (approximately 1mV), the eye has the capability of recording small voltages in the surrounding region, which depend on the position of the eye. The electrodes are tactically located above and below the eye or on the left and right sides of the eye to ensure that horizontal and vertical movements can be recorded independently. The EOG signal is generated by the dynamics of the eye's position even in the absence of actual movement and is affected by other factors, e.g., adapting to darkness and metabolic variations in the eye. It can, however, be prone to drift and other forms of variability, such as the quality of the contact between electrodes and the skin, and its course may even depend on the rate of eye movement [15].

Images techniques are based on the use of video and image processing technologies to automatically infer eye locations in photographs. Other systems use what are known as Purkinje images, which are reflections the light source bounces off the different surfaces in the eye, to be able to find the location of the eye by observing the relative movements of the images. Video images are normally used and linked to computer software in order to determine the position of the pupil and the measurement of the vertical movement and horizontal movement of the eyes. Nevertheless, lower temporal resolutions may be observed in image-based techniques relative to infrared (IR) techniques.

Like Haar, the use of digital image features named "Haar-like" features are used in object recognition. These characteristics were first used in the first real-time face detector, and they helped solve the computational complexity incurred by using basic image intensities. It is discussed that instead of using regular image intensities, a different set of features based on Haar wavelets shall be used. Haar-like features evaluate neighboring rectangular compartments in a detection window and combine their pixel intensities, finding the difference. This contrast is then applied to the grouping of the image, and this is observed by Popelka et al. [13].

### 3.2. Features of Real-Time Eye Gaze Tracking

Powerful cognitive connections are also essential in the process of continuous eye tracking and gaze assessment. Having modern edge computer vision and strong AI on-device provides a large capacity for eye-tracking abilities to perform gaze-based evaluations. The operations include face detection to determine the location of faces, head pose to be used in the development of gaze assessment models, facial landmarks where faces are detected and key points to determine the localization of the eye region, and eye state that identifies the type of state in identified faces, e.g., open or closed eyes. The ongoing assessment is carried out with real-time video sources from surveillance cameras or webcams/USB cameras, and edge AI processing and calculation accuracy guarantee safe computations on the device and remain consistent in both online and offline cases. The use case of eye gaze estimation systems with deep learning can be appreciated in the study by Poomhiran, et al. [18].

Although continuous eye tracking is a complex technology that previously required complex hardware due to the computing resources it needed, recent innovations in fields such as deep learning and edge AI have enabled it to be performed with high accuracy without the need for complicated hardware. Therefore, eye tracking has become applicable and implemented in real-life through several applications:

**Large-scale implementation:** The usage of deep learning algorithms makes the intricate user customization unnecessary. Such models are more robust than other solutions and are stable under circumstances of less-than-ideal image quality and changing lighting.

**Operational efficiency:** Automated eye tracking contributes to stress and attention monitoring, and these two aspects determine the subsequent quality of products/services [19, 20].

**Accident Prevention:** Real-time monitoring of the driver: distractions and loss of attention on the road (e.g., when using a mobile phone, eating, or drinking).

**Security Systems:** Eye gaze tracking has some special applications in the field, allowing measurement of levels of autonomy used in the transportation and manufacturing industries.

**Cost Savings:** The technology promises to provide cost savings in terms of low insurance costs, avoiding accidents, and evading charges (such as ensuring that drivers have proper rest periods). Continuous eye tracking may be used in numerous ways that indicate its depth and efficiency in a number of domains.

Various parametric measures are explicated in Table 1 according to a number of state-of-the-art [10].

**Table 1.** A review of different parametric measures.

Approaches	Parametric measures					Security elements		
	Colossal degree execution	Practical productivity	Disaster revolution	Security systems	Cost speculation supports	Privacy	Integrity	Safety
Kar and Corcoran [2]	✓			✓		✓		
Kanade, et al. [5]	✓			✓		✓		✓
Kottwani and Kumar [8]	✓	✓	✓	✓				
Rubies, et al. [9]	✓	✓		✓				
Inoue, et al. [11]	✓	✓	✓					
Popelka, et al. [13]	✓	✓		✓		✓		✓
Xia, et al. [12]				✓				✓
Lim, et al. [7]					✓		✓	✓

## 4. PROPOSED METHOD

We trained a CNN-based dual-task model to estimate the direction of the gaze by encoding gaze estimation and gaze prediction separately. In comparison with earlier works that foresee single-frame gaze recognition, we leverage spatiotemporal sequences and sophisticated object detection. The state-of-the-art pretrained convolutional neural networks (YOLOv3, SSD, and Mask R-CNN) were used to extract eye and face representations, whereas strong

resistance to variations in head poses and blinking artifacts was provided through blink detection and coordinate transformations. Such a framework is unique in that it combines appearance-based learning and temporal information to enhance the accuracy of recognition in real-world, unconstrained environments.

