Unobtrusive and Pervasive Video-Based Eye-Gaze Tracking

Stefania Cristina[1,*], Kenneth P. Camilleri[1]

Department of Systems and Control Engineering, University of Malta, Msida, MSD 2080, Malta

Abstract

Eye-gaze tracking has long been considered a desktop technology that finds its use inside the traditional office setting, where the operating conditions may be controlled. Nonetheless, recent advancements in mobile technology and a growing interest in capturing natural human behaviour have motivated an emerging interest in tracking eye movements within unconstrained real-life conditions, referred to as pervasive eye-gaze tracking. This critical review focuses on emerging passive and unobtrusive video-based eye-gaze tracking methods in recent literature, with the aim to identify different research avenues that are being followed in response to the challenges of pervasive eye-gaze tracking. Different eye-gaze tracking approaches are discussed in order to bring out their strengths and weaknesses, and to identify any limitations, within the context of pervasive eye-gaze tracking, that have yet to be considered by the computer vision community.

Keywords: Review, eye-gaze tracking, pervasive, unobtrusive, passive

1. Introduction

The notion of measuring and tracking the human eye-gaze has been receiving increasing interest for over ten decades, ever since the pioneering work on eye tracking in the 1870s revealed a wealth of information contained within the eye movements [1]. Research at the earliest time was mainly concerned

---

*Corresponding author

Email addresses: `stefania.cristina@um.edu.mt` (Stefania Cristina), `kenneth.camilleri@um.edu.mt` (Kenneth P. Camilleri)

with studies of scan paths performed by the eyes during reading, and in the absence of sophisticated machines for eye tracking, early recording methods were based on mere visual observation often requiring the insertion of specks or eyecups over the eyeball which moved in response to the ocular motion. Following this emerging interest in eye movement research, many subsequent studies were driven towards the improvement of eye-gaze tracking techniques leading to the development of new concepts such as video-based eye-gaze tracking which captures the eye movements non-intrusively by means of a video camera alone. The introduction of computers in the 1980s broadened the otherwise narrow application of eye tracking technology in clinical studies to user-oriented applications outside the clinical laboratory. The application of eye-gaze tracking within the domain of human-computer interaction (HCI) was among the first to be established to assist physically challenged individuals to operate a computer by eye movement alone [2]. Eye-gaze tracking technology has since found its way into other domains, including automotive engineering for car driver assistance and monitoring [3–6], human-robot interaction [7] and marketing and advertising research [8].

Notwithstanding the extensive development of stationary desktop-based approaches for eye-gaze tracking, recent advancements in mobile technology and a growing interest in capturing natural human behaviour under unconstrained real-life conditions, often referred to as *in the wild*, give rise to challenges that go beyond the controlled settings for which existing tracking methods have been developed. This led to the emergence of *pervasive eye-gaze tracking*, as originally coined by Bulling et al. [9], referring to the endeavour of tracking and analysing the eye movements continuously in daily life settings. The notion behind this paradigm is multi-faceted and typically relates to characteristics that facilitate eye-gaze tracking in uncontrolled real-life scenarios, such as robustness to varying illumination conditions and extensive head rotations, the capability of estimating the eye-gaze at increased distance from the imaging hardware, reduced or implicit calibration in order to allow for situations that do not permit user cooperation and calibration awareness, and the estimation of eye-gaze on mobile devices comprising integrated imaging hardware without requiring further hardware modification. This will potentially broaden the application areas for eye-gaze tracking within scenarios that may not permit for controlled conditions, such as for gaze-based interaction in public spaces [10–12].

This critical review, therefore, aims to investigate the research avenues that are being followed in response to emerging challenges associated with
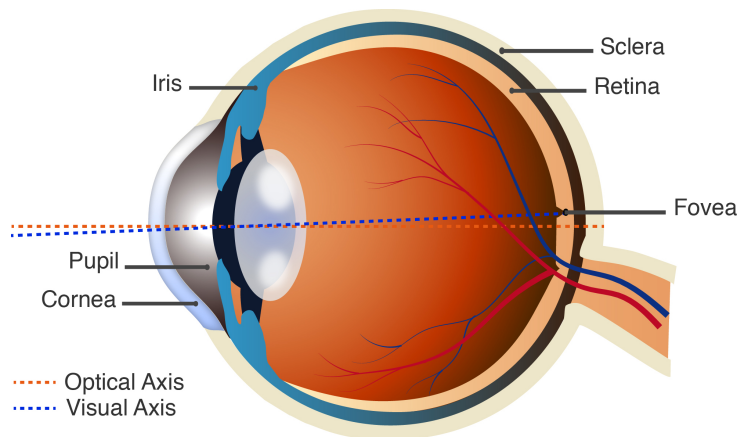
Figure 1: An illustration of the human eye anatomy (left) and the formation of Purkinje images from different components making up the eyeball structure (right).

eye-gaze tracking *in the wild* and to identify research gaps that yet need to be addressed. In the following sections a brief overview of the human eye anatomy is first presented in Section 2, in order to define and illustrate the relevant ocular components that will be referred to in subsequent sections. Section 3 gives an overview of existing eye-gaze tracking methods, followed by a brief description of the contribution of this review in Section 4. This sets the stage for the emerging challenges that will be discussed in Section 5. An overarching discussion is provided in Section 6 in order to summarise the research work that has been carried out so far in light of the emerging challenges, and to identify possible future avenues in Section 7. Finally, Section 8 concludes this review.

2. Anatomical Terminology of the Human Eye

The human eye is an intricate organ despite its small dimensions. Its shape approaches that of a spheroid and is often approximated as such for simplification purposes. The eyeball is free to rotate inside its socket around a centre point lying at a population average of 13.5mm behind the front vertex, termed the eyeball centre. The outer coating of the eyeball is the sclera that is particularly identifiable by its opaque white colour. Light first enters the eye through the cornea, which is the clear dome-shaped layer that covers a portion of the visible part of the eyeball, and which joins the sclera at a border that is known as the corneal limbus. Light then travels through

the iris, which is the coloured portion of the eye and acts as a diaphragm that narrows or widens the dark pupil aperture to regulate the amount of light entering the eye. Light rays are then received at the retina at the back of the eyeball, which contains cones and rods that generate the perceived image. Light rays which fall upon the fovea travel along the visual axis, such that an observer must direct the visual axis towards the object of interest to obtain a sharp image of an object. The angle which the visual axis makes with the optical axis, which passes through the centre of the pupil to the retina, is subject-dependent.

An illustration of the eyeball and its components that shall be referred to within this section is shown in Figure 1.

3. Overview of Eye-Gaze Tracking

Research efforts aimed towards the development of improved eye movement measurement techniques were always driven towards minimising discomfort and direct contact with the user without jeopardising the measurement accuracy. Through technological development, the degree of intrusion of eye movement measurement techniques lessened over the years, progressing from the highly intrusive insertion of specks or search coils onto the eyeball [13], to less intrusive electrooculographic (EOG) methods requiring the attachment of electrodes around the eye regions [14], to video-based eye-gaze tracking methods which track the eye-gaze remotely via digital cameras that capture the ocular movements in a stream of image frames alone [15]. The ease-of-use and practicality of video-based eye-gaze tracking methods outside the research laboratory offer an attractive solution for user-oriented applications, such as HCI, as evidenced by several video-based eye-gaze tracking systems that have been made commercially available for home users [16], and more recently for human authentication [17–20].

