Expert Systems With Applications 199 (2022) 116894

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Eye gaze estimation: A survey on deep learning-based approaches

Primesh Pathirana [a], Shashimal Senarath [a], Dulani Meedeniya [a,*], Sampath Jayarathna [b]

[a] Department of Computer Science and Engineering, University of Moratuwa, Moratuwa, 10400, Sri Lanka
[b] Department of Computer Science, College of Science, Old Dominion University, Norfolk 23529, USA

ARTICLE INFO

Keywords:
Computer vision
Gaze estimation
Deep learning
Eye tracking

ABSTRACT

Human gaze estimation plays a major role in many applications in human–computer interaction and computer vision by identifying the users' point-of-interest. Revolutionary developments of deep learning have captured significant attention in gaze estimation literature. Gaze estimation techniques have progressed from single-user constrained environments to multi-user unconstrained environments with the applicability of deep learning techniques in complex unconstrained environments with extensive variations. This paper presents a comprehensive survey of the single-user and multi-user gaze estimation approaches with deep learning. State-of-the-art approaches are analyzed based on deep learning model architectures, coordinate systems, environmental constraints, datasets and performance evaluation metrics. A key outcome from this survey realizes the limitations, challenges and future directions of multi-user gaze estimation techniques. Furthermore, this paper serves as a reference point and a guideline for future multi-user gaze estimation research.

1. Introduction

Eye gaze plays an important role in identifying the users' point of interest in terms of the direction, location, attention, emotions and interactions. Generally, human gaze estimation is a frequently used approach to gain a better understanding of human cognition and behavior. Many studies have addressed the approaches to trace the position and direction of eye gaze, which is required for different domains like cognitive (Chong et al., 2020), social behavior (Kodama et al., 2018; Sugano et al., 2016), medical health (De Silva et al., 2021, 2019), commercial (Bermejo et al., 2020; Sugano et al., 2016) and other human–computer interaction applications (Zhang et al., 2015). Additionally, gaze estimation environments can be classified as constrained (controlled) or unconstrained (wild). Constrained environments are those that have a fixed set of parameters, such as illumination, subject count and head-angle variation. On the other hand, unconstrained environments are those with a considerable measure of parameter variation. It is clear that with the widespread use of gaze estimating technology across many application domains, gaze estimation has progressed more into unconstrained environments, surpassing constrained environment settings.

Although several eye gaze estimation solutions are available, some of them incur aspects such as expensiveness, requirement of manual interventions, unreliability and inaccuracy in practical deployments. Also, the performance of some traditional approaches is limited by factors such as low image quality and light conditions. In such scenarios, Deep Learning (DL) based eye gaze estimation approaches come into play due to the inherited benefits, such as learning from existing data, automation, flexible process, high accuracies and better decision making. These prevalence DL based approaches have shown success in performance improvements in eye gaze applications.

Human gaze estimation approaches fall into two broad categories: model-based techniques and appearance-based techniques. Model-based methods fundamentally require dedicated devices such as near-infrared (NIR) cameras to manually regress the eye features and build a geometric model (Cheng et al., 2021; Kar & Corcoran, 2017). This method is person-specific and restricted to constrained environments (Akinyelu & Blignaut, 2020; Cheng et al., 2021). In comparison, appearance-based techniques do not necessitate dedicated devices and are not limited to constrained environments. These methods can be subdivided into two categories, namely conventional appearance-based methods and appearance-based methods with DL.

Over the last decade, there has been a surge of interest in eye-tracking literature related to gaze estimating methods based on DL techniques due to their applicability and robustness in unconstrained environments. In contrast to conventional appearance-based methods, DL-based methods exhibit many benefits, such as the ability to extract high-level gaze features from images and the ability to learn a non-linear mapping function directly from the image to eye gaze (Cheng

* Corresponding author.
E-mail addresses: primeshs.17@cse.mrt.ac.lk (P. Pathirana), shashimalsenarath.17@cse.mrt.ac.lk (S. Senarath), dulanim@cse.mrt.ac.lk (D. Meedeniya), sampath@cs.odu.edu (S. Jayarathna).

https://doi.org/10.1016/j.eswa.2022.116894
Received 28 August 2021; Received in revised form 15 January 2022; Accepted 11 March 2022
Available online 21 March 2022
0957-4174/

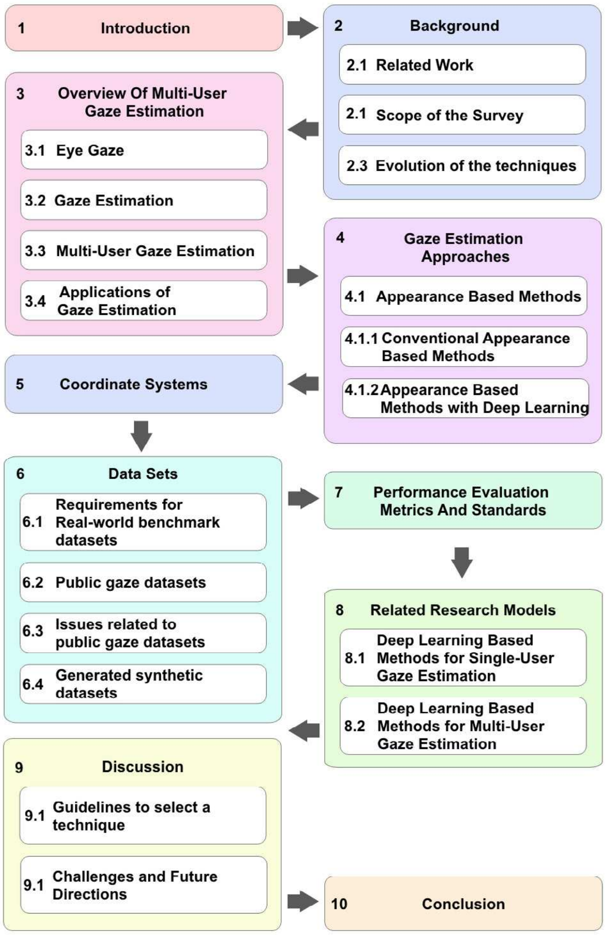


Fig. 1. Structure of the paper.

et al., 2021; Kellnhofer et al., 2019). Deep convolutional neural networks (DCNN) have been utilized in almost every DL-based gaze estimation approach due to their ability to map image features directly, handle large-scale datasets, learn complex non-linear mappings when faced with significant head-pose variations, eye occlusions and illumination conditions.

Appearance-based methods with DL, which is the main focus of this study, can be divided further into two subcategories based on the number of subjects, namely single-user gaze estimation and multi-user gaze estimation. Despite the significant shift in gaze estimation techniques towards applications in unconstrained environments, the demand for multi-user gaze estimation approaches is on the rise. As of the end of year 2021, a limited number of such methods have been researched, specifically regarding time-shifting and space-shifting single-user gaze estimation.

This survey paper explores state-of-the-art methods and techniques used in eye gaze estimation research. We analyze the use of the latest DL techniques, useful public datasets and different approaches used by related studies. The lessons learned from this survey state that the eye gaze applications are evolving with the use of DL techniques due its the inherited benefits. Moreover, this study suggests guidance to follow a DL based process for eye gaze estimation that can be used as a reference. Further, we discuss the challenges and future research directions in eye gaze estimation in several applications. Thus, we aim to inspire the researchers and developers with useful insights to produce effective and efficient eye gaze estimation applications using DL techniques.

Fig. 1 states the survey structure considered for this article, which focuses on the single and multi-user gaze estimation methods in DL. Section 1 states the survey motivation and main contributions of this research. Section 2 explains the scope of the survey and discusses the background of current related studies. Section 3 provides an overview of multi-user gaze estimation by discussing the history, progression and applications of gaze estimation. Section 4 broadly discusses the technical aspects of existing gaze estimation approaches, focusing on appearance-based methods with DL. In Sections 5–7, we present the supplementary knowledge in gaze estimation literature by reviewing the theoretical concepts behind the coordinate systems, describing different gaze datasets and discussing state-of-the-art performance evaluation metrics. Section 8 elaborates and critically analyses the existing single-user and multi-user gaze estimation approaches by summarizing their key outcomes and limitations. Section 9 suggests guidance to select a given approach based on different conditions and discusses the limitations, challenges, future direction in multi-user gaze estimation literature. Finally, Section 10 concludes the study.

2. Background

2.1. Related work

Among many studies that have focused on eye gaze estimation research, only a few survey studies are available that discuss growing aspects in the literature focused on DL techniques. Table 1 summarizes the features addressed by the existing related survey papers. Some of the studies have discussed different gaze estimation approaches like model-based methods, appearance-based methods, DL-based methods and convolutional neural network (CNN) based methods. For instance, Kar and Corcoran (2017) have explored the methods focusing on model-based approaches. They have presented their work under five categories: (1) 2D regression, (2) 3D model, (3) appearance-based, (4) cross Ratio-based and (5) shape based methods. Similarly, Cazzato et al. (2020) have surveyed gaze estimation techniques under two categories, (1) geometric-based and (2) appearance-based methods, by analyzing the advancements in computer vision together with DL. In other perspectives, Akinyelu and Blignaut (2020) and Cheng et al. (2021) have shown different DL-based gaze estimation techniques focusing on CNNs. Many of these studies have further reviewed the calibration techniques, performance evaluation metrics, devices and platforms and datasets in the gaze estimation literature. However, most of the studies have not discussed these approaches in a multi-user gaze estimation perspective considering factors like unconstrained environmental settings, gaze target variations and coordinate systems.

2.2. Scope of the survey

This paper provides a comprehensive survey of single and multi-user gaze estimation methods in DL from 2015 to 2021. The related studies are surveyed from four perspectives: (1) deep neural network model architecture, (2) datasets, (3) environment and (4) performance evaluation. From the deep neural network model architecture perspective, we review the DL-based approaches to include multi-task CNNs, temporal and spatial CNNs and capsule networks. Network backbones, inputs and outputs, optimization techniques are further discussed. From a dataset perspective, the metadata such as the number of images, subject variations, annotation formats and image quality are discussed. The environment perspective describes the coordinate systems used, head-pose variations, illumination variations and other application-specific environmental parameters. Finally, we review and compare the acquired performance aspects. Following are the highlights of this survey paper.

- Present an in-depth analysis of the DL-based gaze estimation approaches from 2015 to 2021 with a focus on multi-user gaze estimation techniques in unconstrained settings.

Table 1
Summary of related survey papers.

Consideration	Survey				
	Kar and Corcoran (2017)	Cheng et al. (2021)	Akinyelu and Blignaut (2020)	Cazzato et al. (2020)	Klaib et al. (2021)
Model-based methods	✓			✓	
Appearance-based methods	✓	✓	✓	✓	✓
DL-based methods	✓	✓	✓	✓	✓
Calibration techniques	✓	✓			
Datasets		✓	✓	✓	
Performance evaluation metrics	✓	✓	✓	✓	
Devices and platforms	✓	✓		✓	✓

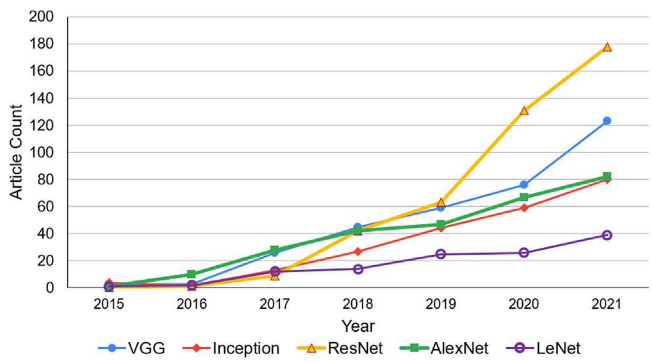


Fig. 2. Evolution of deep learning-based gaze estimation techniques.

- Provide a survey of existing state-of-the-art single-user and multi-user large-scale gaze datasets. Requirements for a standard multi-user gaze dataset, a summary of public and synthetic gaze datasets and issues related to public gaze datasets are discussed and analyzed.
- Explain the theory behind coordinate systems and the possible performance evaluation metrics that can be applied on eye gaze estimation.
- Suggest a guidance for selecting DL-based approaches in eye gaze estimation for researchers and developers. Discuss the open challenges and future opportunities in the field of DL-based multi-user gaze estimation.