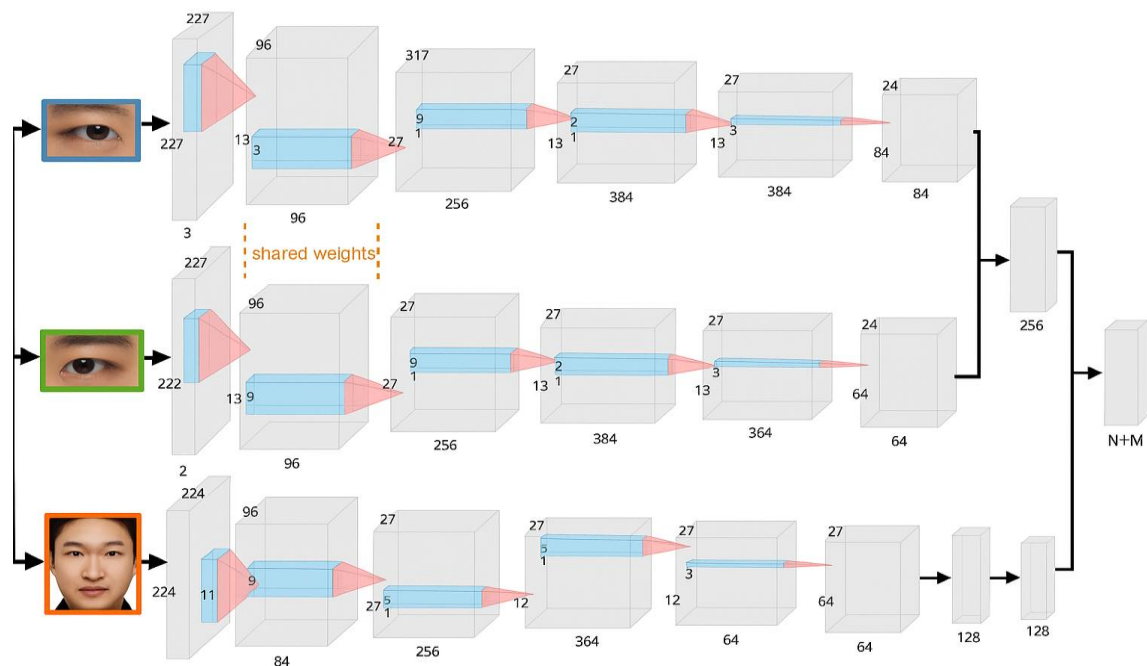


Figure 2. Proposed framework.

This paper is devoted to the evaluation of brain connectivity using the patch of face and eyes, which requires the preprocessing stage as shown in Figure 2. In this procedure, we recognized two eyes and a face in every photo. An open framework using Haar Cascade Classifiers (applied in OpenCV) was used to categorize the edges [13]. The information generated included the location of the top-left corner, the width and height of both eyes, as well as the facial bounding box. To account for changes in perspective, the images of the right eye, left eye, and face were adjusted considering the size of the bounding boxes. The images were then resized back to their original dimensions to ensure uniformity in the data.

The following two subsections outline the two-stage process in assessing the gaze according to the proposal. This method presupposes the use of brain connections to produce two-way assessments and improve the check. It is important to note that the process of critical reversal occurred during the evaluation stage. These layers of convolution have two similar layers below the left and right eye loads. The left and right eye features and pooling levels succeed the two convolutional layers, but the right eye's three convolutional levels work independently. The convolutional layers used to process the face levels are three, followed by a fully connected layer and a covering convolutional layer. The given methodology corresponds to the works by Xia et al. [12] and Inoue et al. [11].

#### 4.1. Model Validation

When evaluating our gaze prediction model performance, we utilized the MPII Gaze dataset, a collection of actual photos taken with web-based cameras built into laptops of 15 maternity users. The exposure and detail are some of the factors that make the image quality in the dataset variable. The core of our suggestion involves predicting eye gaze directions (pitch, yaw) based on single views of people's eyes, including a group of 10 participants, and considering the related information about head position in the training phase.

In the model assessment, we focused on the top five topics and noted the problem of variability in the visual quality of images caused by real-life variations. The performance measures include accuracy, precision, recall, and F1-

measure, which provide an overall understanding of the model's ability to predict gaze direction accurately under different conditions. The most important details of the implementation are as follows:

**D/3D Gaze Conversion:** A D/3D gaze conversion strategy was implemented to assist in transferring 2D gaze measures (on a computer screen) into 3D gaze orientations in the physical environment. This includes a mapping of a two-dimensional gaze target to a three-dimensional gaze direction, which increases the spatial awareness of the model.

**Image Preprocessing:** The first stage is preprocessing of images, where simple operations, available as part of the OpenCV library, are used to standardize the dimensions (size of the eye to 416x416 pixels). As complementary methods, techniques such as Haar cascades and libraries such as Dlib are used in the detection of eye regions.

**Selection of Model:** Different deep learning models such as YOLOv3, SSD, Mask R-CNN, and FreezNet are under consideration in gaze estimation. Difficulties involving such networks usually arise due to the requirement of high volumes of training data. Although the datasets can be used non-commercially, speed and accuracy are essential factors that define eye-tracking activities.

**Blink Detection:** We added blink detection into the model using geometric features that make the model more robust. This means marking certain points on the eyes, drawing boundaries, and studying changes in the length of the lines. The stipulation of blink detection is a factor quite important in the application of gaze tracking realities.

To sum up, the process of model validation includes extensive estimation criteria, real-world dataset considerations, and preprocessing. Blink detection further increases the versatility of the model in dynamic gaze tracking.

#### 4.2. Dataset: A Database Known as MPII Gaze

We used the MPII Gaze dataset [18] in our investigation, and it contains photos taken with the internal cameras of the laptops of 15 real users in real situations. Thus, the dataset includes variations in the quality of the images, especially concerning exposure and detail. We aim to create a predictive model that will determine the direction (pitch, yaw) of a participant's eye gaze based on single-eye photos of 10 different participants. Exploitation of head position data is incorporated during training to enhance the learning process of the model. The trained model is then strictly tested with the top 5 data sets. Figure 3 which is the Model Overview is as shown below.