In a previous survey on video-based eye-gaze tracking techniques [15], Hansen and Ji refer to this field as having two distinct areas, which are in fact two different domains, namely, the image domain and the gaze domain, as illustrated further in Figure 2. Video-based eye-gaze tracking will always require eye detection and localisation in the image domain, hence we follow Hansen and Ji [15] in the categorisation of methods for eye detection and localisation, which are often done simultaneously. These methods are categorised by Hansen and Ji [15] according to the exploited eye image characteristics, such as the shape and contours of the open eye [21–29, 29–36], the

4

appearance of the eye region [37–51], or distinctive features around the eyes [52–65]. A further category is also included for active methods that exploit the physiological properties of the human eye by illuminating the face and eye regions with infra-red (IR) illumination for eye detection and localisation. Nonetheless, within the context of pervasive eye-gaze tracking, the effectiveness of methods that utilise IR illumination diminishes as soon as external factors, such as interfering IR illumination from the surrounding environment [66], affect the image quality or the pupil size and brightness. Hence, while in the office paradigm the use of IR illumination may be preferred, since this simplifies the task of eye detection and localisation, the problems that active methods potentially face in pervasive scenarios outweigh the benefits. In absence of any dependence upon specialised IR illumination sources and imaging hardware, passive methods are potentially better suited for a wider range of pervasive scenarios, permitting increased portability for both indoor and outdoor use, and the capability to capture natural visual behaviour in less constrained scenarios. Hence, this review will not be considering active methods any further.

Specifically, *eye detection and localisation* methods that are categorised as **shape-based**, since they exploit the shape and contours of the open eye, typically aim to fit a geometrical model to the eye region [21–29, 29–36]. These methods approach the problem of eye detection either by fitting the iris or pupil region with a simple circular [21–30] or elliptical model [29, 31, 32], or by the use of more complex models [33–36], which are more computationally demanding than their simpler counterparts but allow for detailed modelling of the eye region by including components such as the eyelids [33–36] and eye corners [36]. Both subcategories of shape-based methods typically require high contrast images that permit reliable feature extraction for shape fitting [21, 23, 24, 26, 27, 32, 33, 35], together with the initialisation of the shape models close to the eye region for successful localisation [28] or the acquisition of close-up eye region images within which the model-fitting is carried out directly [29, 32, 35]. **Appearance-based methods**, in contrast, are independent of the actual geometry since these exploit the photometric properties of the eye region for detection [37–51]. Eye detection by these methods follows different approaches, such as via machine learning [37–44] or template-matching techniques [45–51], which preserve the spatial and intensity information of the eye image pixels. Most appearance-based methods that rely on machine learning require the collection of a sizeable set of training data under the expected tracking conditions, upon which the performance

Image Domain
Eye Detection and Localisation

Shape-based methods: Exploit the shape and contours of the open eye.
Appearance-based methods: Exploit the photometric properties of the eye region.
Feature-based methods: Exploit specific informative local features of the eye region.

Gaze Domain
Calibration

Intrinsic camera calibration: Characterises the geometrical properties of the camera.
Setup calibration: Positions the camera centre within the world coordinate space.
User/personal calibration: Characterises the intrinsic anatomical properties of the eye.
Gaze calibration: Positions the eyeballs within the head and camera coordinate spaces.
Point-of-regard calibration: Positions the gaze on a monitor screen.

Gaze Domain
Gaze Estimation and Tracking

Gaze estimation under different eye and head configurations, across image frames.
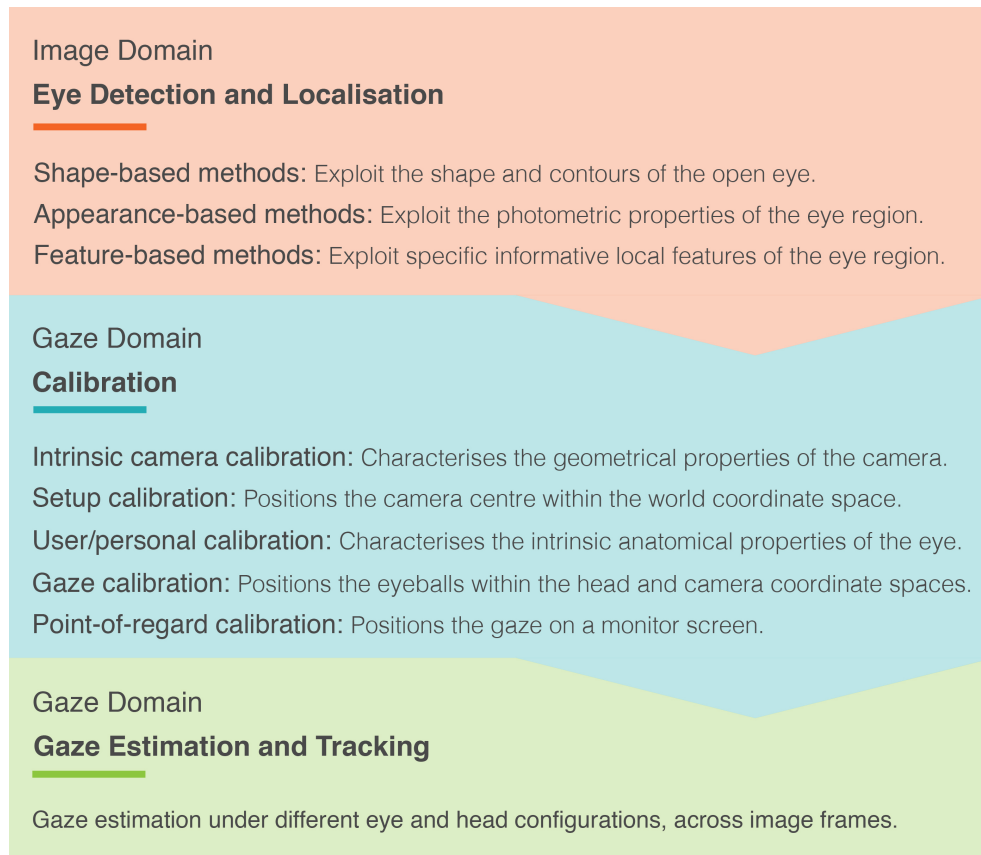
Figure 2: The field of video-based eye-gaze tracking comprises two distinct domains, namely the *image domain* and the *gaze domain*. The gaze domain further consists of two main classes of tasks, namely the *calibration* tasks, and the *gaze estimation and tracking* tasks. This figure summarises different aspects of these domains and tasks.