2.3. Evolution of the techniques

Fig. 2 shows a quantitative view of the use of the techniques in the related literature during the years 2015–2021. We have considered the research papers indexed in Google Scholar for each of the techniques in the related studies. Our search strategy is based on "⟨technique name⟩" + "⟨research consideration⟩". Although the considered data can vary slightly due to the search query's associated noise, we assume the flaws are equally distributed over the search results for all the considered techniques. Thus, the audience can get a comparative view of the usage of the main techniques in this area.

As shown in Fig. 2, there is a similar growth in AlexNet, VGG (Visual Geometry Group) and Inception techniques from the year 2015 to 2021. Similarly, the residual neural network (ResNet) technique has shown a rapid increase in popularity. However, LeNet has decreased its usage, which may be due to the recent advancements in residual networks. Overall, it can be seen that the interest in gaze estimation research with deep CNNs is steadily increasing irrespective of the type of technique.

3. Overview of multi-user gaze estimation

3.1. Eye gaze

Human eye gaze is an active natural form of interaction that gathers information from a visual scene. It provides a wealth of information about human actions even though eye gaze is subtle and straightforward in comparison to gesture and speech. In eye gaze research, eye movements are studied thoroughly based on their type, functionality and characteristics. Analysis of eye movements are used to gather data about the user's intention, cognitive activities and attention (De Silva et al., 2021; Goldberg & Kotval, 1999; Velichkovsky et al., 2014). These eye movements are broadly classified as fixations, saccades, smooth pursuit, scanpath, gaze duration, blink and pupil size change (Kar & Corcoran, 2017).

Fixations are defined as times when eyes are stationary between movements and scan a scene. They have the least movement rate and are helpful for scanning detailed information, reading and attention. Saccades, on the other hand, have the highest movement rates and are helpful for visual search. These are simultaneous movements of both eyes that occur between fixations. Smooth pursuits are eye tracking movements used to follow moving targets of interest. Scanpath is a combination of alternating eye fixations and saccades prior to when the eyes reach a target position. The dimensionality of eye gaze can be classified as 2D gaze and 3D gaze. 2D eye gaze can be calculated using gaze direction from a single eye, while 3D eye gaze estimation requires both gaze direction and gaze depth from both eyes (Kwon et al., 2006).

3.2. Gaze estimation

Gaze estimation is an umbrella term used to assess human intent and interest through the measurement of human eye gaze (Tsukada et al., 2011). The history of human gaze estimation and eye-tracking dates back to the 18th century where researchers used invasive eye-tracking techniques to observe eye movements (Kar & Corcoran, 2017; Khan & Lee, 2019). However, with the evolution of digital signal processing and computer vision fields, more non-invasive gaze estimation approaches have been adopted by utilizing unique, physical characteristics of the eye (Chennamma & Yuan, 2013; Kar & Corcoran, 2017; Khan & Lee, 2019). The photometric and motion characteristics of the human eye have provided essential features required for this task (Akinyelu & Blignaut, 2020; Khan & Lee, 2019).

Gaze direction and point of gaze are two metrics used for gaze estimation. The visual axis, which deviates from the optical axis, determines the gaze direction (Kar & Corcoran, 2017), as shown in Fig. 3. Eye properties such as pupil and corneal reflection derived from eye regions, are used to determine it in the application level (Chennamma & Yuan, 2013). Subsequently, gaze point is defined as the intersection of the of gaze direction and the object's surface (Sun et al., 2016).

Before the emergence of computer vision-based methods, gaze estimation techniques relied on detecting patterns of eye movement including fixations, saccades and smooth pursuits (Young & Sheena, 1975). Methods based on computer vision can be classified into three groups: (1) 2D eye feature regression methods, (2) 3D eye model recovery method and (3) appearance-based methods (Cheng et al., 2021). These methods estimate the gaze using eye image and video data and the eye's geometric model characteristics. Specifically, the first two approaches detect geometric features of the eye, such as corneal reflection, pupil center and build an eye model to estimate gaze (Cheng et al., 2021; Kar & Corcoran, 2017). Coherently, these two

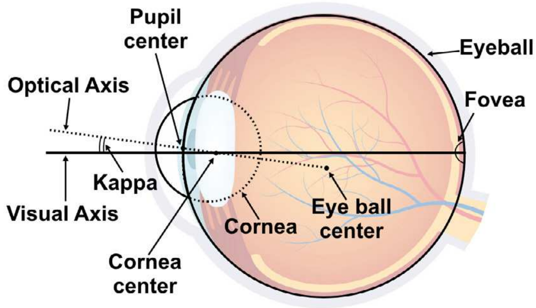


Fig. 3. Model of a human eye ball.

approaches are referred in the literature as model-based approaches. The third strategy considers the eye's photometric appearance to estimate gaze (Chennamma & Yuan, 2013). Model-based methods require the assistance of dedicated devices such as infrared cameras, while methods based on appearance do not require specialized instruments for gaze measurement.

Generally, there are two types of devices used in these methods: (1) remote eye tracker and (2) head-mounted eye tracker. The first type is typically kept at a distance of 60 cm from the user and the cameras. The second type is commonly installed on a frame of glass (Cheng et al., 2021). The user interfaces for gaze estimation are categorized into four groups: active, passive, single, or multi-modal (Kumar, Paepcke et al., 2007; Sibert & Jacob, 2000; Špakov & Miniotas, 2005). Active interfaces utilize the user's gaze to activate a function, while passive interfaces use gathered gaze data to determine a user's level of interest or attention (Kar & Corcoran, 2017).

Depending on the coordinate system used, gaze estimation techniques are divided into 2D and 3D gaze estimation. The existing majority of the studies have been proposed for 2D gaze estimation, while a few studies have focused on 3D gaze for accurate gaze estimation in real-world settings (Kodama et al., 2018; Sugano et al., 2016).

3.3. Multi-user gaze estimation

A growing interest in gaze estimation in unconstrained environments has been noticed alongside the rapid utilization of DL-based approaches in gaze estimation techniques in the last decade. The concept of multi-user gaze estimation has been studied and applied in various domains due to this adaptation (Bermejo et al., 2020; Kellnhofer et al., 2019; Kodama et al., 2018). In contrast to conventional single-user gaze estimation, the multi-user gaze estimation is mostly required in open environmental settings such as retail, public gatherings and public venues. Hence, it requires robust, low-overhead and high-speed gaze estimation approaches.

Existing multi-user gaze estimation studies can be split into two categories: time-sharing approaches and space-sharing approaches (Kodama et al., 2018; Sugano et al., 2016). The time-sharing method distributes the number of users over a time period. On the other hand, the space sharing approach process multiple users at the same time. In literature, time-shifting approaches have not captured much attention due to their unscalability and fewer robustness (Park et al., 2012; Park & Shi, 2015).

3.4. Applications of gaze estimation

Gaze estimation is becoming an increasingly effective technique in a variety of fields including computer vision, medical diagnosis, autonomous vehicles, psychology, human–computer interaction and sports training (De Silva et al., 2021, 2019; Kerr-Gaffney et al., 2019; Raptis et al., 2017; Sugano et al., 2016; Wang et al., 2015; Wang, Pi et al., 2018; Zhang, Sugano, Bulling, 2019). Through eye gaze estimation, valuable information of human behavior such as the object of concentration, internal cognitive state, user intent and attention analysis can be inferred (Kar & Corcoran, 2017). Eye tracking and gaze estimation were limited to psychological and cognitive studies and medical research in the early stages. But with technological breakthroughs in computing power, digital video processing, low-cost hardware and applications in gaze estimation have grown into new domains such as gaming, virtual reality and web advertising (Kar & Corcoran, 2017; Morimoto & Mimica, 2005). In human–computer interaction, gaze location can be used as an input modality to supplement other primary modalities such as a mouse, keyboard and touch. Eye movements reflect the cognition process of a human, as well as the medical and mental condition of that person, which can be used in multiple applications (Guojun & Saniie, 2016).

Kar and Corcoran (2017) have classified the types of devices in which single-user gaze estimation is used in five broad categories as: desktop-based systems, television and large display panels, head-mounted setups, automotive and hand-held devices (smartphones and tablets). In desktop-based systems, gaze estimation is used for computer communication such as mouse pointer control, gaze-based object selection, password entry and psychoanalysis (Ghani et al., 2013; Kasprowski & Harężlak, 2014; Kumar, Garfinkel et al., 2007; Sibert & Jacob, 2000; Zhai et al., 1999). In television and large display, a panel's gaze estimation can be applied for navigating menus, modifying display properties in TVs, switching channels and understanding user interests (Gwon et al., 2013; Lee et al., 2010). Gaze trackers installed on the head are commonly employed in portable platforms and have a variety of uses in domains such as augmented reality, virtual reality, sports training, computer gaming and psychological research (Lee et al., 2009, 2011; Piumsomboon et al., 2017; Sidorakis et al., 2015; Thies et al., 2018). In automotive systems, gaze estimation is vital for driver alertness detection, driver fatigue detection and cognitive state estimation (Ji et al., 2004; Sun et al., 2007; Zheng et al., 2015). In the context of hand-held devices, smartphone and tablet interaction has been immensely improved with the assistance of gaze estimation for tasks such as controlling the device, gaze-based user authentication and keyboard typing (Liu et al., 2015; Velichkovsky et al., 2014). Consequently, while single-user gaze estimation has expanded to a broad range of domains and applications, at the research level, multi-user gaze estimation is still a novel concept.

4. Gaze estimation approaches

Existing gaze estimation approaches are classified into two broad categories: appearance-based techniques and model-based techniques. Model-based gaze estimation techniques make use of a geometric model of the eye that includes a number of ocular components to include the cornea, optical and visual axes. While model-based gaze estimation methods are more precise, they typically require time-consuming personal calibration for each participant.

Appearance-based methods usually require user eye appearance images to directly learn a mapping function from eye appearance image to gaze estimation (Fischer et al., 2018; Huang et al., 2017; Kellnhofer et al., 2019; Xu et al., 2015). Appearance-based methods typically do not require camera calibration and geometry data, since the mapping is made directly on the image of the user's eye. Appearance-based methods can be divided into two categories: conventional appearance-based methods and appearance-based methods with DL. Their abstract concepts are depicted in Figs. 4 and 5, respectively.

4.1. Conventional appearance-based methods

Conventional appearance-based approaches treat whole images as features and deduce eye gaze directly from them. Conventional

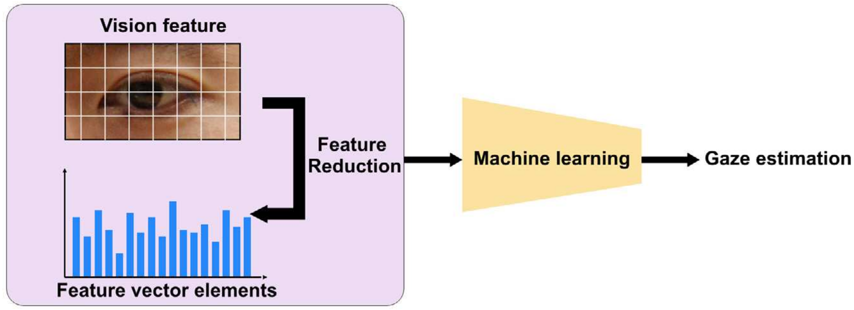


Fig. 4. Conventional appearance-based methods.