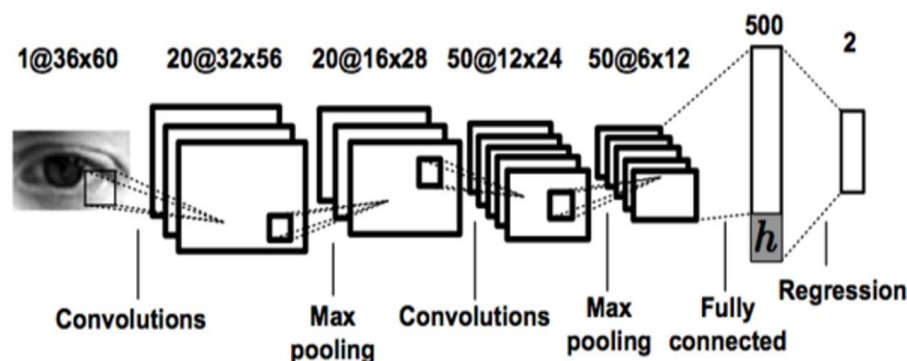


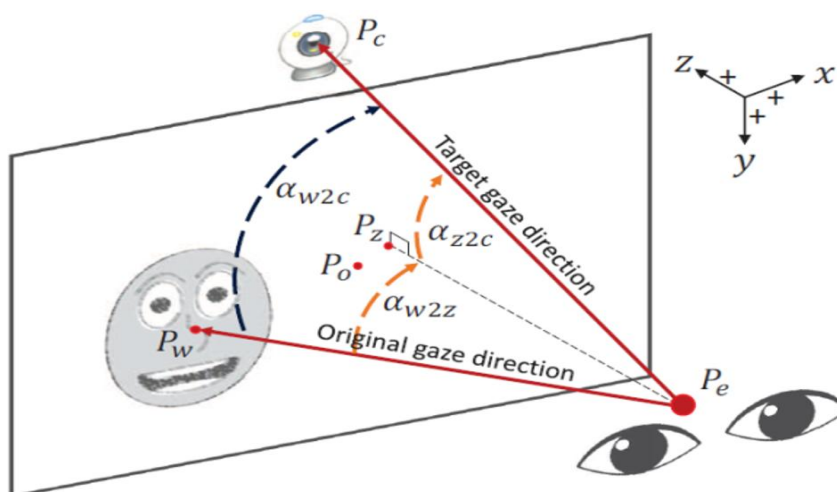
Figure 3. Model overview.

The 2D gaze assessment algorithm primarily predicts gaze points on a computer screen [13], whereas the 3D gaze assessment algorithm predicts gaze directions in three-dimensional space. We introduce the process of converting between 2D gaze and 3D gaze. Given a 2D gaze target  $p=(u,v)$  on the screen, our objective is to calculate the corresponding 3D gaze direction  $g=(g_x,g_y,g_z)$ . In the processing pipeline, we initially establish the 3D gaze target and the 3D gaze origin in the camera coordinate system (CCS).

The gaze direction can then be computed as follows:

$$S_g = (g_x, g_y, g_z) \cdot (i + u, j + v))^2 \quad (1)$$

Below is **Figure 4** Angle of coordination.



**Figure 4.** Angle of co-ordination.

The initial phase of developing a model for the specified task involves image preprocessing. Image preprocessing facilitates the extraction of valuable information, which is essential for guiding the assessment and analysis of the method's effectiveness. The most commonly used preprocessing components from the OpenCV library were applied to the eye image, with the goal of resizing the resolution to 416x416 pixels, as illustrated in [Figure 1](#). However, the results obtained using Haar cascade and the dlib library to monitor gaze position were significantly low and inaccurate. Despite this, such libraries may still serve as secondary tools for defining eye regions.

Viola and Jones [21], in a discovery in the field of artificial intelligence, invented the Haar cascade, an object detection method. During the classification process, a pre-trained Haar cascade accepts an image as input and determines whether the image contains the target object or not, classifying the input image into one of two categories. The success identification algorithm described offers an implementation of the Regression Tree Ensemble (ERT) presented by Kazemi and Sullivan [22] and Pathirana et al. [17]. This algorithm uses an efficient and fast operation to determine the position of a landmark. The same issue arises in the estimation of positions, which is again carried out iteratively by a cascade of regressors, each at an iteration attempting to minimize the alignment error of the estimated points at that iteration.

Although there are quite a few deep neural networks in the market, some of these neural networks are highly effective in tracking activities that require speed and accuracy. Here, we examined YOLOv3, SSD, Mask R-CNN, and FreezNet. One of the main challenges in deep neural networks is the necessity to employ large datasets during training. There are several training datasets available (although not all are commercial) for training neural networks.

The gaze estimation method relates to two horizontal positions defined on both eyes and setting a boundary between them. Thereafter, there is an illustration of four points, two at the upper side of the eye and two at the lower side of the eye, defining a vertical boundary.

The absolute lengths of the horizontal and the vertical lines are then calculated, and the ratio to the original circle in each eye is found out. Crossed horizontal and vertical lines are established in both eyes, left or right, and the ratio of the lines is determined. When one covers his eyes, the length of the horizontal line remains constant, but the length of the vertical line tends to zero. Hence, this ratio can be utilized to detect blinking, as demonstrated in [Figure 5 \[23\]](#) also below show is [Figure 6: Data Processing Steps](#).