of eye detection is subsequently contingent [44]. Furthermore, methods that rely on image templates for eye detection are often inherently limited by rotation and scale constraints [48]. **Feature-based methods**, on the other hand, seek to identify specific informative local features of the eye region that may be less susceptible to illumination changes and variations in viewpoint [52–65]. Detection of these local features may either exploit the grey-level differences at the feature boundaries, such as the limbus boundary [52–55], or alternatively the dark and distinctive colour of the pupil [56–65]. In order to reduce the number of eye candidates that may be captured by wide field-of-view imaging hardware and which may exhibit similar features to the
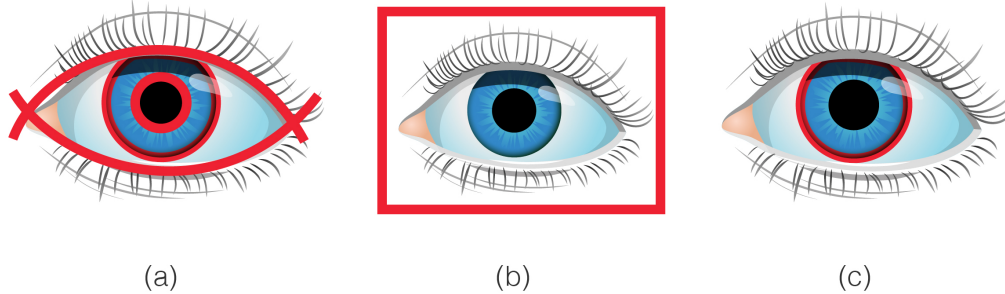
Figure 3: Eye detection and localisation methods typically exploit (a) the shape and contours of the open eye by fitting a geometrical model to the eye region, (b) the photometric appearance of the eye by employing machine learning or template-matching techniques, or (c, d) specific informative local features of the eye region, such as the limbus boundary.

eyes, feature detection is often performed within close-up eye region images [63–65]. Different methods for eye detection and localisation are summarised in Figure 2 and illustrated in Figure 3.

The gaze domain is a second component of video-based eye-gaze tracking consisting of two main classes of tasks, which are the calibration tasks and the gaze tracking tasks, as outlined in Figure 2. Calibration is typically performed to estimate parameter values that permit information from the image domain to be converted to the gaze domain, hence allowing for *gaze estimation* and subsequently *gaze tracking*, where the latter refers to the process of estimating the eye-gaze through time. We distinguish between the following calibration procedures:

(a) *Intrinsic camera calibration* determines parameter values that characterise the geometric properties of the imaging hardware, such as the focal length, the principal point offset and the skew coefficients [67].

(b) *Setup calibration* determines parameter values that position the camera centre within the world coordinate space, and hence permit transformation of points from the camera image space to the 3-dimensional world space [67].

(c) *User/personal calibration* determines parameter values that characterise the intrinsic anatomical properties of the human eye [49, 68–76]. These may include the eyeball and cornea radii, the angular offset between the optical and visual axes, and the refraction parameters.

(d) *Gaze calibration* determines parameter values that position the eyeballs within the coordinate spaces of the head and camera [64, 68–77].

(e) *Point-of-regard calibration* determines parameter values that position the gaze on a monitor screen [22, 25, 27, 37, 40, 45, 56, 57, 61, 62, 70–73, 78–87] or on a planar surface positioned at a distance from the user in physical space [43, 68, 69, 74–77], hence permitting the estimation of a point-of-regard (PoR). This may be carried out by:

 (i) *Geometric mapping* which estimates the PoR by intersecting two gaze vectors, each projecting from each eyeball, in 3-dimensional space [77], or by intersecting a single gaze vector, often estimated by averaging the two gaze vectors projecting from each eyeball [47, 68–76, 88], with a prior defined surface. This is illustrated in Figure 4(a).

 (ii) *Implicit mapping* which determines a mapping function that estimates the PoR by mapping the image contents directly to screen coordinates. This may be achieved by means of machine learning techniques [22, 37, 38, 40, 43, 45, 56, 57, 61, 62, 82–87], or by mapping the distance of a specific eye region feature from a reference image point onto a corresponding fixation point on a monitor screen [25, 27, 78–81]. Figure 4(b) illustrates this calibration procedure.

The outlined procedures require different levels of calibration effort per session. For instance, calibration procedures (a) and (b) may be carried out once prior to use, and remain valid as long as the properties pertaining to the geometry of the setup and the imaging hardware remain unchanged. Subject-dependent parameter values characterising the anatomical properties of the human eyeballs may also be re-used between sessions after having been estimated via calibration procedure (c), since these also remain unchanged. The parameter values estimated by calibration (d), on the other hand, are typically re-initialised at every session since, while the eyeball positions remain fixed within the head coordinate space of the same user, their placement within the camera coordinate space depends on the head pose. Calibration (e) is also typically performed at every session, since this takes into consideration the positioning of the user with respect to the surface of interest in order to estimate the PoR. During calibration procedures (c) to
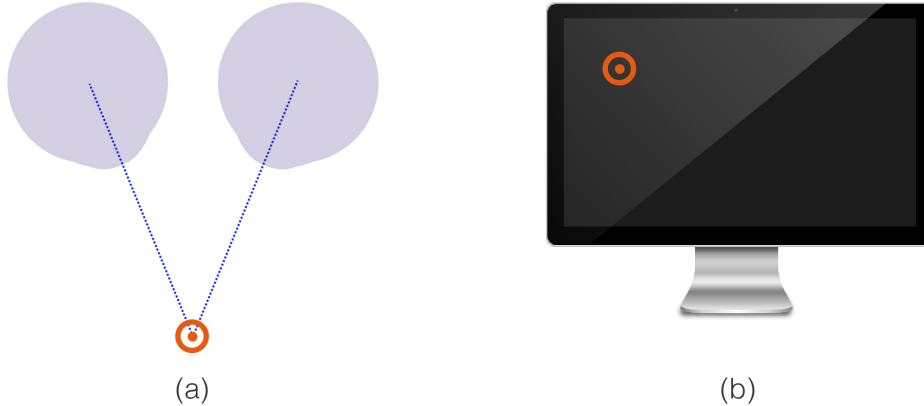
Figure 4: Point-of-regard calibration may be carried out by (a) geometric mapping, which estimates the point-of-regard from gaze vectors projecting from each eyeball, or (b) implicit mapping, which determines a mapping function that estimates the point-of-regard by mapping the image contents directly to screen coordinates.

(e), the user is either required to maintain a frontal eye and head pose [68], or perform specific eye [25, 27, 37, 40, 43, 45, 49, 56, 57, 61, 62, 64, 69–73, 79–86] or head movements [62, 64, 69, 77] by fixating at markers appearing sequentially on a monitor screen [25, 27, 40, 45, 49, 56, 57, 61, 62, 64, 70–73, 79–86], or positioned at a distance from the user in physical space [43, 77]. This may require substantial user effort, especially if calibration requires the collection of hundreds [38, 57, 86] or thousands [40, 43, 56, 61, 83, 85] of data samples under the expected tracking conditions, in order to serve as training data for a machine learning algorithm.

The second class of tasks within the gaze domain deals with *gaze estimation and tracking*. In the process of estimating and tracking the gaze, the head pose plays an important role in conjunction with the eyeball orientation: the head pose typically defines a coarse estimate of the gaze direction, while the eyeball orientation refines upon this estimate to define the gaze direction at a finer level. The estimation of gaze under different eye and head configurations is a desirable aspect of pervasive eye-gaze tracking, since this would lift any constraints on the user and permit natural head movement. While more recent eye-gaze tracking methods are increasingly allowing more freedom of movement, as will be further discussed in Section 5.4, many other methods have conventionally resorted to stationary head pose constraints

[22, 45, 49, 57, 75, 78, 79, 81–84], often by the use of a chin-rest [45, 82], or made use of wearable setups that circumvent the need for head pose estimation but cannot track the eye-gaze remotely [43, 78, 85]. Furthermore, the appearance of the face and eye regions changes significantly with head motion, resulting in many possible combinations of face appearance and eyeball orientation which correspond to the same gaze direction. This is especially taxing on methods that exploit the photometric appearance of the eye region for gaze estimation, since these would typically require a large corpus of training data that captures the eye appearance under many head orientations [38, 40, 44].