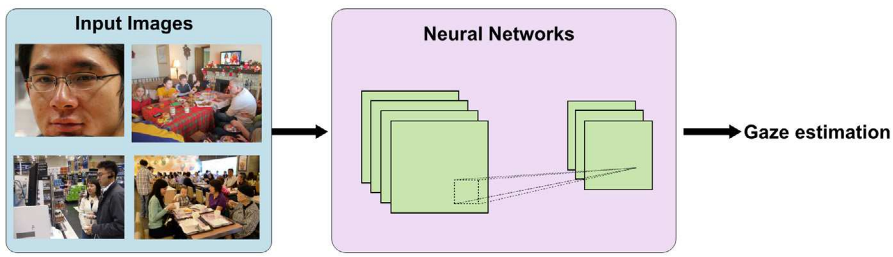
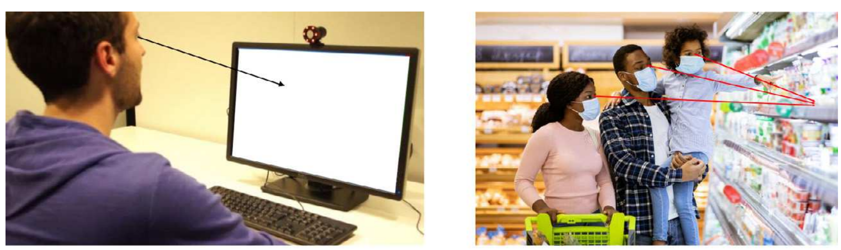


Fig. 5. Appearance-based methods with deep learning.



(a) Constrained environments.

(b) Unconstrained environments.

Fig. 6. Constrained environment and Unconstrained environment.

appearance-based methods have used mapping functions, such as Adaptive linear regression, K-Nearest-Neighbor, Random Forest regression, Artificial Neural Networks, Gaussian Processes, Support Vector Machines. Lu et al. (2014b) have proposed the adaptive linear regression (ALR) technique for mapping high-dimensional features of the ocular image to low-dimensional gaze positions, which significantly reduces the number of training samples for high accuracy estimation. K-Nearest Neighbors has become a standard method in the conventional appearance-based method for predicting gaze using the mean of neighbor samples' gaze angles. Wang, Zhao et al. (2018) have presented a gaze estimation framework that is a combination of neighbor selection and neighbor regression. It makes extensive use of information about the head's position, the pupil center and the appearance of the eyes. Kacete et al. (2016) have proposed an approach based on an ensemble of trees grouped in a single forest to learn the highly non-linear mapping function between the gaze information and the RGB eye image appearances, including depth cues. Yu et al. (2016) have proposed a method based on particle swarm optimization BP neural network. These methods endure many challenges. Most conventional appearance-based methods require a fixed head pose or a limited range of head movements as represented in Fig. 6(a). Furthermore, this method has difficulties in handling subject differences, especially in the unconstrained environment.

4.2. *Appearance-based methods with deep learning*

In computer vision, it has been demonstrated that DL techniques outperform earlier state-of-the-art machine learning techniques. Recently, research on gaze estimation has concentrated on methods based on DL. They have the ability to overcome challenges, such as significant head motion, subject differences and unconstrained environmental settings, as represented in Fig. 6(b). CNNs are the most widely used algorithm in this regard. An in-depth discussion on appearance-based methods with DL is presented in Section 8.

5. Coordinate systems

This section focuses on the main types of coordinate systems that have been addressed in the literature on gaze estimation. Mainly the coordinate systems can be categorized: (1) image coordinates, (2) subject and camera coordinates, and (3) screen coordinates, as shown in Fig. 7.

Image Coordinate System: An image coordinate system is a 2D coordinate system that enables a coordinate to specify a location in a 2D image (Chong et al., 2020; Fang et al., 2021; Recasens, 2016). There are two types of image coordinates namely, pixel coordinate and spatial coordinate. The image is treated as a grid composed of discrete elements in the pixel coordinates, ordered from top to bottom and left to right. Spatial coordinates provide for more precise location specification in an image than pixel coordinates do. Also, they describe

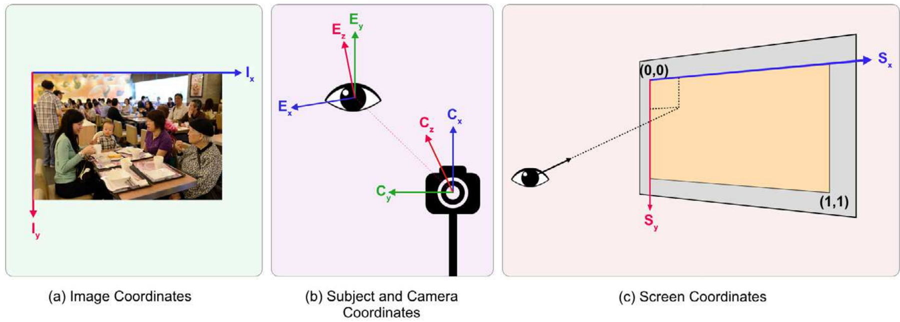


Fig. 7. Coordinate systems used for 2D and 3D gaze estimation; (a) Image coordinates, (b) Subject and camera coordinates, (c) Screen coordinates.

image positions in terms of partial pixels. Image coordinate systems are especially used in gaze-following systems (Recasens, 2016). In gaze-following dataset, a single image contains one or more persons. The pixel or spatial coordinate system is used to annotate the locations of the center of eyes, head, gaze point of each person in the image. In addition, some datasets contain object bounding boxes, segmentation masks and other boundaries. These are also annotated using the image coordination system (Tomas et al., 2021). Fig. 7(a) depicts the standard image coordinate system.

Subject and Camera Coordinate System: Subject coordinates represent the coordinate of the world from the perspective of the user's eyes (Bermejo et al., 2020; Kellnhofer et al., 2019). Camera coordinates share an origin with the subject coordinate system, but the coordinate axis orientation may be different as shown in Fig. 7(b). The coordinates of a camera are expressed in terms of points with the origin at the optical center of the camera. Subject coordinates and camera coordinate systems are 3D coordinate systems and are specifically used in gaze direction estimation systems to target positioning and, express the gaze orientation.

Screen Coordinate System: When using an eye tracker with a screen (gaze point estimation), all gaze estimations are mapped into a screen coordinate system (Kar & Corcoran, 2017; Sugano et al., 2016). This two-dimensional coordinate system corresponds to the physical coordinates of pixels on the computer screen based on the current screen resolution. The origin of the screen coordinate system is the screen's top left corner and the point (0, 0) signifies the screen's upper left corner, while (1, 1) denotes the screen's bottom right corner as shown in Fig. 7(c).

6. Datasets

6.1. Requirements for real-world benchmark datasets

The general requirements for a real-world benchmark dataset for multi-user gaze estimation can be listed as follows.

Environment: Different light conditions (Kellnhofer et al., 2019) that exist in unconstrained environments should be captured to improve the generality of the dataset. Bright light, night light, dawn, dusk and shadows are a few of the varying illumination conditions under which the images should be captured. The dataset should include a broad diversity in scenarios such as different head poses, body poses, in-frame gaze points, out-frame gaze point (Chong et al., 2020; Kellnhofer et al., 2019). Furthermore, these scenarios should be captured with different backgrounds, patterns and textures.

Target variation: In gaze estimation literature, a variety of targets such as gaze point, gaze direction and gazed object, have been studied. 2D and 3D gaze points are required by multiple applications to include desktop scenarios and public displays (Recasens, 2016; Sugano et al., 2016). 2D and 3D gaze directions are required to calculate the respective gaze points (Fang et al., 2021). Gazed object is a target associated with the novel concept of gaze object prediction, which requires annotating gazed object bounding boxes (Tomas et al., 2021). A multi-user perspective of these targets is a necessary requirement in a multi-user benchmark dataset.

Subject variation: Substantial subject variations should be captured by considering aspects such as collecting images with a sufficient number of subjects, male and female subjects, subjects from different regions of the world representing different skin colors, face and eye shapes (Kellnhofer et al., 2019; Tomas et al., 2021).

Viewpoint: Different viewpoints have been studied in the gaze literature. 2D image coordinates, 2D screen coordinates, 3D subject coordinates and 3D camera coordinates are the used viewpoints of coordinate systems. Head pose is captured in different viewpoints such as constraint head poses, unconstrained head poses, broad head yaw and pitch variations (Zhang et al., 2020). Similarly, ocular regions are captured in multiple viewpoints to include without occlusion, partial occlusion and total occlusion (Kellnhofer et al., 2019). Either head-mounted displays or remote cameras such as webcams, kinect and surveillance cameras are used to collect images from the different viewpoints.

Challenging conditions: Multi-user gaze estimation in unconstrained settings introduce numerous challenging conditions. Datasets should capture these challenging conditions such as eye, face and body occlusion (Kellnhofer et al., 2019), and other subject distortions such as scenarios where subjects are wearing spectacles (Tomas et al., 2021). Furthermore, datasets should capture scene images with varying camera-to-subject distances (Mishra & Lin, 2020) and different illumination conditions (Zhang et al., 2015).

6.2. Public gaze datasets

Recent research on eye gaze estimation have used different types of datasets with the growth of DL techniques. Most of the publicly available datasets have used head mounted devices, surveillance camera and other desktop and mobile eye trackers to capture images for eye tracking, head pose detection and pupil tracking. Table 2 provides a summary of common gaze estimation datasets. Most of the existing datasets support the single-user eye gaze estimation process and there is a lack of datasets related to multi-user eye gaze images. Some of the datasets are captured in controlled environments, whereas the others are acquired in uncontrolled (wild) settings. Fig. 8 includes sample images from the publicly available datasets, namely (a) MPIIGaze (b) Columbia Gaze, (c) Gaze360, (d) GazeFollow, (e) Gaze on objects. Some of the gaze estimation datasets that are widely used in related studies are listed as follows.

MPIIGaze: Zhang et al. (2015) have presented the MPIIGaze dataset. It is a novel in-the-wild gaze dataset and one of the most widely used datasets for estimating gaze using appearance-based methods. This dataset was collected utilizing laptops over a three-month period that demonstrate significant variations in eye appearance. Even though the original dataset only contains binocular eye images, the improved version of the dataset includes face images (Zhang et al., 2017) and

(a) MPIIGaze (b) Columbia Gaze (c) Gaze360 (d) GazeFollow (e) Gaze on Object

Fig. 8. Sample images from publicly available datasets.

Table 2
Summary of gaze estimation datasets.

Dataset	Year	Subjects	Total	Annotations	Type	Environment
MPIIGaze (Zhang et al., 2015)	2015	15	213,659	2D and 3D gaze directions	Single	Wild
Columbia Gaze (Smith et al., 2013)	2013	56	5880	3D gaze direction	Single	Controlled
Gaze360 (Kellnhofer et al., 2019)	2019	238	172,000	3D gaze direction	Single	Wild
GazeFollow (Recasens, 2016)	2015	130,339	122, 143	2D gaze direction, target	Multi	Wild
GOOReal (Tomas et al., 2021)	2021	100	9552	2D gaze direction, target	Single	Wild
UTMultiview (Sugano et al., 2014)	2014	50	1,100,000	2D and 3D gaze direction	Single	Controlled
EyeDiap (Mora et al., 2014)	2014	16	94 (videos)	2D and 3D gaze direction	Single	Controlled
GazeCapture (Lu et al., 2014a)	2016	1474	2,400,000	2D gaze direction	Single	Wild
RT-Gene (Fischer et al., 2018)	2018	15	123,000	2D gaze direction	Single	Controlled
ETH-XGaze (Zhang et al., 2020)	2020	110	1,100,000	2D and 3D gaze direction	Single	Controlled
NVGaze (Kim et al., 2019)	2020	30	4,500,000	2D gaze direction	Single	Controlled
TabletGaze (Huang et al., 2017)	2017	51	816 (videos)	2D gaze direction	Single	Controlled

manually annotated landmarks (Zhang, Sugano, Fritz, Bulling, 2019) as well. It contains 213,659 images that were gathered from fifteen participants. It also includes both 2D and 3D annotations. Additionally, MPIIGaze provides a standard evaluation dataset that includes 15 participants and 3000 images of each participant's left and right eyes. Most current gaze datasets restrict the head pose range. However, MPIIGaze includes an extensive head-pose range and a gaze angle range (Mora et al., 2014; Sugano et al., 2014).

Columbia Gaze: Smith et al. (2013) have developed a large publicly available dataset for appearance-based gaze estimation. The collection contains 5880 high-quality images of 56 subjects (32 males and 24 females), with a resolution of 5184 × 3456 pixels for each image. Participants ranged in age from 18 to 36 years and 21 of them wore glasses. Twenty-one participants were Asian, nineteen were Caucasian, eight were South Asian, seven were African and four were Hispanic or Latina, indicating a greater range of eye appearances. For each subject, they collect images for each of the seven horizontal gaze directions, five horizontal head poses and three vertical gaze directions. In the data collection setting, participants were seated in a fixed place in front of a black background. They were asked to focus on a dot shown on a wall while their eye gaze was recorded. The 3 × 7 grid of dots was placed in 10 increments vertically and ten increments horizontally.