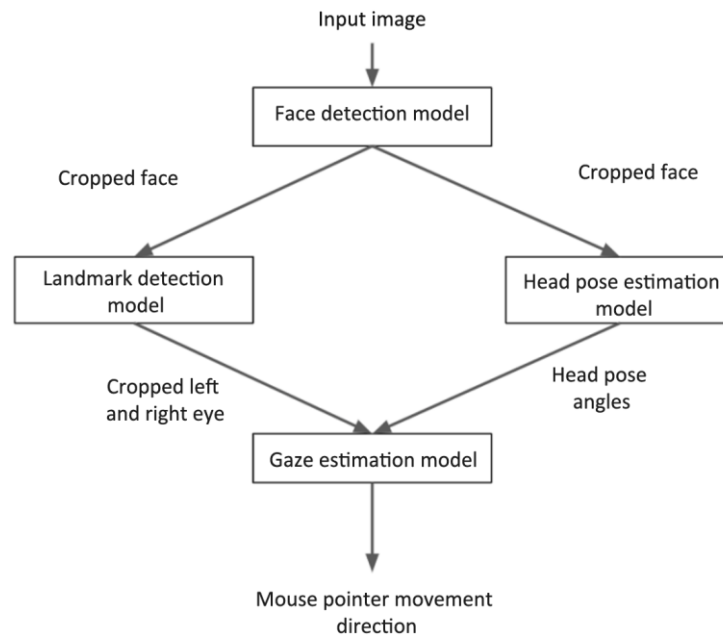


Figure 5. Pipeline of algorithm.

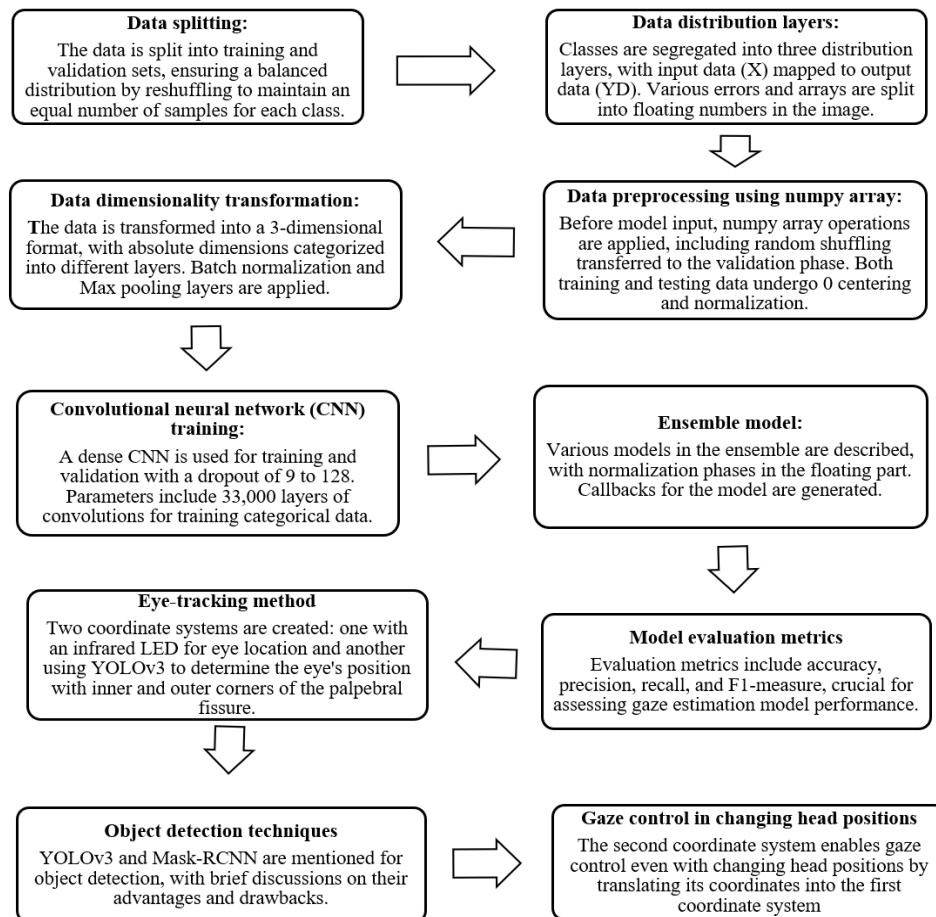


Figure 6. Data processing steps.

## 5. RESULT AND DISCUSSION

Following the visualization of the plugins and the images of training and testing data, the next step is to reshuffle the training and validation data so that data classes are nearly equal in the number of data sets. After the data has been separated into training and validation sets, the counterparts are divided following three different layers of

distribution, whereby input X results in YD, different errors, and sets of arrays defined accordingly [19, 24]. There are floating numbers showing these counterparts in the picture. The data undergoes a preprocessing procedure with the help of the array Numpy prior to feeding it into the model. The random shuffle parameter of the image is used in the validation stage, and the zero-centering and normalization of the training set are carried out for the ensemble component. Zero-centering the testing data and normalizing the data follow the same process. This three-dimensional step involves grouping the absolute dimensions of the data into various strata, resulting in 53,521 to 215,457 categorical data points. There are batch normalization layers and max pooling layers in this stage, and the training and validation are performed using a dense convolutional neural network with dropout of nine to 128. The dense layer contains 33,000 convolutional parameters to train categorical data, and the process of batch normalization is illustrated within the ensemble component of the various models.

The full model can be seen as depicted in Figures Two and Three, which show the different normalization stages in the floating region, and model callbacks are created [12].