Most of the overviewed video-based methods focus mainly on the application of eye-gaze tracking as a desktop technology. Hence, most research effort in past years has been mainly dedicated towards the development of methods that fit controlled conditions where the head movement, if any, is mostly constrained to a small volume [22, 35, 45, 49, 57, 75, 78, 79, 81–84, 89], the illumination conditions are stable or controllable by the projection of visible [62] or IR [29, 31, 57, 90] illumination, and both calibration [22, 25, 27, 35, 43, 47, 49, 64, 64, 69–80, 88, 91] or the collection of a training data set [37, 38, 40, 43, 45, 52, 56, 57, 61, 62, 82–86, 92] may be carried out as required. Such conditions limit the applicability of these video-based eye-gaze tracking methods within the uncontrolled settings associated with pervasive eye-gaze tracking and, hence, call for methods that address the challenges which emerge from these limitations.

We identify five emerging challenges of pervasive eye-gaze tracking, namely the estimation of gaze from low-resolution eye images; the estimation of gaze from sparse, synthesised or person-independent training samples; reduced or implicit calibration; head pose invariant gaze estimation; and eye-gaze tracking on mobile platforms. The next sections, therefore, define our research contribution and subsequently discuss emerging work and challenges within the context of pervasive eye-gaze tracking.

4. Our Contribution

This paper presents a critical review of recent technical advancements in eye-gaze tracking within the context of pervasive applications. In light of emerging demands for pervasive applications, most state-of-the-art video-based eye-gaze tracking methods that had previously been reviewed by Hansen and Ji [15], face new challenges that go beyond the classical desktop setting

10

for which they were developed. Hence, in contrast to the survey of Hansen and Ji [15], our review aims to investigate and highlight the research avenues that are specifically being followed in response to emerging challenges associated with eye-gaze tracking *in the wild.*

The following sections categorise emerging passive and unobtrusive video-based eye-gaze tracking methods in recent literature, with the aim to identify different research avenues that are being followed in response to the identified challenges of pervasive eye-gaze tracking. We aim to bring out their strengths and weaknesses, and to identify any limitations, within the context of pervasive eye-gaze tracking, that have yet to be considered further by the computer vision community.

5. Challenges of Pervasive Eye-Gaze Tracking

5.1. Challenge A: Gaze Estimation from Low-Resolution Eye Images

The shift towards pervasive eye-gaze tracking has brought an increasing interest in the development of low cost alternatives to commercial eye-gaze tracking systems using consumer electronics, such as webcams. In comparison with the image quality captured by higher-end cameras used in commercially available eye-gaze trackers, consumer-grade cameras typically acquire images of inferior quality and wider field-of-view, which capture an expansive portion of the background. If the camera is placed at a distance from the user [37, 38, 40, 81, 82, 84, 86, 93–99] rather than mounted close to the eye regions [85, 100, 101], its wide field-of-view accommodates larger head movement that is advantageous within the context of pervasive eye-gaze tracking, in comparison with the narrower field-of-view that is typically associated with commercial eye-gaze trackers. Nonetheless, this reduces the perceived resolution of the eye region images, which calls for methods that address the challenge of estimating the eye-gaze from low-quality images.

Appearance-based methods have been receiving increasing attention in this regard, since these methods exploit the photometric properties of the eye that are preserved at low-resolution, rather than the geometric details that may be difficult to extract. A subset of the methods that address the problem of gaze estimation from low-resolution images aim to exploit the eye region appearance at the detection stage, in order to extract the pupil [81, 93] or iris [94–96] centre coordinates as a prior step to gaze estimation. In this regard, Ince and Kim [93] and Janko and Hajder [94] propose to estimate the pupil [93] or iris [94] centre coordinates by employing a circular template to

capture the gradient change across the boundary of the eye region under consideration. Wojciechowski and Fornalczyk [95], on the other hand, localise the iris region by performing adaptive thresholding on the eye region image. Kim et al. [81] extract the pupil centre coordinates within an image frame in two successive stages, by initially extracting an approximate location of the iris region due to its larger footprint and subsequently using this estimate to detect the pupil position via a machine learning approach. Their method [81] first localises the iris region by summing across the image rows and columns in order to compute the respective horizontal and vertical intensity projections that delineate the change in intensity between the iris and surrounding sclera. This gives a coarse indication of the iris centre coordinates, which are subsequently improved by a neural network trained on pairs of iris and pupil centre coordinates. Lu et al. [96] address the problem of iris localisation from low-resolution eye images, where a reliable iris contour is difficult to extract. To this end, they synthesise virtual iris appearances for possible gaze directions using a simple 3-dimensional model of the eyeball, and subsequently optimise over this synthetic space to find the solution that best fits the observed eye image and iris appearance. The resulting output of this method is a 3-dimensional gaze estimate. Within the context of pervasive eye-gaze tracking, a common downside to these methods is their requirement for suitable image contrast at the boundaries that separate different eye region components, such as the iris, pupil and sclera, a condition which may not be necessarily fulfilled under pervasive conditions. In comparison, Cristina and Camilleri [97, 98] propose to extract the iris region by means of a Bayes' classifier trained to classify between the intensity values of iris and non-iris pixels based on labelled samples belonging to the two classes. This results in a binary image that permits the estimation of the iris centre coordinates as the centre of mass of the segmented iris region, followed by image-to-screen mapping in order to compute the screen coordinates of the PoR. While good image contrast is not a strict requirement for this method to operate, its performance is susceptible to illumination variations that alter the intensity values of the iris and non-iris pixels. The methods of Cristina and Camilleri [97, 98] alleviate this susceptibility to illumination variations by populating the training data set with iris and non-iris samples acquired under different illumination conditions.

The aforementioned methods [81, 93–95, 97, 98] use the information extracted at the detection stage to compute a point-of-regard on a monitor screen via an image-to-screen mapping function. Hence, the gaze estima-

tion performance of these approaches [81, 93–95, 97, 98] is highly contingent upon the accuracy of detection of single feature coordinates, namely the iris or pupil centre coordinates. Alternatively, appearance-based methods that map high-dimensional [38, 40, 82, 84, 99, 102] or low-dimensional [37, 86, 103] image information directly to gaze parameters, utilise the entire image content for gaze estimation irrespective of the resolution and hence do not require the explicit identification of small scale features. In this regard, Mansanet et al. [86] compare the use of k-Nearest Neighbour regression, support vector regression and random forest regression to map a set of principal components extracted from an eye region image, to a PoR on a monitor screen. Wojke et al. [103] employ Gaussian process regression to map an observed eye patch to a PoR on a monitor screen. A mapping relationship between screen coordinates and a low-dimensional manifold of eye patches, generated by applying a Gaussian process latent variable model to the image data, is established following a calibration phase during which sets of on-screen gaze points and corresponding eye patches are recorded. The method of Huynh [37] seeks to obtain a discrete classification of the gaze direction by feeding a neural network with a set of principal components extracted from eye region images, based on the availability of a training data set. George and Routray [102] seek a discrete classification of the eye-gaze direction as well, by feeding a convolutional neural network with low-resolution eye images at the input stage, following a training step on a publicly available image data set. Similarly, Holland and Komogortsev [84] and Sewell and Komogortsev [82] feed a neural network with a vector of raw pixel intensities from an input eye region image to estimate the PoR. The methods of Wojke et al. [103], Holland and Komogortsev [84], and Sewell and Komogortsev [82] necessitated the collection of a training data set via a calibration procedure that required the user to gaze at visual markers appearing in sequence on a display screen.