Gaze360: Most of the available datasets are not suited for developing a model capable of reliably assessing 3D gaze in the wild. Kellnhofer et al. (2019) have proposed Gaze360, a large-scale gaze estimate dataset for unconstrained 3D gaze estimation. Gaze360 is unique for its combination of numerous gaze poses, head poses, 3D gaze annotations, a variety of indoor and outdoor locations and a diversity of subjects like age, sex, ethnicity. The dataset contains 172 000 images of 238 participants and each image has a resolution of 3382 × 4096 pixels. Dataset has collected in 5 indoor (with 53 participants) and 2 outdoor (with 185 participants) locations. This dataset consists of 58% and 42% of female and male participants, respectively. The dataset enables gaze estimate up to the limit of eye visibility, which in certain circumstances corresponds to gaze yaws of around ±140°. The Gaze360 dataset collecting arrangement was centered on a Ladybug5 360° panoramic camera in the scene's center, a moving target board marked with an AprilTag (Wang & Olson, 2016), and a cross on which participants were asked to gaze constantly. Participants were located at a distance of approximately 1–3 m from a camera.

GazeFollow: This is a large-scale dataset labeled with the 2D image location of where participants in the images are looking at (Recasens, 2016). The dataset contains 122,143 images that utilize people as a source of imagery. These images contain individuals engaged in a variety of ordinary tasks, and each image contains a single person or multiple people. Since the images do not consist of ground truth gaze, they have labeled images using Amazon's Mechanical Turk and their online tool. GazeFollow dataset is designed to capture different fixation scenarios. Several images depict multiple people paying attention to one another, while others depict individuals looking at each other (Everingham et al., 2009; Lin et al., 2014; Russakovsky et al., 2015; xiong Xiao et al., 2010; Yao et al., 2011; Zhou et al., 2014).

GOOReal: Most of the gaze estimation datasets only have the pixel being looked at instead of the boundaries of a particular object of interest. This lack of object annotations presents an opportunity for advanced gaze estimation research. Tomas et al. (2021) have introduced the task of gaze object prediction along with the Gaze On Object (GOO) dataset for the retail environment in order to address this issue. The GOO-Real dataset consists of 9552 images of 100 participants (32 female and 68 male), and each image is composed of shelves packed with 24 different classes of product items. Each participant was instructed to enter the grocery environment, and they would then fixate on each item for a few seconds. Two images were collected for each item stared at, and annotators were attached a ground truth label (grocery item identifier) for each image. All objects were annotated with their class, bounding box (product items, head area) and segmentation mask.

6.3. Issues in public gaze datasets

Many public datasets have several issues and challenges when using in real-world applications. The majority of datasets are suited for physically constrained applications such as desktop and mobile phone gaze estimation. Typically, these datasets are collected using a static recording setup, which allows higher accuracy. However, they may lack the diversity in illumination and motion blur. Therefore, these datasets are not valid for general applications. On the other hand, these datasets contain relatively small head pose angles and gaze variation and are restricted to frontal views. Most of the existing gaze datasets are not annotated for multi-user gaze estimation. Therefore, additional effort is required to annotate the images using these datasets in the multi-user gaze estimation process.

6.4. Generated synthetic datasets

Generally, publicly available datasets are primarily used to train and evaluate gaze estimation models. Collecting accurate gaze estimation data and creating a dedicated gaze estimation dataset require time, effort and cost. Additionally, public datasets are not always suitable and sufficient for a particular task. Tomas et al. (2021) have presented a synthetic dataset called GOO-Synth with 192,000 images. They have used Unreal Engine to create a realistic-looking replica of the scene that is used in the real dataset. Moreover, Bermejo et al. (2020) have created a synthetic dataset with 50 subjects to improve the back head detection task in their models. Different approaches based on techniques like Mask-RCNN (Shashirangana et al., 2021) and StyleGAN (Karras et al., 2020) have been used in the literature to generate synthetic datasets.

7. Performance evaluation metrics and standards

Different performance evaluation metrics and standards have been used in the literature to assess the 2D and 3D gaze estimation techniques. The type of evaluation metrics is depend on the nature of gaze estimation, which can be further classified into two broad categories, namely 2D gaze estimation and 3D gaze estimation. Furthermore, these metrics differ depending on the gaze estimation task performed including gaze point estimation, gaze direction estimation and gaze object prediction.

Area Under Curve (AUC): Area Under the ROC curve is one of the primary metrics used to evaluate the accuracy of 2D gaze point estimation (Chong et al., 2020; Fang et al., 2021; Recasens, 2016; Tomas et al., 2021). Judd et al. (2009) have presented Area Under Curve criteria from a ROC curve to predict the performance of human saliency maps in gaze fixations. The saliency map is treated as a binary classifier for each image pixel in this metric. The classification threshold is determined in such a way that a specified percentage of picture pixels are categorized as fixated, while the remainder are classed as unfixed. AUC of the value 1 indicates a model that behaves perfectly, while average performance is 0.5.

L_2 Distance: L_2 distance is another primary metric used to evaluate the accuracy of 2D gaze point estimation (Chong et al., 2020; Fang et al., 2021; Recasens, 2016; Tomas et al., 2021). The mean Euclidean distance between the gaze predictions and their respective ground-truth gaze annotations is defined as L_2 distance in 2D gaze estimation literature (Fang et al., 2021; Recasens, 2016). L_2 distance can be obtained from Eq. (1), where gt_x_i and gt_y_i refers to the ground truth gaze annotations along x-axis and y-axis, respectively. The notations x_i and y_i denotes the gaze predictions in 2D image coordinates.

$$L_2 distance = \frac{1}{n}\sum_{i=1}^{n}\sqrt{(gt\_x_i - x_i)^2 + (gt\_y_i - y_i)^2} \tag{1}$$

Angular Error: Some studies have used angular error to determine the accuracy of 2D and 3D gaze direction estimation techniques (Fang et al., 2021; Kellnhofer et al., 2019; Recasens, 2016; Tomas et al., 2021). The angular difference between the predicted and true gaze direction vectors is defined as the angular error. The predicted gaze direction vector is produced by connecting the head point to the predicted gaze point. This metric is calculated in both 2D and 3D vector spaces.

Average Precision: The average precision metric is used in scenarios, where out of frame gaze binary classification has been considered (Chong et al., 2020; Fang et al., 2021). The area under the precision–recall curve is defined as the average precision as stated in Eq. (2).

$$AP = \int_0^1 p(r)\,dr \tag{2}$$

Classification Accuracy: The classification accuracy metric is reported in scenarios, where gaze estimation has been represented as a classification problem (Akinyelu & Blignaut, 2020; Mahanama et al., 2020). It is the ratio of correct predictions to total predictions. The accuracy of binary classification is expressed in terms of positives and negatives, as given in Eq. (3). The notations TP, TN, FP, FN denote true positive, true negative, false positive, false negative, respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Mean Squared Error (MSE): The mean squared error is another metric used to determine the accuracy of 2D and 3D gaze direction estimation techniques (Fang et al., 2021; Kellnhofer et al., 2019; Recasens, 2016; Tomas et al., 2021). MSE is defined as the average squared difference between the ground truth and the prediction (Handelman et al., 2019). In gaze estimation literature, MSE can be obtained from Eq. (4), where y_i and gt_y_i refers to the predicted gaze and ground truth gaze, respectively.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - gt\_y_i)^2 \tag{4}$$

8. Related research models

Deep learning-based techniques have been widely used in the field of gaze estimation due to their ability to map high-level gaze features directly from images and produce results in real-world settings (Kellnhofer et al., 2019; Wang & Shen, 2017; Zhang et al., 2015). CNNs are at the backbone of most of these techniques incorporating other DL architectures and techniques such as Capsule Networks, Recurrent-Neural Networks, Residual Neural Networks, Multi-Task CNNs and Transfer Learning (Chong et al., 2018; Fang et al., 2021; Kellnhofer et al., 2019; Lian et al., 2018; Mahanama et al., 2020).

This section explores DL-based gaze estimation methods with a focus on multi-user gaze estimation. These methods are introduced in two main perspectives, DL-based methods for single-user gaze estimation and DL-based methods for multi-user gaze estimation. The surveyed studies are further categorized according to the coordinate system and environmental settings. The studies on single-user and multi-user eye gaze estimation approaches are summarized in Tables 3 and 4, respectively. Moreover, we discuss the recent research studies that have used 2D and 3D DL architectures as demonstrated in Figs. 9–14.

8.1. Deep learning based methods for single-user gaze estimation

8.1.1. 2D deep learning methods in constrained-environments: Single-user

The extraction of the ocular regions is a challenging task in naturalistic environments due to occlusion (Saad et al., 2020). Also, extracting head-pose information from ocular regions has not been explored in detail. Among the related studies, Mahanama et al. (2020) have proposed an appearance-based 2D gaze estimation model named Gaze-Net, using capsule networks for decoding, representing and estimating gaze information from ocular region images. Capsule networks have been used in contrast to CNNs with pooling due to their capability to learn equivariant representations of objects. They have followed a two-step approach combining the classification of gaze direction into six classes and reconstructing the original ocular image in order to construct and train the deep neural network. In their work, it has been hypothesized that a single eye image consisting of sufficient information can reliably estimate the gaze. They have used two publicly available datasets, MPIIGaze (Zhang, Sugano, Fritz, Bulling, 2019) and the Columbia Gaze (Smith et al., 2013), to train and test the model. Further, they have incoperated PoseNet (Oved et al., 2018) to obtain x,y coordinates of ocular regions in the images. An accuracy of 62% and a mean absolute error of 2.84 were recorded for the gaze estimation task.

8.1.2. 2D deep learning methods in the wild: Single-user

Existing gaze-related dataset annotations only contain the pixel of the gaze, instead of the area of a specific object of interest. In a related study, Tomas et al. (2021) have addressed this issue by introducing a challenging task called gaze object prediction. Moreover, they have presented the Gaze On Objects dataset based on the retail environment for training and evaluation. The dataset consists of a smaller set of real images (GOO-Real) and a larger synthetic set of images (GOO-Synth). GOO-Real consists of 100 human and 9552 grocery item images. Goo-Synth consists of 192,000 images created with Unreal Engine. All Objects in the frame are annotated with their class, bounding box and segmentation mask. The GOO dataset can be used in gaze following, gaze object prediction and domain adaptation. Several baselines are benchmarked on the GOO dataset (Chong et al., 2020; Lian et al., 2018). They have been evaluated using standard metrics such as the area under the ROC curve (AUC), L_2 distance and the angular error. Baseline evaluation results consistently show the models training on the GOO-Synth dataset, prior to the training on a GOO-Real dataset to achieve higher performance on all metrics.

8.1.3. 3D Deep learning methods in constrained-environments: Single-user

Most appearance-based eye gaze estimation methods have only used encoded features from eye images. In addition, gaze estimation tasks are limited to 2D screen mapping. Zhang et al. (2017) have proposed a 2D and 3D appearance-based gaze estimation method that uses face images as the input. The proposed model architecture is based on CNNs. They have introduced additional layers that learn spatial weights to activate the last convolutional layer in order to efficiently use the face information. The spatial weights mechanism forces the network to understand and learn the importance of various face regions for gaze estimation. This mechanism has implemented using the concept of the 1 × 1 convolutional layer and the rectified linear unit layer. The obtained results have outperformed the state-of-the-art for both 2D and 3D gaze estimation, reaching an accuracy of 6° and 4.8°, improvements of up to 27.7% and 14.3% on EYEDIAP (Mora et al., 2014) and MPIIGaze (Zhang, Sugano, Fritz, Bulling, 2019) for 3D gaze estimation, respectively.