Accuracy refers to the number of true positive tests compared to the total number of predicted positives. The formula is given below:

$$\text{Accuracy} = \text{Number of correct predictions} / \text{Total number of predictions} \quad (2)$$

Review: The number of authentic positive tests among the true positive instances is considered.

$$\begin{aligned} P/TP + FN &= TP/TP + FN = TP/TP + FN = TP/TP + FN = TP/TP + FN = TP/TP + FN = \\ &TP/TPFN = TP/TP \quad (3) \end{aligned}$$

F1-measure, the weighted average of precision and recall, and review represent the two measures. It may give precedence to pieces of information over accuracy due to the lopsidedness of the classifications. The model summary is shown in Figure 6 and Figure 7.

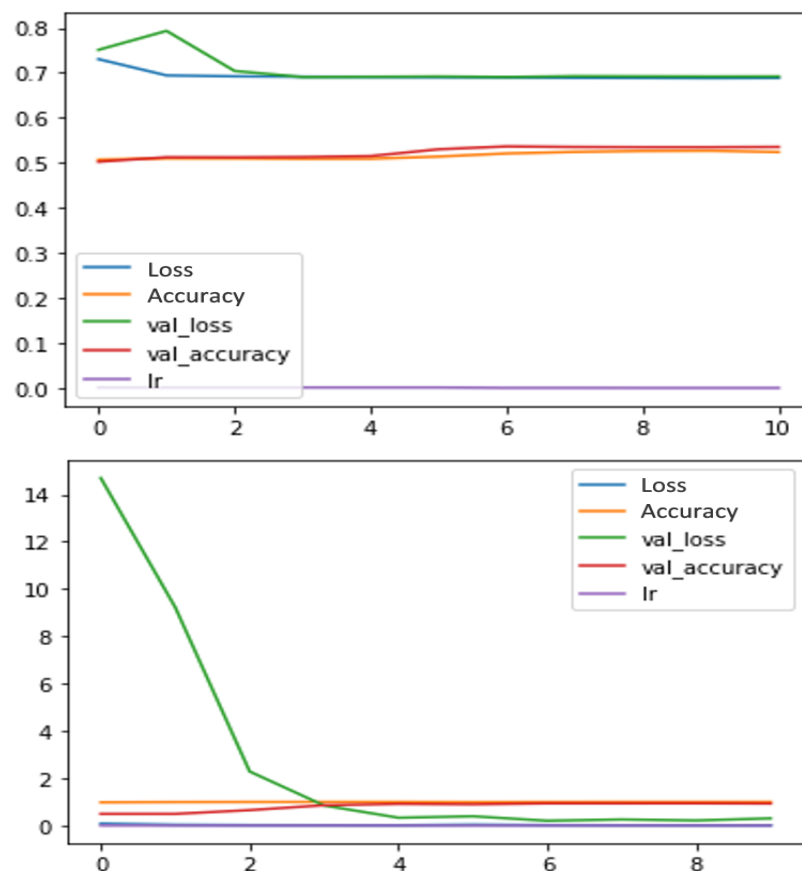


Figure 7. Model accuracy.

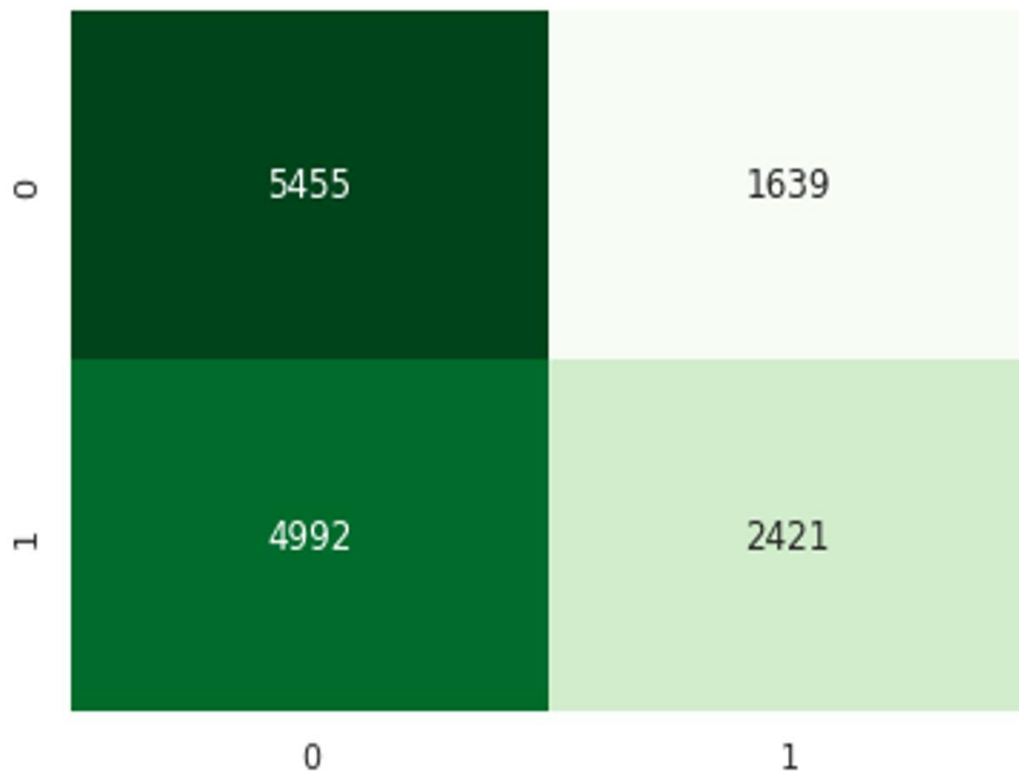


Figure 8. Confusion matrix.