The aforementioned methods for gaze estimation [37, 82, 84, 86, 102, 103] do not compensate for changes in eye appearance that arise due to head movement, hence constraining the user to a stationary head pose [37, 82, 84, 86, 102, 103] that is often supported by the use of a chin-rest [86]. This constraint is incompatible in the context of pervasive eye-gaze tracking since it greatly limits the natural movement of the user that is otherwise characteristic of real-life scenarios, as will be further discussed in Section 5.4. To counteract this limitation, Koutras and Maragos [99] and Dhyawala et al. [40] compensate for head movement by including information related to the head pose in the estimation of gaze, where Koutras and Maragos [99] obtain this infor-

13

mation by fitting an active-appearance model to the face, while Dhyawala et al. [40] infer the head pose from the relative image distances between salient facial features. These methods [40, 99] also require the availability of a training data set in order to train several Gaussian mixture models [99] or a neural network [40] to estimate the PoR. Nguyen et al. [38] employ Gaussian process regression to estimate a PoR from input data consisting of the pixel intensities of the eye region image. They propose to handle head movement by employing two separate calibration procedures; the first procedure collects pairs of eye images and corresponding screen coordinates while the user maintains a stationary head pose, whereas the second procedure requires the user to perform head rotations in four different directions in order to capture the change in eye region appearance under a changing head pose.

In order to estimate the gaze from image information, most of the appearance-based methods that have been discussed within this section require the collection of a large corpus of training data, especially if the method aims to handle head movement by collecting a training data set that captures the changes in eye region appearance under a changing head pose, as in [38, 40, 99]. For instance, Nguyen et al. [38] reportedly perform ten calibration sessions in order to collect sufficient training data that permit a suitable reduction in the gaze estimation error under head movement. The collection of such a large corpus of data requires extensive user cooperation, which is highly limiting in a pervasive scenario where prolonged user cooperation cannot be guaranteed. Techniques to reduce the size of the training data sets typically required by such appearance-based methods and, hence, improve the suitability of these methods for pervasive eye-gaze tracking, will be further discussed in Section 5.2.

5.2. Challenge B: Gaze Estimation from Sparse, Synthesised or Person-Independent Training Samples

Within the context of pervasive eye-gaze tracking, appearance-based methods have been receiving increasing attention since these potentially permit the estimation of gaze from low-resolution eye region images without requiring explicit identification of small scale features. Nonetheless, the computation of the mapping function necessitates the collection of large data sets comprising hundreds [38, 57, 86] or thousands [40, 43, 56, 61, 83, 85] of training samples, which often requires considerable user cooperation prior to gaze estimation. Furthermore, the collected training samples are often person-specific and, hence, the computed mapping function may not suitably

14

compensate for appearance variations across a cohort of users [87, 104, 105]. The following sections discuss recent research work that specifically aims to address these issues.

5.2.1. *Gaze Estimation from Sparse Data*

The availability of large sets of training data typically allows for an improvement in the gaze estimation accuracy [45] at the expense of lengthy data collection that may be taxing on the user. Recent methods that specifically address this problem, seek to learn a mapping function from sparsely collected training samples without compromising the gaze estimation accuracy [87, 104, 105].

In this regard, Lu et al. [87, 104] propose an adaptive linear regression technique (ALR) that optimally selects a small number of training samples for gaze estimation. Since the training samples are sparsely collected, local linearity of the eye appearance manifold cannot be assumed, as would be the case if dense training samples had otherwise been collected. Alternatively, Lu et al. [87, 104] propose to estimate the gaze information for a newly observed eye image from an optimal subset of the training data that is closely correlated to the observed eye image and can interpolate it linearly. Liang et al. [105] similarly seek to address the problem of gaze estimation from a limited number of training samples. To this end, they propose a method to increase the number of available training samples by employing semi-supervised learning on a set of labelled data, which consists of pairs of eye appearance vectors and the corresponding gaze coordinates, and unlabelled data, which comprises the eye appearance vectors alone. The objective of the semi-supervised learning, based on Gaussian process regression, is to infer the gaze coordinates of the unlabelled data from an active set of limited labelled samples. This bypasses the need to collect further training data by increasing the number of calibration points displayed to the user, which procedure may be taxing and time consuming. These methods [87, 104, 105] all reportedly achieved mean gaze estimation accuracies below 1° of visual angle, using sparse training data collected during calibration procedures comprising no more than 33 points [87, 104] and as few as 5 points [105]. It is, however, important to underline that the method of Lu et al. [104] does not consider the variation of eye image appearance under head movement instead focusing on a fixed and frontal head scenario, which limits the applicability of this method in the context of pervasive eye-gaze tracking. The methods of [87, 105] handle slight head movement only, where Lu et al. [87] compensate

for the resulting image distortion via a rigid transformation consisting of translational and rotational components. The challenges associated with head pose-free gaze estimation will be further discussed in Section 5.4.

5.2.2. *Gaze Estimation from Synthesised Training Data*

Rather than opting to collect a sparse data set in order to lessen the data collection effort [87, 104, 105], as discussed in the previous section, the synthesis of photo-realistic training data has been alternatively proposed in order to allow the creation of a large body of training data, bypassing the need for data collection [106, 107].

The methods of Wood et al. [106, 107] propose an anatomically inspired eye region model that can represent variations in facial shape and appearance, and which may be rendered at different eye and head rotation angles and under variable lighting to produce a training data set that covers a wide range of tracking conditions. This morphable shape model is built upon high resolution 3-dimensional head scans captured by professional photogrammetry from a cohort of participants [106, 107]. To improve realism, the model proposed by Wood et al. [106, 107] includes elements such as the eyelashes and eye wetness, and animates the eyelid to simulate its motion during eye saccades. In order to quantify the ability of the proposed model in estimating the eye-gaze, Wood et al. [107] match the rendered eye images to those contained within a benchmark data set comprising real image data, using a k-Nearest-Neighbour technique trained on the rendered images. They report a comparable performance to state-of-the-art appearance-based methods employing deep learning algorithms, which were trained on synthesised or real image data and tested on the same benchmark data set.

The shape and appearance variability of the training data synthesised by the model of Wood et al. [106, 107] is reportedly contingent upon the number of captured 3-dimensional head scans. In their work, this number is increased from ten [106] to twenty [107] in order to include more of such variations, nonetheless the need for further head scans, which would capture additional shape and appearance variations, has been reported necessary [107]. This leads on to the challenge of person-independent gaze estimation in the following section.

5.2.3. *Person-Independent Gaze Estimation*

Another challenging aspect of appearance-based gaze estimation relates to variation in the facial characteristics of different users, which often affects the

gaze estimation accuracy. Indeed, an investigation carried out by Zhang et al. [85] revealed a drop in the gaze estimation accuracy for an unseen subject whose samples were not contained within the training data set in use, where the data was characterised by large variation of eye appearance from persons of different gender, age and ethnicity. Hence, data collection for training and testing is often carried out for every individual user [87, 104, 105], which is often limiting if this data is not available or a data collection session cannot be performed. Head movement offers a further challenge in this regard by introducing further changes in facial appearance under different head poses.