Another approach for 3D single-user gaze estimation has been proposed by Lian et al. (2019) using multi-task CNNs. In this work, 3D gaze estimation task has introduced as RGBD gaze estimation by incorporating the depth channel. A generative adversarial network (GAN) has been used for depth image generation to reduce noise and black holes. The proposed network architecture combines an eyeball feature extractor, a head pose extractor and a 3D eye position encoder to predict the gaze point by taking two single eye images and an RGBD (Red, Green, Blue, Depth) head image as inputs.

8.1.4. 3D Deep learning methods in the wild: Single-user

Many related studies have explored gaze target detection without incorporating the depth estimation of gaze prediction (Chong et al., 2020; Recasens, 2016). As a solution, Fang et al. (2021) have proposed a method for gaze target detection in the unconstrained environments based on deep CNNs. As shown in Fig. 9, the authors have introduced a novel architecture for the task by incorporating 3D gaze estimation and a dual attention module (DAM) consisting of a field of view mask and a gaze-depth channel. The model used a single image in the wild as the input and outputs a 2D saliency map.

In another study by Ranftl et al. (2020), a priori depth map has been employed to generate the depth map of the image. A coarse-to-fine strategy has been developed for 3D gaze estimation, which can cope with completely occluded eyes and faces. The task of gaze target prediction has presented as a combination of two sub-tasks; (1) identifying whether the gaze target is inside or out of the image, (2) locating the target if inside. The output from the DAM and the scene image has passed to a ResNet-50 backbone and then to a binary classification head and a heatmap regression head to obtain the two results. They have used Gaze360 (Kellnhofer et al., 2019), GazeFollow (Recasens, 2016) datasets and VideoAttentionTarget (Chong et al., 2020) dataset to train, test and fine-tune the model, respectively. The proposed method has produced on par results as a single human, achieving 14.9° angular error, 0.922 AUC and 0.896 average precision. This work has shown promising results for single-user gaze target detection using 2D images in the wild in spite of head-eye inconsistency and occlusion.

Robustly estimating gaze in the wild with varying-camera person distances is another challenge for CNN backbones. Mishra and Lin (2020) have proposed a novel solution for the task by aggregating multiple zoom scales of the same input image using the center-cropping technique. Moreover, they have introduced a sine–cosine transform to avoid the yaw angle discontinuity in 360° backward gaze estimation, which penalizes DL models with substantial losses. The aggregation of center cropped input images with multiple sizes has been carried out by spatial-max pooling and has fed into a ResNet-18 (He et al., 2016) backbone and other backbone variants. The pinball loss function inspired by Gaze360 (Kellnhofer et al., 2019) has been used to output the uncertainty of the predictions further. A sequential model using bidirectional LSTM has been proposed, and a sequence of multi-crops has achieved better performance on the Gaze360 dataset. The best mean angular errors achieved for all 360°, front 180°, and back in Gaze360 dataset are 12.4, 10.7 and 18.9, respectively, using the sequential model with Hard-net (Chao et al., 2019) as the backbone. Validation of the model on the RT-GENE dataset has achieved a state-of-the-art mean angular error of 6.7 using the static model.

Another study on 3D single-user gaze estimation in the wild can be summarized as follows. Chong et al. (2018, 2020) have published two consecutive studies using state-of-the-art deep CNNs to predict heatmaps of gazed targets by a single-user. The predecessor has achieved near single-human performance on the GazeFollow dataset for single image gaze target prediction. A summary of single-user gaze estimation approaches is given in Table 3.

8.2. Deep learning based methods for multi-user gaze estimation

Gaze estimation of multiple people is a relatively new research area that has been emerging with the adaptation of DL-based methods for gaze estimation (Sugano et al., 2016). A summary of multi-user gaze estimation approaches is presented in Table 4.

Existing methods of multi-user gaze estimation can be placed into two categories: (1) techniques that analyze the gazes of multiple people sharing time and space, and (2) techniques that explore the gazes of multiple people sharing only space (Kodama et al., 2018). The first type requires several people to be wearing head-mounted cameras to estimate each of their gazes, thus hindering its practicality in real-world scenarios due to the requirement of a head-mounted camera for each person (Kodama et al., 2018; Park et al., 2012; Park & Shi, 2015). The approaches for the second type are discussed under this section, comparing their performance, reliability and challenges. These approaches are presented under two sections based upon the dimensionality of gaze estimation and the nature of constraints in the environment.

8.2.1. 2D deep learning-based methods in the wild: Multi-user

Multi-user gaze estimation in a 2D image coordinates system is a timely approach due to the potential of DL techniques in determining gaze direction in unconstrained settings. Recasens (2016) has proposed a deep neural network-based approach using CNNs for the novel task gaze-following in the wild. A benchmark dataset GazeFollow has been further presented. Gaze-following is the task of following a person's gaze to predict the object being looked at, which had not received prominent attention until this point. As shown in Fig. 10, the head pose and the gaze orientation are extracted from the scene image.

The location of different objects being looked at by different people in the scene is predicted in 2D image coordinates. Unlike previous work, this approach have used only a single third-person view of the

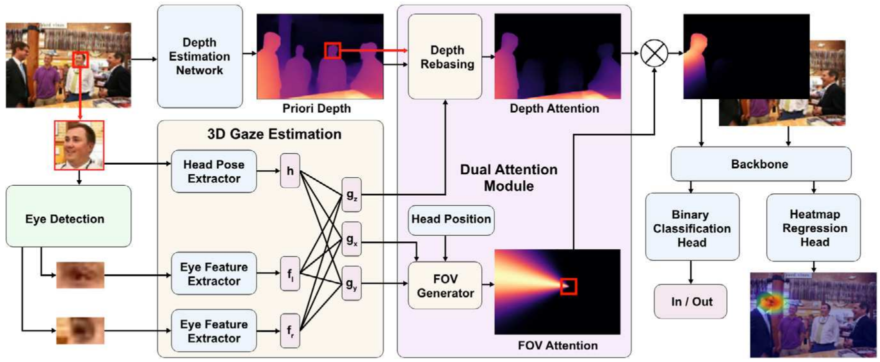


Fig. 9. Representation of the model architecture presented by Fang et al. (2021).

Table 3
Summary of single-user gaze estimation approaches.

Ref.	Architecture	Backbone	Dataset	Performance	Coordinate system	Environment
Zhang et al. (2017)	CNN-Spatial	AlexNet	Own dataset	Ang - 4.8°	2D, 3D	Controlled
Chong et al. (2018)	Multi-task CNN	ResNet-50	EYEDIAP, GazeFollow, SynHead	AUC - 0.896 L_2 - 0.187 Ang. - 6.4°	3D	Wild
Lian et al. (2019)	Multi-task CNN	ResNet-34, Own	EYEDIAP, Own	AUC - 0.906 L_2 - 0.145 MAng - 8.8°	3D	Controlled
Chong et al. (2020)	CNN-LSTM	ResNet-50	GazeFollow, VideoAttentionTarget, VideoCoAtt	AUC - 0.924 L_2 - 0.096 Out of Frame AP - 0.925	3D	Wild
Mahanama et al. (2020)	Capsules, CNN	Own architecture	MPIIGaze, Columbia Gaze	Accuracy - 62%	2D	Controlled
Mishra and Lin (2020)	CNN-LSTM	ResNet-18, Hardnet	Gaze360, RT-GENE	MAng - 12.4°	3D	Wild
Tomas et al. (2021)	CNN-static	ResNet-50	GOO, GazeFollow	AUC - 0.889 L_2 - 0.150 Ang. - 29.1°	2D	Wild
Fang et al. (2021)	CNN-static	ResNet variants	Gaze360, GazeFollow, VideoAttentionTarget	AUC - 0.922 L_2 - 0.124 Ang. - 14.9°	3D	Wild

Table 4
Summary of multi-user gaze estimation approaches.

Ref.	Architecture	Backbone	Dataset	Performance	Coordinate system	Environment
Recasens (2016)	CNN with shifted grids	AlexNet	GazeFollow	AUC - 0.878 L_2 - 0.190 Ang. - 24°	2D	Wild
Sugano et al. (2016)	CNN spatio-temporal	AlexNet	Own, Coutrot, Hollywood2	–	3D	Wild
Kodama et al. (2018)	CNN	LeNet-5	Own	MAE - 10.39 m	3D	Wild
Kellnhofer et al. (2019)	CNN-LSTM	ResNet-50	Gaze360	MAng - 13.5°	3D	Wild
Lian et al. (2018)	CNN	ResNet-50	GazeFollow, DLGaze	AUC - 0.906 L_2 - 0.081 MAng. - 8.8°	2D	Wild
Bermejo et al. (2020)	CNN	ResNet-18	UcoHead, Own	MAE - 19° FPS - 0.52	3D	Wild

scene, including the person and the object being gazed at to infer gaze. They have introduced a large-scale dataset, GazeFollow, annotated with the gaze object annotations by accumulating 122,143 image data consisting of 130,339 people from several significant datasets for model training and evaluation tasks. An in-depth survey of the dataset is given in Section 6.2. The dataset is designed to capture various fixation scenarios in which the people count varied from a single person to a crowd of people. They have described the gaze-following of humans using a gaze pathway that detects the gaze direction and a saliency pathway that identifies the salient objects. Also, a CNN architecture based on AlexNet (Krizhevsky et al., 2017) is used as the backbone.

The model is designed to support multi-modal predictions for reliable predictions of gaze objects in ambiguous scenarios. The problem is formulated as a classification task by quantizing the fixation location into an N×N grid where the size of N is selected using a shifted grids approach. The experimental results of the study show that the model achieves an AUC of 0.878 and L_2 distance of 0.190 for the gaze fixation prediction task, where the measured single-human level performance for the task is 0.924 ACU and 0.096 L_2 distance. Even though the results show that the model is robust to inaccurate head detection, the lack of 3D understanding has generated incorrect predictions in their work.

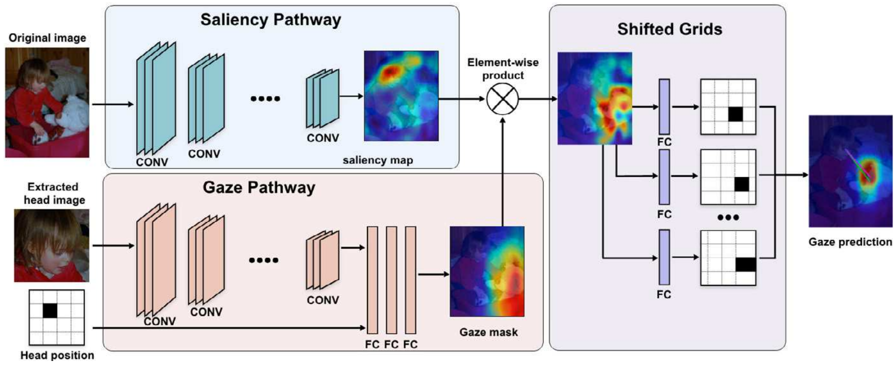


Fig. 10. Representation of the GazeFollow network proposed by Recasens (2016).

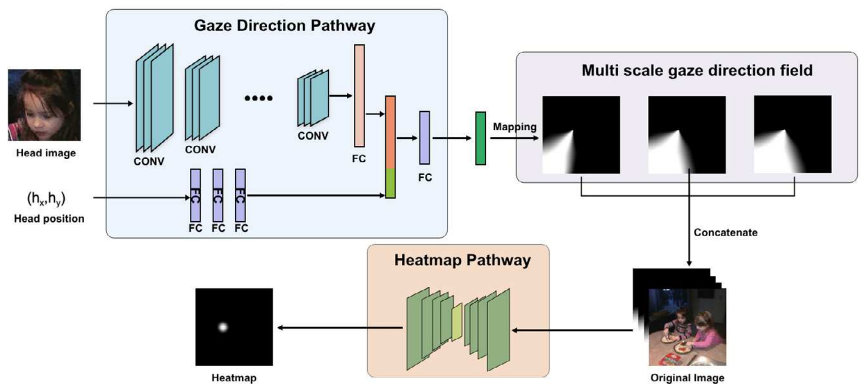


Fig. 11. Representation of the model architecture presented by Lian et al. (2018).