The extension was attempted but proved unsuccessful. The provided directory contains boundary outboxes and masks representing detected objects within the image, revealing the highest confidence score for each prediction. Consequently, the model identified these masks as vacant spaces within the specific geographical area where the objects are depicted. Each instance of the convolution highlights the masked object in question.

Furthermore, the authors of the experiments often resort to using a conventional convolutional network for deep learning or tools from the OpenCV library, despite the growing interest in employing neural networks for image processing. A limitation of this tracking technique is the requirement for labeling. YOLOv3's network is structured in a way that allows only rectangles for labeling at a given time. In contrast, the use of Mask R-CNN allows for more accurate object tracking. However, the drawback of Mask R-CNN is its slow processing speed, limiting its capability to handle streaming video. Table 2 shows the comparison between mean fixation duration (MFD) and mean saccade amplitude (MSA).

The subsequent objective is to train the neural network entirely from scratch without relying on external weights developed using different types of images, as depicted in Figure 8 [23, 25].

Table 2. Comparison of mean fixation duration (MFD) and mean saccade amplitude (MSA).

Measure	Condition	Algorithm	Mean	SD
MFD overall	True	I-VT	214.65	±124.59
MFD overall	False	I-VT	214.9	±100.48
MFD overall	True	I-DT	197.34	±64.02
MFD overall	False	I-DT	202.8	±65.55
MSA overall	True	I-VT	12.65	±7.15
MSA overall	False	I-VT	11.9	±6.74
MSA overall	True	I-DT	8.41	±7.43
MSA overall	False	I-DT	8.07	±7.00

**Note:** MFD = Mean fixation duration (in ms), MSA = Mean saccade amplitude (in degrees or pixels, as applicable); SD = Standard deviation.

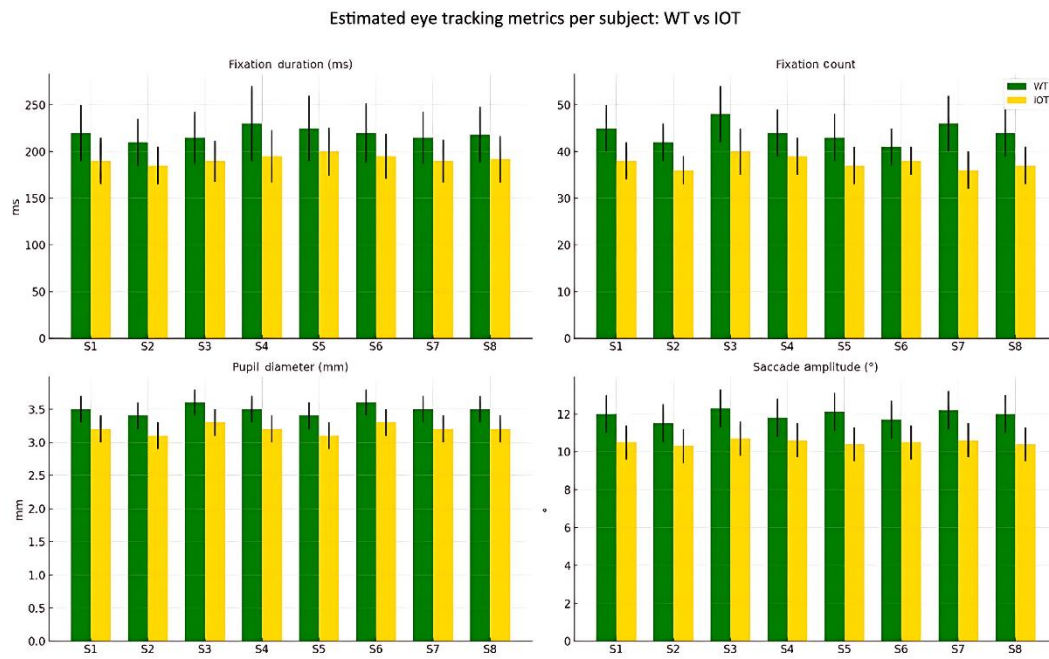


Figure 9. Scratch without using external weights developed.



Figure 10. Measures in eye gaze estimation.

The eye-tracking system we developed utilized two coordinate systems. The first coordinate system considers OpenCV and the range function in determining the position of the eye when the infrared LED is in front of the display. This system works to adjust the position of the head relative to the monitor. The location of the eye is measured at the inner and outer corners of the palpebral fissure with the help of YOLOv3 in the second coordinate system. The advantage of the second coordinate system is that it allows for gaze control even as the head moves. This is achieved through the transformation of the coordinates from the second system to the first. Figure 9 shows the schematic diagram of the transmission of eye coordinate data to the computer. Meanwhile, Figure 10 shows the measurement of eye gaze estimation provided by a model in coordinating the eye position.

These findings were consistent with the results of previous observation (e.g., [13, 17]), since CNN-based models perform better compared to rule-based models in a free environment. It is curious that our results are partly contradictory to the findings by Troya et al. [15], who observed a lower gaze detection accuracy when they changed

the illumination, but we did not see that problem when we used spatiotemporal features. Moreover, Singh et al. [19] also argued that the model should be robust in dynamic situations that our dual-task intervention managed to solve. Such comparisons outline the novelty and usefulness of our approach.

## 6. CONTRIBUTION TO RESEARCH

A number of important contributions made by this work regarding the area of gaze estimation using Convolutional Neural Networks (CNNs) are listed below. The main contributions include the following:

The standard approaches applied in gaze estimation worked only in a single-user and constrained environment, being not adaptable to a variety of circumstances. This paper has considered this shortcoming by using deep learning techniques so that the system can be used in a free operating environment where more than one user exists. This extension widens the scope of gaze estimation and demonstrates the flexibility of deep learning algorithms.