Methods that aim to address the challenge of person-independent gaze estimation generally extract information from an existing data set of training samples and exploit this information to compensate for the change in appearance of an unseen user that is not included within the data set [108–110]. In this regard, Schneider et al. [108] propose a method for submanifold reconstruction of the low-dimensional features and corresponding gaze labels for an unseen user, via a linear interpolation of neighbouring data points on the same manifold. Their method does not address the challenges associated with head movement, alternatively focusing on frontal eye images alone aligned according to the relative positions of the inner and outer eye corners. Lu and Chen [109] aim to extract from the eye image, eye features that are less person-dependent in order to learn a person-independent relationship between changes in gaze and eye appearance variation. To this end, they partition eye images into smaller patches of different sizes and subsequently employ a sparse auto-encoder to learn a codebook. This codebook carries essential structural information that can reconstruct the original image patches from coefficient vectors computed for every image patch. Gaze prediction is then carried out by employing a multi-class support vector machine trained on sets of eye features, produced by pooling the coefficients at different image scales, and the corresponding gaze directions. Similar to Schneider et al. [108], Lu and Chen [109] do not handle head movement and are restricted to frontal eye images alone. Mora and Odobez [110] propose to reconstruct a newly observed eye image from a linear combination of a subset of training samples, selected according to their relevance to the test image based on the computation of a weighting value. Any variation to the eye appearance due to head movement is handled beforehand by re-rendering the texture of an RGB-D frame to a frontal head pose, according to head pose parameters computed by fitting a 3-dimensional face model to depth information acquired by a Kinect sensor. Hence, the proposed appearance-based method

17

estimates the eye rotation from frontal eye images, which is then combined with the estimated head pose as a final step to compute the gaze direction.

Mora and Odobez [110] compare the results obtained by the proposed method, under stationary and free head pose conditions, to the estimation of gaze by specific training and testing for an individual user alone, and by cross-testing of an unseen user that is not included within the data set. Specific user training and testing is reported to produce the highest gaze estimation accuracy whereas the use of cross-estimation leads to the least accurate results, in agreement with the results of Zhang et al. [85] as previously reported. Furthermore, the results report a degradation in the gaze estimation accuracy under a changing head pose, indicating the need for robust techniques that address the challenges of appearance-based methods under less constrained conditions.

5.3. Challenge C: Reduced or Implicit Calibration

One of the limitations associated with eye-gaze tracking that is often considered to hamper the use of the technology, especially in pervasive scenarios such as public spaces, is the need to perform calibration prior to gaze estimation. Several types of calibration procedures have been discussed in Section 3, where it has been seen that different calibration procedures typically require active user participation in order to estimate parameter values that permit information from the image domain to be converted to the gaze domain. Nonetheless, user participation cannot necessarily be guaranteed especially if the users are infants, such as in studies related to autism [111], or exhibit involuntary movements as by-product of a disability [112], or if the users simply regard the process of calibration as a tedious task [113]. Furthermore, *setup calibration* typically imposes a static setup geometry that cannot be re-configured to accommodate personal needs during tracking. In light of these constraints, recent research work on eye-gaze tracking has been directly aimed at the reduction of calibration, as will be further discussed in the following sections.

5.3.1. *Reduction of Point-of-Regard Calibration*

One approach towards the reduction of calibration effort performs implicit calibration while the user is induced to execute a specific action, such as following a moving target [114] or gazing at a static visual stimulus [115], or while the user is interacting with a personal computer through mouse

clicks [116–119] and keyboard presses [118, 119]. Such user activities permit the unobtrusive collection of data in the background without the user being explicitly aware of the calibration task. Pfeuffer et al. [114] designed three different applications to investigate different capabilities of their implicit calibration approach for gaze mapping, namely the smooth blend between the calibration task and the application, the unawareness of the user during calibration and the capability of resuming calibration at the occurrence of distractions. The proposed method detects when the user attends to the moving target by computing a Pearson's product-moment correlation measure between moving windows that aggregate the target coordinates and uncalibrated gaze coordinates over time. The sampled coordinate-pairs are then fitted to a homography model that represents a perspective projection between the gaze estimates and the screen plane. Pfeuffer et al. [114] report short calibration times for most of the participants, nonetheless the display area covered by the calibration target together with its size and speed were reported to affect the calibration accuracy. Alnajar et al. [115] exploit the fact that different individuals tend to produce similar gaze patterns when looking at the same visual stimulus. Hence, they present different users with the same image to capture an initial set of uncalibrated gaze patterns and subsequently match these patterns across users to correct their image space positions. Two methods have been proposed for the computation of the initial gaze patterns; the first of which computes the gaze points corresponding to newly captured eye region images from pairs of eye region templates and gaze information acquired from a template subject, while the second method follows the first one, however the template and newly captured eye region images are projected to a lower-dimensional space. Matching between gaze patterns is subsequently carried out by a closest point search algorithm that applies different translation and scaling transformations, hence permitting the computation of a mapping transformation across different users. Head movements have not been handled at this stage, nonetheless these reportedly affect the gaze mapping accuracy by altering the eye region appearance upon which the proposed method depends.

Methods that exploit user interaction often assume that the PoR coincides with the on-screen mouse cursor position when the user clicks the mouse button [116, 117]. Papoutsaki et al. [116] collect pairs of training data that consist of specific eye features together with corresponding on-screen click coordinates, while users navigate the web unaware of the calibration process. The collected data is then used to train a linear regression model to map the

19

eye features to screen coordinates. Similarly, Sugano et al. [117] assume that clicked screen positions coincide with a PoR, and create training samples using appearance features of the eye and head together with the corresponding screen coordinates. In order to handle head movement, training samples with similar head poses are clustered and a local manifold for each cluster is created. The output gaze for a newly observed eye image and the associated head pose is then estimated as a weighted average of candidate gaze predictions from multiple clusters. Huang et al. [118, 119], in contrast, argue that the moment of best alignment between the user activity and the corresponding PoR varies according to the type and context of the interaction, and the individuality of different users. They investigate the relationship between gaze patterns and interaction cues across a cohort of users. Based on this investigation, Huang et al. [118, 119] propose data validation techniques that utilise head pose and eye features to identify instances when the PoR and an interaction event are most likely to be aligned. These head pose and eye features together with the corresponding on-screen gaze point are then used to train a random forest regression model. These methods typically report effective use of the relationship between the eye-gaze and the interactive behaviour of the user. Nonetheless, they are designed for a desktop scenario where the user actively interacts with on-screen content through the use of peripheral devices, which may limit the applicability of such methods for pervasive applications.