A similar approach to GazeFollow (Recasens, 2016) has been proposed by Lian et al. (2018) for multi-user gaze point prediction of the target person in a scene. As demonstrated in Fig. 11, they have proposed a two-stage solution consisting of a gaze direction pathway and a heatmap pathway by mimicking the gaze-following the behavior of a human. In the first stage, gaze direction was estimated by head images and its position to generate multi-scale gaze direction fields. In the second stage, multi-direction gaze fields have concatenated with the original image to regress the heatmap. Unlike in GazeFollow (Recasens, 2016), two pathways have been associated with each other to mimic gaze following the behavior of a human. Furthermore, more robust gaze heatmap prediction has been proposed to replace gaze point estimation. ResNet-50 based DCNN has been used along with a three fully connected layer network for gaze direction prediction. Adam optimizer has been used to optimize the model training. The heatmap pathway has used a feature pyramid network (Lin et al., 2017) with a Sigmoid activation function. The GazeFollow dataset and their own video dataset named DLGaze, have been used for model training, validation and evaluation. The experimental study has shown a mean angular error of 8.8°, which has surpassed the 11.6° result in Recasens (2016). The authors have stated that the two-stage architecture inspired by human behavior is the reason for the improved performance.

8.2.2. 3D Deep learning-based methods in the wild: Multi-user

Application independent 3D gaze estimation in the wild serves as a good entry point for many applications in the domain. Kellnhofer et al. (2019) have proposed a robust appearance-based method for 3D gaze estimation in unconstrained images of large diversity using bidirectional Long Short-Term Memory capsules (LSTM) (Graves et al., 2005). As shown in Fig. 12, the authors have presented Gaze360, a large gaze estimation dataset containing 172k images with 238 subjects. It consists of a wide range of gaze and head pose angles, significant variation in natural illumination, diverse and arbitrary environments for the task. The gap between leveraging the full potential of DCNN and the lack of sufficient annotated diverse data for the task is bridged through the approach.

The proposed model emphasizes the temporal nature and the continuity of gaze as a signal by aggregating seven image frames to predict the gaze of the central frame using LSTM capsules. ImageNet-pretrained ResNet-18 (He et al., 2016) architecture is used as the CNN backbone to predict the gaze in real-world 3D spherical coordinates. An uncertainty value for a gaze prediction is introduced and measured using quantile regression (Koenker, 2005) by the pinball loss function. The uncertainty prediction, as well as not relying on eye or face detectors, allowed the model to robustly estimate a gaze direction even in fully occluded eyes. Mean angular errors (MAE) are calculated for various static and temporal models to validate the gaze estimation and calculated the correlation between the actual error and the predicted uncertainty using Spearman's rank correlation. MAE values 13.5, 11.4 and 11.1 were obtained for all 360°, front 180° and front-facing scenarios, respectively. Further, an uncertainty correlation of 0.45 was obtained using the proposed method.

Kodama et al. (2018) have proposed a method for localizing the common gaze target focused on by a crowd of people in a tennis stadium using low-resolution images by aggregating the individually estimated 3D gaze of each person, as shown in Fig. 13. This study has further analyzed the relationship between the number of people involved in the aggregation and the localization accuracy of the common gaze target estimation. They have constructed a dataset of 12,792 images, which consists of 96 participants in a tennis stadium using two cameras. Each image consists of 48 people. The dataset further contains 454,739 face images annotated with 3D real-world coordinates with yaw angle and pitch angle ranging from −74.02 to 74.02 and −20.09 to −3.01, respectively.

The authors have used a multi-task cascaded CNN-based face detector to detect the faces, which were then used to train the Le-Net-5 (LeCun et al., 1998) based gaze angle estimator. In the experimental

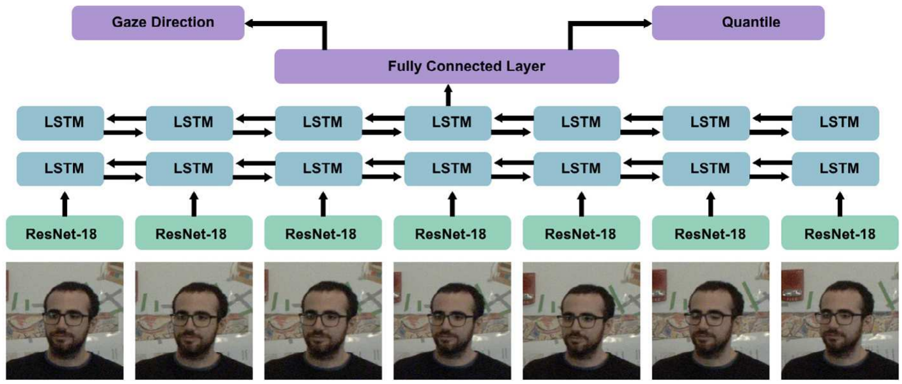


Fig. 12. Representation of the Gaze360 model architecture by Kellnhofer et al. (2019).

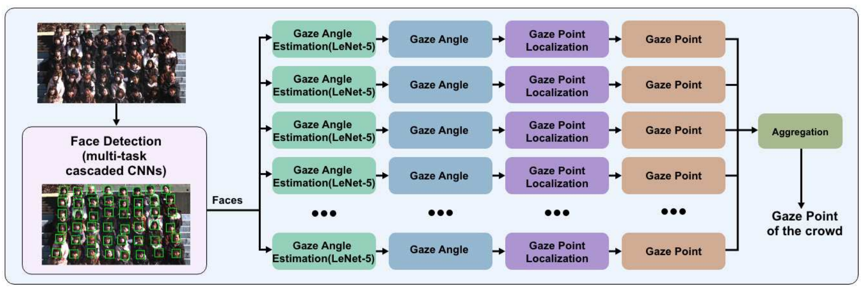


Fig. 13. Representation of the model architecture by Kodama et al. (2018).

study, the method's performance has been studied with respect to the number of people involved in the aggregation, considering the single-person case as the baseline. They achieved a 13.99 m MAE of estimated gaze point for the baseline and reduced it to 10.39 m by aggregating 24 people. Their comprehensive experimental study indicates promise for aggregating individual gaze estimations for more accurate common gaze target prediction in the wild. However, a more robust aggregation method still needs to be developed where individual gaze estimations contain significant biases.

An application-specific method for 3D Multi-user gaze estimation in the wild has been explored by Bermejo et al. (2020). They have proposed an exciting approach EyeShopper, to analyze customer behavior in retail stores using gaze estimation from back-head images of shoppers as shown in Fig. 14. They have further generated a synthetic back-head image dataset of 144,000 images consisting of 50 subjects and ±90° head yaw and pitch variations due to the unavailability of public back-head datasets in the wild. In this work, they have assumed that the customer's gaze can be predicted based on the customer's head position when the subject's face is not visible. With this assumption, they have proposed an accurate DCNN based architecture for gaze estimation using head-pose from back-head images and a novel loss function. A fine-tuned version of You Only Look Once (YOLO) v3 model is used as the back-head detector and a hybrid coarse-fine approach using a static ResNet-18 backbone as the head pose estimator has been used. The coarse-fine approach combines a four-class head-pose classification layer and a fine regression layer implemented using fully connected layers. The proposed model has been trained with 122,092 images and validated on 26,184 images by combining images from the UcoHead (Muñoz-Salinas et al., 2012) dataset, a manually labeled dataset and the synthetic dataset. For backhead gaze estimation, a mean absolute error of 19°, which is 10% lower than Hopenet (Ruiz et al., 2018) has been achieved along with an average of 0.52 frames per second (FPS).

9. Discussion

9.1. Criteria for selecting a research approach

We provide our suggestions for the selection of the DL-based gaze estimation approach in a practical point of view, as shown in Fig. 15. The criteria are based on the performance metrics and implementation issues in gaze estimation literature. By considering the majority vote of surveyed papers, we assumed that DL-based gaze estimation approaches are employed in unconstrained settings, while model-based methods are used in constrained situations. The effectiveness of the aforementioned techniques depends on the environment settings, head angle, distance variance, subject count, available computational resources and the other constraints. Additionally, our selection criteria are confined to methodologies based on DL and do not consider the availability of datasets for decision making. These guidelines can be used as an advisory for the practitioners and should not consider as a rigid criterion.

9.2. Open challenges and future research directions

The existing appearance-based gaze estimation methods can be broadly divided into single-user gaze estimation and multi-user gaze estimation. Multi-user gaze estimation has not received considerable attention in the literature thus far. With the adaptation of DL-based techniques in this domain, most of the studies have progressed into gaze estimation in real-world scenarios with unconstrained settings in the last decade. As per this adaptation, the field has been confronted with numerous challenges and future opportunities. Achieving real-time inference speeds for multi-user gaze estimation has not yet been explored and remains a significant challenge in the field. The application-specific approach of Bermejo et al. (2020) for estimating shoppers' gaze in retail has reported an average of 0.52 FPS for the task.

Moreover, a generalized DL model for multi-user gaze estimation in unconstrained settings has not been explored and remains a challenge.

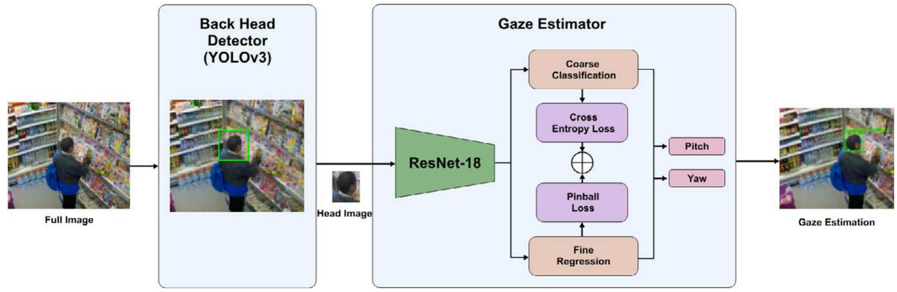


Fig. 14. Representation of the EyeShopper system architecture by Bermejo et al. (2020).

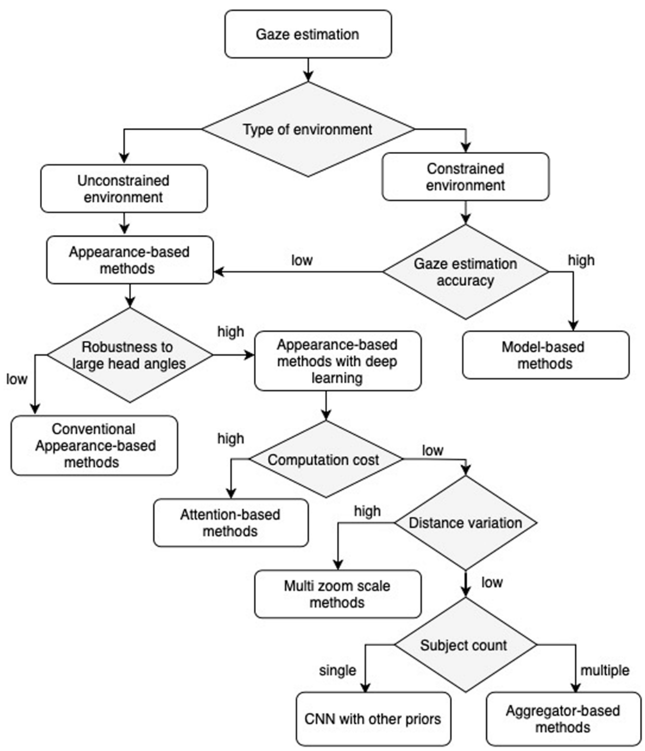


Fig. 15. Guide to select a deep learning-based gaze estimation approach.

This generalized model should not be restricted to a specific application, environmental constraints and a given number of users. In another point of view, the eye gaze estimation solutions can be integrated with other related fields like human tracker (Gamage et al., 2018) and face detection (Meedeniya & Ratnaweera, 2007) systems to provide a complete product in a low-cost environment, where the models can be deployed in edge devices (Shashirangana et al., 2021). Also, a standard framework can be developed for performance evaluation of eye gaze systems (Kar & Corcoran, 2017).

Furthermore, the success of DL algorithms is due to the availability of large-scale datasets and computational resources. In this field, the requirement of a large-scale generalized dataset remains a substantial challenge. The currently available, GazeFollow dataset is limited to 2D multi-user gaze annotations in the image coordinates. These challenges are the basis for future research. Thus, the future directions can be summarized as follows.