The present study adopts a dual-task approach, which involves the disintegration of the issue into two tasks: gaze estimation and gaze prediction. This method allows for a more in-depth understanding of each process, leading to improvements in accuracy and effectiveness.

The research performs an in-depth survey of different Convolutional Neural Network (CNN) architectures in the perspective of gaze estimation. Such comprehensive research can be used by future studies as a reference to understand what the optimal model for a specific activity should be, as well as to help determine the most suitable model for this activity.

A new feature of this research is the use of already existing eye-image sequences to generate eye-gaze predictions. In such a manner, the predictions remain not only dependent on the last glimpse but also have a temporal dimension, which contributes to a better estimate of the gaze.

The implications of such studies in real life are wide-ranging. The paper provides instances of moves such as security and education to demonstrate the possible applications of improved performance and predictive capabilities. Notably, the implications of the discovery are more far-reaching, as the advancements can be smoothly incorporated into the domains of gaming and health diagnostics, demonstrating their versatility.

## 7. CONCLUSION

### 7.1. Practical Implications of the Study

The developed deep learning model offers significant potential to enhance real-time eye gaze estimation by integrating it into practical applications across various fields. In healthcare, it can support diagnostic devices that monitor the visual attention of patients with neurological or cognitive issues, aiding in early diagnosis and treatment planning. Gaze tracking can be combined with intelligent tutoring systems in education to monitor engagement, comprehension, and learning behaviors in real-time. In security sectors, gaze-based systems or behavioral surveillance can be employed in high-risk situations. Such user intent and engagement insights can be utilized by eye-tracking technology to develop adaptive systems in gaming, big data analysis, European Commission projects, and human-computer interaction (HCI). The robustness of the model under diverse conditions including variable lighting, head positions, and multiple users makes it suitable for real-life, consumer-oriented applications that require minimal calibration and hardware dependence.

### 7.2. Limitations of the Study

Although the results have demonstrated effectiveness, there are several limitations of the available framework that should be noted. Extreme changes in illumination conditions, occlusions (e.g., eyeglasses or eyelashes), and oblique camera angles can also affect the performance of the model, and these are not well represented in the MPII dataset. Moreover, engineering deep learning models and tuning often require ample computation and the availability of large-scale, well-labeled data, which can restrict reproducibility or real-time implementation in low-resource

settings. The other constraint is the apparent gap between the domain used to train the models by controlling or semi-natural data and the real situations characterized by high dynamics or diverse cultures.

### 7.3. Future Research Directions

The ongoing limitations can be addressed through future research, focusing on lightweight, energy-efficient CNN or transformer-based architectures that can be deployed on mobile and embedded platforms. Enhancing the user-friendliness of proliferated systems and interfaces could be achieved by researching calibration-less gaze estimation algorithms suitable for multi-user or public settings. Generalizability will be improved by expanding datasets to include ethnically and demographically diverse samples, varying environmental conditions, and low-quality imaging settings. Additionally, incorporating real-time head-pose estimation and compensation libraries can increase resistance to user movement. Lastly, applying transfer learning and domain adaptation schemes to related visual tasks, such as facial recognition or attention modeling, can reduce training time and enhance performance, especially when large annotated gaze datasets are unavailable.

**Funding:** This study received no specific financial support.

**Institutional Review Board Statement:** The study involved minimal risk and adhered to ethical guidelines for social science fieldwork. Ethical approval was obtained from the UTM, Malaysia Research Ethics Committee (Approval No. UTMREC-2024-94). Informed verbal consent was obtained from all participants, and all data were anonymized to protect participant confidentiality.

**Transparency:** The authors state that the manuscript is honest, truthful, and transparent, that no key aspects of the investigation have been omitted, and that any differences from the study as planned have been clarified. This study followed all writing ethics.

**Competing Interests:** The authors declare that they have no competing interests.

**Authors' Contributions:** All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.

## REFERENCES

- [1] N. Zdarsky, S. Treue, and M. Esghaei, "A deep learning-based approach to video-based eye tracking for human psychophysics," *Frontiers in Human Neuroscience*, vol. 15, p. 685830, 2021. <https://doi.org/10.3389/fnhum.2021.685830>
- [2] A. Kar and P. Corcoran, "A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms," *IEEE Access*, vol. 5, pp. 16495-16519, 2017. <https://doi.org/10.1109/ACCESS.2017.2735633>
- [3] Z. Orman, A. Battal, and K. Erdem, "A study on face, eye detection and Gaze estimation," *International Journal of Computer Science & Engineering Survey*, vol. 2, no. 3, pp. 29-46, 2011. <https://doi.org/10.5121/ijcses.2011.2303>
- [4] Y. Cheng, H. Wang, Y. Bao, and F. Lu, "Appearance-based gaze estimation with deep learning: A review and benchmark," *arXiv:2104.12668*, 2021. <https://doi.org/10.48550/arXiv.2104.12668>
- [5] P. Kanade, F. David, and S. Kanade, "Convolutional neural networks (CNN) based Eye-Gaze tracking system using machine learning algorithm," *European Journal of Electrical Engineering and Computer Science*, vol. 5, no. 2, pp. 36-40, 2021. <https://doi.org/10.24018/ejece.2021.5.2.314>
- [6] Y. Wang, X. Ding, G. Yuan, and X. Fu, "Dual-cameras-based driver's eye gaze tracking system with non-linear Gaze point refinement," *Sensors*, vol. 22, no. 6, p. 2326, 2022. <https://doi.org/10.3390/s22062326>
- [7] J. Z. Lim, J. Mountstephens, and J. Teo, "Emotion recognition using eye-tracking: Taxonomy, review and current challenges," *Sensors*, vol. 20, no. 8, p. 2384, 2020. <https://doi.org/10.3390/s20082384>
- [8] A. Kottwani and A. Kumar, "Eye gaze estimation model analysis," *arXiv (Cornell University)*, 2022. <https://doi.org/10.48550/arXiv.2207.14373>
- [9] E. Rubies, J. Palacin, and E. Clotet, "Enhancing the sense of attention from an assistance mobile robot by improving eye-gaze contact from its iconic face displayed on a flat screen," *Sensors*, vol. 22, no. 11, p. 4282, 2022. <https://doi.org/10.3390/s22114282>