Another approach exploits the visual saliency of video frames in order to reduce or eliminate the necessity for active PoR calibration [120, 121]. For instance, Valenti et al. [121] reduce the required calibration effort without jeopardising the gaze estimation accuracy, by exploiting visual saliency to re-calibrate during runtime without necessitating user intervention. Visual saliency is extracted by the computation of displacement vectors from isophote curvatures that indicate the positions of interesting or structured objects inside the image space. Following the estimation of a gaze point on the image, a meanshift algorithm iterates towards the closest saliency peak, which subsequently determines the adjusted gaze estimate and provides information regarding the calibration error. The proposed method compensates for affine transformations of the gaze estimates and potentially permits changes in head pose, which is desirable in a pervasive setting. Sugano et al. [120] generate saliency maps for every input image frame by extracting low-level saliency features, such as colour, intensity, orientations, flicker and motion, and higher level saliency features of the human face. Following the

20

collection of a training data set under stationary head conditions, consisting of synchronised pairs of saliency maps and eye images, gaze probability maps are computed from sets of saliency maps clustered according to the similarity of corresponding eye images. The mapping relationship between eye images and gaze points is established by Gaussian process regression based on the training data set, permitting the computation of gaze for a newly acquired eye image. The method of Sugano et al. [120] eliminates the need for calibration at the expense of requiring the collection of a training data set, which task may be difficult to achieve under pervasive conditions.

Pervasive scenarios in which user interaction happens spontaneously and varies in duration often do not allow for implicit or explicit person-dependent calibration. Interaction with public displays fits this scenario, where different users that walk up to the display individually or simultaneously expect to be able to navigate the displayed content immediately. Within this context, Zhang et al. [10, 11] eliminate the need for gaze mapping calibration entirely by proposing a relative rather than absolute mapping of the eye movements onto the display screen. Conventionally, gaze mapping methods are absolute since these estimate a point on the screen for every gaze direction directly [22, 25, 27, 37, 38, 40, 43, 45, 56, 57, 61, 62, 64, 78–86, 92]. In comparison, Zhang et al. [10, 11] propose to relate the movements of the eyes to the scrolling direction of information presented on the screen, bypassing the computation of a PoR. Changes in eye movements are identified by the computation of the distance between the pupil centres and the inner eye corners, which they term as pupil-canthi ratio in previous work [122], under the assumption that the head is held stationary and in a frontal and upright position. A change in eye movements away from the central position of the display triggers the displayed content to scroll towards the centre, coming to a halt at the display centre if the user is detected to follow the scrolling movement. In absence of head pose estimation, several participants were reported to find the stationary head constraint difficult to maintain [10, 11], which underlines another important challenge associated with pervasive eye-gaze tracking, that is to permit natural head movement as will be discussed in Section 5.4. Indeed, Zhang et al. [10, 11] reported that head rotations were a main cause of scrolling failure, causing the changes in the distance between the pupil centres and inner eye corners due to head movement to be misinterpreted by the proposed method.

In absence of any information on visual saliency or user interaction cues, Lu et al. [123] propose a method to recover patterns of gaze movement, up

to a scaling and translation ambiguity, without necessitating calibration or training data. As Lu et al. had previously shown [87], eye images captured under a stationary head pose constitute an appearance manifold that may be approximated by a 2-dimensional surface, with a similar 2-dimensional structure that comprises the corresponding gaze positions. If the gaze position for a newly observed eye image is recovered by Euclidean measurements, this appearance-gaze relationship will hold true within local surface regions where the similarity between eye images is high. This would require the availability of a large set of calibration samples to permit the recovery of gaze inside each local region. In order to address this problem, Lu et al. [123] propose to extend this local linearity by introducing the Geodesic distance, which measures the shortest path between two nodes in a graph by adding up small Euclidean distances of neighbouring nodes. Lu et al. [123] report that this Geodesic distance maintains a global linear relationship between the eye appearance and the corresponding gaze movement, and exploit this relationship to recover an uncalibrated pattern of gaze movements. This uncalibrated gaze pattern may be useful for pervasive applications that do not require an absolute mapping of the eye movements onto a physical scene object, such as for the previously mentioned interaction with public displays. If an absolute mapping is required, this may be achieved through a calibration procedure that requires the user to gaze at calibration markers to which the gaze pattern is subsequently aligned.

5.3.2. *Reduction of User/Personal Calibration*

Calibration or user-related parameters is typically required by methods that project a 3-dimensional gaze vector from the eyeball centre to estimate the gaze direction [35, 43, 47, 49, 64, 69–77, 88, 91]. These methods generally define a 3-dimensional geometric model of the eyeball and perform calibration to estimate subject-dependent parameter values, such as the eyeball centre coordinates and radius.

To reduce the complexity of the eyeball model and hence the number of parameter values to be estimated, the visual axis of the eyeball is often assumed to coincide with the optical axis, eliminating the need to calculate the position of the fovea and the angle between these two axes [35, 47, 49, 64, 69, 70, 74–77]. Nonetheless, active user participation is generally still required to estimate the remaining model parameters, which is often undesirable in a pervasive scenario. To this end, the methods of Kohlbecher et al. [124] and Klefenz et al. [125] seek to eliminate the need for personal calibration

by reconstructing the pupil region in 3-dimensional space via stereo-vision, and calculating the gaze direction from the pupil normal. Kohlbecher et al. [124] reconstruct the pupil region by projecting two conics through the camera centres and the pupil images, defining the 3-dimensional pupil ellipse by their intersection in space. Similarly, Klefenz et al. [125] project the pupil ellipse and centre coordinates extracted inside the image into 3-dimensional space by stereo-vision, denoting the gaze vector by the pupil normal. The downside of these methods in the context of pervasive eye-gaze tracking is their general requirement for eye images of good resolution and contrast, which permit accurate extraction of the pupil boundary for ellipse-fitting. In this regard, Kohlbecher et al. [124] report that in the absence of well-defined ellipse parameters due to small pupil projections, the gaze vector estimates exhibited considerable variability.

A further downside to the use of stereo-vision is the need for multiple cameras and additional camera calibration, which reduce the portability of the setup in a pervasive setting and limit the immediate use of integrated imaging hardware on off-the-shelf mobile devices. In the absence of any redundant information acquired by multiple cameras, the complexity of estimating the 3-dimensional parameter values from a single camera, without resorting to personal calibration, increases. Urano et al. [12] propose a method based on a single camera without the need for calibration, aimed for an objective evaluation of adverts situated in public spaces. Parameter values that personalise the 3-dimensional eyeball model but cannot be directly measured from image information, such as the eyeball radius, are estimated based on the ratio between the image space interpupillary distance in pixels and the corresponding anthropometric measure in millimetres. This allows the anthropometric measure of the eyeball radius to be converted to pixels, hence permitting estimation of the 3-dimensional eyeball centre coordinates. The gaze direction is defined by the 3-dimensional vector that joins the eyeball centre with the iris centre, previously extracted by an ellipse-fitting technique. The method proposed by Urano et al. [12] does not compensate for head movements, which is a serious limitation in public environments where individuals are expected to move naturally. In comparison, Cristina and Camilleri [126] permit free head movement by proposing a method based upon the definition of a spherical eye-in-head rotation model that permits gaze estimation by compensating for the change in eye region appearance due to head rotation. The proposed method requires minimal user cooperation by holding an upright and frontal head and eye pose during a brief initialisation stage. Similarly

to Urano et al. [12], the estimation of parameter values that personalise the rotation model is carried out by relating image space measurements in pixels to the corresponding anthropometric measures in millimetres. This permits the estimation of a 3-dimensional vector that joins the eyeball and iris centre coordinates, and hence the estimation of gaze direction as a combination of head and eyeball rotation components.