- DCNN based approaches for multi-user gaze estimation have only been explored to a small extent in the literature. Current work has considered a few application domains like retail industry and crowd behavior analysis (Bermejo et al., 2020; Kodama et al., 2018; Tomas et al., 2021). Therefore, future research can consider applying multi-user gaze estimation in different application domains.
- Most CNN-based techniques reviewed in this paper for multi-user gaze estimation have not focused on the throughput of the approach. Future work can focus on acquiring a trade-off between the accuracy and the inference rate to produce a viable solution in unconstrained environments.
- Although multiple datasets exist for single-user gaze estimation, a standard publicly available dataset for multi-user gaze estimation remains a limitation. Future work can produce a large-scale generalized multi-user gaze dataset considering different head poses, illumination conditions, facial and head occlusions, subject variations and target variations.
- In the multi-user gaze estimation literature, a standard performance evaluation framework has not been observed. Hence, future work can develop a framework for performance evaluation in multi-user gaze estimation considering unconstrained environments, target variations, subject variations, accuracy and inference rates.

This survey presented an in-depth overview of DL-based gaze estimation techniques focusing on multi-user gaze estimation in real-world conditions by highlighting their advantages and limitations. Furthermore, we provided critical analysis on the related models, described available datasets, coordinate systems, performance evaluation metrics and standards, together with the challenges and future opportunities in the field. Although only a few studies are done in the specific field of multi-user gaze estimation, our study discussed state-of-the-art research with a comprehensive benchmark to encourage more work in this field. We believe that this field possesses a high potential in demand for gaze estimation applications in real-world settings. Finally, this survey can be used as a guideline for DL-based gaze estimation research.

10. Conclusion

Eye gaze estimation solutions are beneficial to many application domains including commercial, social and medical health. This survey mainly explored state-of-the-art approaches used in eye gaze research focusing on deep learning techniques. This study critically analyzed the related models in appearance-based gaze estimation approaches using deep learning techniques. In comparison to model-based methods and conventional appearance-based methods, appearance-based methods with deep learning perform robustly in unconstrained environment settings including extreme head-pose variations, illumination conditions, eye and face occlusions. Furthermore, they can learn a complex non-linear mapping function directly from image data to gaze without the requirement of a dedicated device. It was observed that single-user gaze estimation approaches have been broadly studied in constrained and unconstrained environments, achieving near-human performance. However, multi-user gaze estimation studies have been explored in few application domains including retail and crowd-behavior analysis.

Moreover, we have presented the strengths and challenges in related techniques and the features of publicly available datasets. Finally, we have provided suggestions for selecting eye gaze estimation approaches and discussed possible future research directions, which can be beneficial for researchers and developers in the field.

CRediT authorship contribution statement

Primesh Pathirana: Material preparation, Analysis, Writing. **Shashimal Senarath:** Material preparation, Analysis, Writing. **Dulani Meedeniya:** Conceptualization, Design of the study, Supervised the entire study and provided critical feedback and refined the manuscript completely. **Sampath Jayarathna:** Research idea, Methodology, Fine-tune the paper.

Declaration of competing interest

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

References

Akinyelu, A. A., & Blignaut, P. (2020). Convolutional neural network-based methods for eye gaze estimation: A survey. *IEEE Access*, *8*, 142581–142605. http://dx.doi.org/10.1109/ACCESS.2020.3013540.

Bermejo, C., Chatzopoulos, D., & Hui, P. (2020). EyeShopper: EStimating shoppers' gaze using CCTV cameras. In *MM 2020 - Proceedings of the 28th ACM international conference on multimedia* (pp. 2765–2774). http://dx.doi.org/10.1145/3394171.3413683.

Cazzato, D., Leo, M., Distante, C., & Voos, H. (2020). When i look into your eyes: A survey on computer vision contributions for human gaze estimation and tracking. *Sensors (Switzerland)*, *20*(13), 1–42. http://dx.doi.org/10.3390/s20133739.

Chao, P., Kao, C.-Y., Ruan, Y.-S., Huang, C.-H., & Lin, Y.-L. (2019). Hardnet: A low memory traffic network. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3552–3561).

Cheng, Y., Wang, H., Bao, Y., & Lu, F. (2021). Appearance-based gaze estimation with deep learning: A review and benchmark. (pp. 1–21). arXiv:2104.12668 URL: http://arxiv.org/abs/2104.12668.

Chennamma, H. R., & Yuan, X. (2013). A survey on eye-gaze tracking techniques, Vol. 4, pp. 388–393. arXiv:1312.6410 URL: http://arxiv.org/abs/1312.6410.

Chong, E., Ruiz, N., Wang, Y., Zhang, Y., Rozga, A., & Rehg, J. M. (2018). Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* In *LNCS*: In *LNCS*: *Vol. 11209*, 397–412. http://dx.doi.org/10.1007/978-3-030-01228-1_24.arXiv:1807.10437,

Chong, E., Wang, Y., Ruiz, N., & Rehg, J. M. (2020). Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5396–5406).

De Silva, S., Dayarathna, S., Ariyarathne, G., Meedeniya, D., Jayarathna, S., & Michalek, A. (2021). Computational decision support system for ADHD identification. *International Journal of Automation and Computing (IJAC)*, *18*(2), 233–255. http://dx.doi.org/10.1007/s11633-020-1252-1.

De Silva, S., Dayarathna, S., Ariyarathne, G., Meedeniya, D., Jayarathna, S., Michalek, A., & Jayawardena, G. (2019). A rule-based system for ADHD identification using eye movement data. In *Proceedings of the moratuwa engineering research conference (MERCon)* (pp. 538–543). http://dx.doi.org/10.1109/MERCon.2019.8818865.

Everingham, M., Gool, L. V., Williams, C. K. I., Winn, J., & Zisserman, A. (2009). The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, *88*, 303–338.

Fang, Y., Tang, J., Shen, W., Shen, W., Gu, X., Song, L., & Zhai, G. (2021). Dual attention guided gaze target detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11390–11399).

Fischer, T., Chang, H. J., & Demiris, Y. (2018). Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 334–352).

Gamage, G., Sudasingha, I., Perera, I., & Meedeniya, D. (2018). Reinstating dlib correlation human trackers under occlusions in human detection based tracking. In *Proceedings of the 18th international conference on advances in ICT for emerging regions (ICTer)* (pp. 92–98). http://dx.doi.org/10.1109/ICTER.2018.8615551.

Ghani, M. U., Chaudhry, S., Sohail, M., & Geelani, M. N. (2013). Gazepointer: A real time mouse pointer control implementation based on eye gaze tracking. In *Proceedings of the 16th international multi-topic conference* (pp. 154–159). IEEE.

Goldberg, J. H., & Kotval, X. P. (1999). Computer interface evaluation using eye movements: methods and constructs. *International Journal of Industrial Ergonomics*, *24*(6), 631–645.

Graves, A., Fernández, S., & Schmidhuber, J. (2005). Bidirectional LSTM networks for improved phoneme classification and recognition. In *International conference on artificial neural networks* (pp. 799–804). Springer.

Guojun, Y., & Saniie, J. (2016). Eye tracking using monocular camera for gaze estimation applications. In *IEEE international conference on electro information technology, Vol. 2016-August* (pp. 292–296). http://dx.doi.org/10.1109/EIT.2016.7535254.

Gwon, S. Y., Cho, C. W., Lee, H. C., Lee, W. O., & Park, K. R. (2013). Robust eye and pupil detection method for gaze tracking. *International Journal of Advanced Robotic Systems*, *10*(2), 98.

Handelman, G. S., Kok, H. K., Chandra, R. V., Razavi, A. H., Huang, S., Brooks, M., Lee, M. J., & Asadi, H. (2019). Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. *American Journal of Roentgenology*, *212*(1), 38–43.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Huang, Q., Veeraraghavan, A., & Sabharwal, A. (2017). Tabletgaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. *Machine Vision and Applications*, *28*(5), 445–461.

Ji, Q., Zhu, Z., & Lan, P. (2004). Real-time nonintrusive monitoring and prediction of driver fatigue. *IEEE Transactions on Vehicular Technology*, *53*(4), 1052–1068.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision* (pp. 2106–2113). IEEE.

Kacete, A., Séguier, R., Collobert, M., & Royan, J. (2016). Unconstrained gaze estimation using random forest regression voting. In *Asian conference on computer vision* (pp. 419–432). Springer.

Kar, A., & Corcoran, P. (2017). A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms. *IEEE Access*, *5*(November), 16495–16519. http://dx.doi.org/10.1109/ACCESS.2017.2735633.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8110–8119).

Kasprowski, P., & Harężlak, K. (2014). Cheap and easy PIN entering using eye gaze. *Annales Universitatis Mariae Curie-Sklodowska, Sectio AI–Informatica*, *14*(1), 75–84.

Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., & Torralba, A. (2019). Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6912–6921).

Kerr-Gaffney, J., Harrison, A., & Tchanturia, K. (2019). Eye-tracking research in eating disorders: A systematic review. *International Journal of Eating Disorders*, *52*(1), 3–27.

Khan, M. Q., & Lee, S. (2019). Gaze and eye tracking: techniques and applications in ADAS. *Sensors*, *19*(24), 5540.

Kim, J., Stengel, M., Majercik, A., De Mello, S., Dunn, D., Laine, S., McGuire, M., & Luebke, D. (2019). Nvgaze: an anatomically-informed dataset for low-latency, near-eye gaze estimation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–12). New York, NY, USA: Association for Computing Machinery, URL: https://doi.org/10.1145/3290605.3300780.

Klaib, A. F., Alsrehin, N. O., Melhem, W. Y., Bashtawi, H. O., & Magableh, A. A. (2021). Eye tracking algorithms, techniques, tools, and applications with an emphasis on machine learning and Internet of Things technologies. *Expert Systems with Applications*, *166*, 114037. http://dx.doi.org/10.1016/j.eswa.2020.114037.

Kodama, Y., Kawanishi, Y., Hirayama, T., Deguchi, D., Ide, I., Murase, H., Nagano, H., & Kashino, K. (2018). Localizing the gaze target of a crowd of people. In *Proceedings of the 14th Asian conference on computer vision* (pp. 15–30). Springer.

Koenker, R. (2005). *Quantile regression: economic society monograph serie*. New York: Cambridge University Press.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications on ACM*, *60*(6), 84–90. http://dx.doi.org/10.1145/3065386.

Kumar, M., Garfinkel, T., Boneh, D., & Winograd, T. (2007). Reducing shoulder-surfing by using gaze-based password entry. In *Proceedings of the 3rd symposium on usable privacy and security* (pp. 13–19).

Kumar, M., Paepcke, A., & Winograd, T. (2007). Eyepoint: practical pointing and selection using gaze and keyboard. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 421–430).

Kwon, Y.-M., Jeon, K.-W., Ki, J., Shahab, Q. M., Jo, S., & Kim, S.-K. (2006). 3D Gaze estimation and interaction to stereo display. *International Journal of Virtual Reality*, *5*(3), 41–45. http://dx.doi.org/10.20870/ijvr.2006.5.3.2697.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Lee, E. C., Ko, Y. J., & Park, K. R. (2009). Gaze tracking based on active appearance model and multiple support vector regression on mobile devices. *Optimization and Engineering*, *48*(7), Article 077002.

Lee, S.-H., Lee, J.-Y., & Choi, J.-S. (2011). Design and implementation of an interactive HMD for wearable AR system. In *2011 17th Korea-Japan joint workshop on frontiers of computer vision (FCV)* (pp. 1–6). IEEE.

Lee, H. C., Luong, D. T., Cho, C. W., Lee, E. C., & Park, K. R. (2010). Gaze tracking system at a distance for controlling IPTV. *IEEE Transactions on Consumer Electronics*, *56*(4), 2577–2583.

Lian, D., Yu, Z., & Gao, S. (2018). Believe it or not, we know what you are looking at!. In *LNCS*: *Vol. 11363*, *Proceedings of the 14th Asian conference on computer vision* (pp. 35–50). Springer, http://dx.doi.org/10.1007/978-3-030-20893-6_3.