- [10] M. Q. Khan and S. Lee, "Gaze and eye tracking: Techniques and applications in ADAS," *Sensors*, vol. 19, no. 24, p. 5540, 2019. <https://doi.org/10.3390/s19245540>
- [11] Y. Inoue, K. Koshikawa, and K. Takemura, "Gaze estimation with imperceptible marker displayed dynamically using polarization," in *the 2022 Symposium on Eye Tracking Research and Applications*, 2022.
- [12] Y. Xia, B. Liang, Z. Li, and S. Gao, "Gaze estimation using neural network and logistic regression," *The Computer Journal*, vol. 65, no. 8, pp. 2034-2043, 2022. <https://doi.org/10.1093/comjnl/bxab043>
- [13] S. Popelka, A. Vondrakova, and P. Hujnakova, "Eye-tracking evaluation of weather web maps," *ISPRS International Journal of Geo-Information*, vol. 8, no. 6, p. 256, 2019. <https://doi.org/10.3390/ijgi8060256>
- [14] T. Kumar and R. Ponnusamy, "Robust medical x-ray image classification by deep learning with multi-versus optimizer," *Engineering, Technology & Applied Science Research*, vol. 13, no. 4, pp. 111406-11411, 2023. <https://doi.org/10.48084/etasr.6127>
- [15] J. Troya *et al.*, "The influence of computer-aided polyp detection systems on reaction time for polyp detection and eye gaze," *Endoscopy*, vol. 54, no. 10, pp. 1009-1014, 2022. <https://doi.org/10.1055/a-1770-7353>
- [16] I. S. Shehu, Y. Wang, A. M. Athuman, and X. Fu, "Remote eye gaze tracking research: A comparative evaluation on past and recent progress," *Electronics*, vol. 10, no. 24, p. 3165, 2021. <https://doi.org/10.3390/electronics10243165>
- [17] P. Pathirana, S. Senarath, D. Meedeniya, and S. Jayarathna, "Eye gaze estimation: A survey on deep learning-based approaches," *Expert Systems with Applications*, vol. 199, p. 116894, 2022. <https://doi.org/10.1016/j.eswa.2022.116894>
- [18] L. Poomhiran, P. Meesad, and S. Nuanmeesri, "Improving the recognition performance of lip reading using the concatenated three sequence keyframe image technique," *Engineering, Technology & Applied Science Research*, vol. 11, no. 2, pp. 6986-6992, 2021. <https://doi.org/10.48084/etasr.4102>
- [19] C. Singh, T. Imam, S. Wibowo, and S. Grandhi, "A deep learning approach for sentiment analysis of COVID-19 Reviews," *Applied Sciences*, vol. 12, no. 8, p. 3709, 2022. <https://doi.org/10.3390/app12083709>
- [20] S. M. Usha and H. B. Mahesh, "Monitoring and analysis of agricultural field parameters in order to increase crop yield through a colored object tracking robot, image processing, and IoT," *Engineering, Technology & Applied Science Research*, vol. 12, no. 4, pp. 8791-8795, 2022. <https://doi.org/10.48084/etasr.5028>
- [21] P. Viola and M. Jones, "Fast and robust classification using asymmetric AdaBoost and a detector cascade," *Advances in Neural Information Processing Systems*, vol. 14, pp. 1311-1318, 2001.
- [22] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1867-1874.
- [23] Y. Erel, C. E. Potter, S. Jaffe-Dax, C. Lew-Williams, and A. H. Bermano, "iCatcher: A neural network approach for automated coding of young children's eye movements," *Infancy*, vol. 27, no. 4, pp. 765-779, 2022. <https://doi.org/10.1111/inf.12468>
- [24] S. Tanaka, A. Tsuji, and K. Fujinami, "Poster: A preliminary investigation on Eye Gaze-based concentration recognition during silent reading of text," presented at the 2022 Symposium on Eye Tracking Research and Applications, 2022.
- [25] J. P. Prasad, "IRJET- implementation of machine learning algorithm in Eye-Gaze tracking using WSN based convolutional neural networks," Retrieved: [https://www.academia.edu/45598716/IRJET\\_Implementation\\_of\\_Machine\\_Learning\\_Algorithm\\_in\\_Eye\\_Gaze\\_Tracking\\_using\\_WSN\\_based\\_Convolutional\\_Neural\\_Networks](https://www.academia.edu/45598716/IRJET_Implementation_of_Machine_Learning_Algorithm_in_Eye_Gaze_Tracking_using_WSN_based_Convolutional_Neural_Networks), 2021.

*Views and opinions expressed in this article are the views and opinions of the author(s), Journal of Asian Scientific Research shall not be responsible or answerable for any loss, damage or liability etc. caused in relation to/arising out of the use of the content.*