Lian, D., Zhang, Z., Luo, W., Hu, L., Wu, M., Li, Z., Yu, J., & Gao, S. (2019). RGBD based gaze estimation via multi-task CNN. In *Proceedings of the AAAI conference on artificial intelligence, Vol. 33* (pp. 2488–2495).

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117–2125).

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Proceedings of the 13th European conference on computer vision* (pp. 740–755). Springer.

Liu, D., Dong, B., Gao, X., & Wang, H. (2015). Exploiting eye tracking for smartphone authentication. In *International conference on applied cryptography and network security* (pp. 457–477). Springer.

Lu, F., Okabe, T., Sugano, Y., & Sato, Y. (2014a). Learning gaze biases with head motion for head pose-free gaze estimation. *Image Vision Comput.*, *32*, 169–179. http://dx.doi.org/10.1016/j.imavis.2014.01.005.

Lu, F., Sugano, Y., Okabe, T., & Sato, Y. (2014b). Adaptive linear regression for appearance-based gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*, 2033–2046.

Mahanama, B., Jayawardana, Y., & Jayarathna, S. (2020). Gaze-net: Appearance-based gaze estimation using capsule networks. In *Proceedings of the 11th augmented human international conference* (pp. 18–21). http://dx.doi.org/10.1145/3396339.3396393.

Meedeniya, D., & Ratnaweera, A. (2007). Enhanced face recognition through variation of principle component analysis (PCA). In *Proceedings of international conference on industrial and information systems (ICIIS)* (pp. 347–352). Peradeniya, SriLanka: http://dx.doi.org/10.1109/iciinfs.2007.4579200.

Mishra, A., & Lin, H.-T. (2020). 360-Degree gaze estimation in the wild using multiple zoom scales. arXiv preprint arXiv:2009.06924.

Mora, K. A. F., Monay, F., & Odobez, J. (2014). Eyediap: a database for the development and evaluation of gaze estimation algorithms from RGB and RGB-d cameras. In *Proceedings of the symposium on eye tracking research and applications* (pp. 255–258). http://dx.doi.org/10.1145/2578153.2578190.

Morimoto, C. H., & Mimica, M. R. (2005). Eye gaze tracking techniques for interactive applications. *Computer Vision and Image Understanding*, *98*(1), 4–24.

Muñoz-Salinas, R., Yeguas-Bolivar, E., Saffiotti, A., & Medina-Carnicer, R. (2012). Multi-camera head pose estimation. *Machine Vision and Applications*, *23*(3), 479–490.

Oved, D., Alvarado, I., & Gallo, A. (2018). Real-time human pose estimation in the browser with tensorflow. *TensorFlow Medium*, URL: https://blog.tensorflow.org/2018/05/real-time-human-pose-estimation-in.html.

Park, H. S., Jain, E., & Sheikh, Y. (2012). 3D social saliency from head-mounted cameras. In *Proceedings of the 25th international conference on neural information processing systems - Volume 1* (pp. 422–430).

Park, H. S., & Shi, J. (2015). Social saliency prediction. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition, Vol. 07-12-June* (pp. 4777–4785). http://dx.doi.org/10.1109/CVPR.2015.7299110.

Piumsomboon, T., Lee, G., Lindeman, R. W., & Billinghurst, M. (2017). Exploring natural eye-gaze-based interaction for immersive virtual reality. In *2017 IEEE symposium on 3d user interfaces (3DUI)* (pp. 36–39). IEEE.

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., & Koltun, V. (2020). Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, http://dx.doi.org/10.1109/TPAMI.2020.3019967.

Raptis, G. E., Katsini, C., Belk, M., Fidas, C., Samaras, G., & Avouris, N. (2017). Using eye gaze data and visual activities to infer human cognitive styles: method and feasibility studies. In *Proceedings of the 25th conference on user modeling, adaptation and personalization* (pp. 164–173).

Recasens, A. R. C. (2016). *Where are they looking?* (Ph.D. thesis), Massachusetts Institute of Technology.

Ruiz, N., Chong, E., & Rehg, J. M. (2018). Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 2074–2083).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., C, A., Berg, & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252. http://dx.doi.org/10.1007/s11263-015-0816-y.

Saad, A., Elkafrawy, D. H., Abdennadher, S., & Schneegass, S. (2020). Are they actually looking? identifying smartphones shoulder surfing through gaze estimation. In *ACM symposium on eye tracking research and applications* (pp. 1–3).

Shashirangana, J., Padmasiri, H., Meedeniya, D., Perera, C., Nayak, S. R., Nayak, J., Vimal, S., & Kadry, S. (2021). License plate recognition using neural architecture search for edge devices. *International Journal of Intelligent Systems*, 1–38. http://dx.doi.org/10.1002/int.22471.

Sibert, L. E., & Jacob, R. J. (2000). Evaluation of eye gaze interaction. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 281–288).

Sidorakis, N., Koulieris, G. A., & Mania, K. (2015). Binocular eye-tracking for the control of a 3D immersive multimedia user interface. In *2015 IEEE 1st workshop on everyday virtual reality (WEVR)* (pp. 15–18). IEEE.

Smith, B. A., Yin, Q., Feiner, S. K., & Nayar, S. K. (2013). Gaze locking: passive eye contact detection for human-object interaction. In *Proceedings of the 26th annual ACM symposium on user interface software and technology* (pp. 271–280).

Špakov, O., & Miniotas, D. (2005). Gaze-based selection of standard-size menu items. In *Proceedings of the 7th international conference on multimodal interfaces* (pp. 124–128).

Sugano, Y., Matsushita, Y., & Sato, Y. (2014). Learning-by-synthesis for appearance-based 3D gaze estimation. In *2014 IEEE conference on computer vision and pattern recognition* (pp. 1821–1828).

Sugano, Y., Zhang, X., & Bulling, A. (2016). AggreGaze: COllective estimation of audience attention on public displays. In *UIST 2016 - Proceedings of the 29th annual symposium on user interface software and technology* (pp. 821–831). http://dx.doi.org/10.1145/2984511.2984536.

Sun, W., Sun, N., Guo, B., Jia, W., & Sun, M. (2016). An auxiliary gaze point estimation method based on facial normal. *Pattern Analysis and Applications*, *19*(3), 611–620.

Sun, X., Xu, L., & Yang, J. (2007). Driver fatigue alarm based on eye detection and gaze estimation. In *MIPPR 2007: Automatic target recognition and image analysis; and multispectral image acquisition, Vol. 6786*. International Society for Optics and Photonics, Article 678612.

Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2018). Facevr: Real-time facial reenactment and eye gaze control in virtual reality. *ACM Transactions on Graphics*, *37*, http://dx.doi.org/10.1145/3182644.

Tomas, H., Reyes, M., Dionido, R., Ty, M., Mirando, J., Casimiro, J., Atienza, R., & Guinto, R. (2021). Goo: A dataset for gaze object prediction in retail environments. arXiv:2105.10793.

Tsukada, A., Shino, M., Devyver, M., & Kanade, T. (2011). Illumination-free gaze estimation method for first-person vision wearable device. *Proceedings of the IEEE international conference on computer vision*, 2084–2091. http://dx.doi.org/10.1109/ICCVW.2011.6130505.

Velichkovsky, B. B., Rumyantsev, M. A., & Morozov, M. A. (2014). New solution to the midas touch problem: Identification of visual commands via extraction of focal fixations. *Procedia Computer Science*, *39*, 75–82.

Wang, H., Dong, X., Chen, Z., & Shi, B. E. (2015). Hybrid gaze/EEG brain computer interface for robot arm control on a pick and place task. In *2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (pp. 1476–1479). IEEE.

Wang, J., & Olson, E. (2016). AprilTag 2: Efficient and robust fiducial detection. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 4193–4198).

Wang, H., Pi, J., Qin, T., Shen, S., & Shi, B. E. (2018). SLAM-based localization of 3D gaze using a mobile eye tracker. In *Proceedings of the 2018 ACM symposium on eye tracking research & applications* (pp. 1–5).

Wang, W., & Shen, J. (2017). Deep visual attention prediction. *IEEE Transactions on Image Processing*, *27*(5), 2368–2378.

Wang, Y., Zhao, T., Ding, X., Peng, J., Bian, J., & Fu, X. (2018). Learning a gaze estimator with neighbor selection from large-scale synthetic eye images. *Knowledge-Based Systems*, *139*(C), 41–49. http://dx.doi.org/10.1016/j.knosys.2017.10.010.

xiong Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition* (pp. 3485–3492).

Xu, P., Ehinger, K. A., Zhang, Y., Finkelstein, A., Kulkarni, S. R., & Xiao, J. (2015). Turkergaze: Crowdsourcing saliency with webcam based eye tracking. arXiv preprint arXiv:1504.06755.

Yao, B., Jiang, X., Khosla, A., Lin, A., Guibas, L., & Fei-Fei, L. (2011). Human action recognition by learning bases of action attributes and parts. In *2011 international conference on computer vision* (pp. 1331–1338).

Young, L. R., & Sheena, D. (1975). Survey of eye movement recording methods. *Behavior Research Methods & Instrumentation*, *7*(5), 397–429.

Yu, L., Xu, J., & Huang, S. (2016). Eye-gaze tracking system based on particle swarm optimization and bp neural network. In *2016 12th world congress on intelligent control and automation (WCICA)* (pp. 1269–1273). http://dx.doi.org/10.1109/WCICA.2016.7578296.

Zhai, S., Morimoto, C., & Ihde, S. (1999). Manual and gaze input cascaded (MAGIC) pointing. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 246–253).

Zhang, X., Park, S., Beeler, T., Bradley, D., Tang, S., & Hilliges, O. (2020). Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *Proceedings of the 16th European conference on computer vision* (pp. 365–381). http://dx.doi.org/10.1007/978-3-030-58558-7_22.

Zhang, X., Sugano, Y., & Bulling, A. (2019). Evaluation of appearance-based methods and implications for gaze-based applications. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–13).

Zhang, X., Sugano, Y., Fritz, M., & Bulling, A. (2015). Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4511–4520).

Zhang, X., Sugano, Y., Fritz, M., & Bulling, A. (2017). It's written all over your face: Full-face appearance-based gaze estimation. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)* (pp. 2299–2308).

Zhang, X., Sugano, Y., Fritz, M., & Bulling, A. (2019). Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*, 162–175.

Zheng, R., Nakano, K., Ishiko, H., Hagita, K., Kihira, M., & Yokozeki, T. (2015). Eye-gaze tracking analysis of driver behavior while interacting with navigation systems in an urban area. *IEEE Transactions on Human-Machine Systems*, *46*(4), 546–556.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning deep features for scene recognition using places database. In *Proceedings of the 27th advances in neural information processing systems* (pp. 487–495).

Mr. **Primesh Pathirana** is an undergraduate in the Department of Computer Science and Engineering, Faculty of Engineering, at the University of Moratuwa, Sri Lanka. His main research interests are computer vision, and deep learning.

Mr. **Shashimal Senarath** is an undergraduate in the Department of Computer Science and Engineering, Faculty of Engineering, at the University of Moratuwa, Sri Lanka. His main research interests are computer vision, and deep learning.

Prof. **D. Meedeniya** is a Professor in Computer Science and Engineering at the University of Moratuwa, Sri Lanka. She holds a Ph.D. in Computer Science from the University of St Andrews, United Kingdom. She leads the Bio-Health Informatics group at her department and engages in many collaborative research. She is a co-author of 100+ publications in indexed journals, peer reviewed conferences and international book chapters. She serves as a reviewer, program committee and editorial team member in many international conferences and journals. Her main research interests are Software modeling and design, Bio-Health Informatics, Deep Learning and Technology-enhanced learning. She is a Fellow of HEA (UK), MIET, MIEEE, Member of ACM and a Chartered Engineer registered at EC (UK).

Dr. **Sampath Jayarathna** is an Assistant Professor of Computer Science at Old Dominion University, where he is associated with Web Science and Digital Libraries (WS-DL) research group. Dr. Jayarathna earned a Ph.D. in Computer science from the Texas A&M University College Station in 2016. His research interest includes machine learning, information retrieval, data science, eye tracking, and brain–computer interfacing. Dr. Jayarathna is a member of ACM, IEEE, and Sigma XI.