While the 3-dimensional eyeball model may be simplified by defining the optical axis as the line of gaze, the true gaze direction is defined by the visual axis, which differs from the optical axis by an angular offset. Computation of this offset should allow for improved gaze estimation, hence several methods include the calculation of the angle between the optical and visual axes during the calibration procedure [71–73]. Methods that are better suited for pervasive environments seek to estimate this angle implicitly without requiring explicit user awareness of the underlying calibration [127–130]. Nagamatsu et al. [127] propose to estimate the yaw and pitch angular components of the left and right eyeballs while the user gazes at a minimum of four directions into the distance. Their calculations are based upon the availability of the optical axes of both eyes, which they estimate beforehand via an active eye-gaze tracking technique, under the assumption that the visual axes are parallel as the user gazes at distant visual stimuli. The proposed method is mostly suitable for driving situations that require the user to gaze into the distance for most of the time. If the visual stimuli are alternatively situated closer to the user, Model and Eizenman [128] exploit the constraint that the visual axes should instead intersect on the observed surface at the PoR. Hence, the yaw and pitch angular components of both eyes are estimated by minimising the distance between the intersections of the left and right visual axes with the surface of interest for several gaze points collected as the user gazes at the display. Model and Eizenman [128] report that the proposed solution becomes less sensitive to noise for increasing angles between the visual axes and surface normal, hence requiring either a larger display that provides a wider range of horizontal and vertical viewing angles or an observation surface that consists of multiple planes. In a pervasive context, both conditions limit the applicability of the proposed method on small surfaces, such as the display screens of mobile devices. The methods of Chen and Ji [129, 130] do not require the user to perform specific actions other than the natural interaction with the system, in order to calculate and gradually improve the angular offset and gaze estimates. Chen and Ji [129, 130] exploit visual saliency to approximate the gaze probability distribution for static images displayed to

24

the user. Eye parameter estimation is solved as an inference problem in a Bayesian network, which estimates the posterior probability of the angular offset given the optical axis and the displayed image, and the computed gaze probability distribution. Gaze estimation is also based on the same Bayesian network model, where the gaze posterior probability is computed through the prior probability of gaze and the gaze likelihood, which may be derived from the angular offset probability. Training data is required to estimate the angular offset probability beforehand, which data is collected by requesting the users to gaze at a set of images for a specified period of time. In order to eliminate the need for data collection, Chen and Ji [129, 130] propose an incremental learning algorithm that extends the Bayesian network to a dynamic model and recursively updates the angular offset probability, which is more suitable for a pervasive scenario.

5.3.3. *Reduction of Setup Calibration*

The estimation of a 2-dimensional PoR on elements of the surrounding scene from 3-dimensional gaze information is often carried out by intersecting the computed 3-dimensional gaze vector with objects inside the world space [70, 72, 75]. This typically needs the inclusion of a further setup calibration stage that relates the position and orientation of the camera with other units in the setup, such as a monitor screen [70, 72, 75]. This calibration procedure is performed prior to use and imposes a static setup geometry, hence constraining the flexibility of re-configuring the setup during use and requiring re-calibration if the setup configuration is changed.

Nitschke et al. [131] work around the need for additional imaging hardware and calibration by proposing a method to estimate the PoR directly from corneal surface images of the reflected scene. Similar to the methods in Section 5.3.2, the method of Nitschke et al. [131] defines a 3-dimensional eyeball model, characterised by anthropometric measures of the human eye, that permits the estimation of gaze direction. The gaze-reflection point is defined by the model as the intersection point between the incoming light ray, parallel to the optical axis, and the corneal surface. The image coordinates of the gaze-reflection point under weak perspective projection are then computed from the known gaze direction, and the angle which the vector that joins the gaze-reflection point to the corneal centre makes with the optical axis. Nitschke et al. [131] additionally introduce a one-point calibration method to further improve the PoR accuracy by compensating for the angular offset between the optical and visual axes. Nonetheless, as discussed in

25

Section 5.3.2, the inclusion of user calibration is undesirable in a pervasive scenario and methods have been proposed to estimate this angle implicitly without active user participation [127–130]. Takemura et al. [132] follow a similar method for PoR estimation from corneal surface images, however they introduce an additional scene camera in order to obtain a higher resolution image of the focused object in the scene. In absence of any setup calibration between the eye and scene cameras, the method of Takemura et al. [132] identifies the PoR in the scene by matching an unwarped image of the corneal reflection to the scene image via normalised cross-correlation. The methods of Nitschke et al. [131] and Takemura et al. [132] require eye images of suitable resolution that permit the extraction of scene detail, hence these may be better suited for applications that allow short distance setups, such as wearable cameras.

5.4. Challenge D: Head Pose Invariant Gaze Estimation

The ability to infer the orientation of a human head within a global coordinate framework from image information, hereby referred to as head pose estimation, presents a challenging problem in itself that has generated significant interest within the computer vision community [133]. Nonetheless, supplementing gaze estimation methods based on the computation of eye rotation alone [22, 45, 49, 57, 75, 78, 79, 81–84] with information regarding the head pose, permits the estimation of gaze under different eye and head configurations. This is especially desirable in the context of pervasive eye-gaze tracking where the estimation of head pose lifts any restrictions on the natural head movement of the user, without resorting to stationary head pose constraints [22, 45, 49, 57, 75, 78, 79, 81–84], often by the use of a chin-rest [45, 82], or wearable setups that circumvent the need for head pose estimation but cannot track the eye-gaze remotely [43, 78, 85]. Indeed, as discussed in Section 5.3.1 in the context of user interaction with public displays, among others, the absence of any compensation for head movement constrains the user to an unnatural fixed head pose that is difficult to maintain, and often leads to tracking inaccuracies or failure if this condition is violated [10, 11].

5.4.1. *Normalisation of Observed Eye Images to a Frontal Pose*

One of the challenging aspects relating to the estimation of eye-gaze under different head orientations is the introduction of significant changes in eye image appearance that arise with a changing head pose. A subset of the methods that seek to address this problem propose to compensate for such

changes in appearance by normalising the eye images to a frontal view based on an inverse rigid transformation of pre-computed head pose parameters [134–138]. The gaze direction, based on the eyeball rotation alone, is then estimated from the pose-normalised image and subsequently transformed by the pre-computed head pose parameters to compute a combined gaze estimate.

Pose-normalisation allows for a reduction in the eye model complexity of shape-based methods that, otherwise, seek to capture the shape variations of different eye region components under a changing head pose [32, 34]. In this regard, normalisation of eye images to a frontal view permits the definition of a geometrical model that represents the shape of different components that constitute the open eye, such as the iris [26, 27, 136] and eyelids [137–139], as seen from a frontal perspective. A number of methods [26, 27, 136–139] first estimate the head pose parameters by fitting a face model to specific facial landmarks [26, 27, 136, 139] or a 3-dimensional morphable model to depth information acquired by a Kinect sensor [137, 138]. Under a frontal head pose assumption, Baek et al. [136] subsequently generate several elliptical iris templates by rotating a 3-dimensional eyeball model and computing its image space projection at different yaw and pitch angles. Following template matching to determine the best fitting template inside the pose-normalised images, the corresponding iris centre coordinates are transformed according to the head pose parameters and mapped to a PoR on a monitor screen. Mora and Odobez [137, 138] propose a method that infers the eye movement parameters by maximising an eye image likelihood measure, constrained by the eyeball geometry and frontal eyelid configuration, and subsequently transforms the pose-rectified eye movement estimate according to the head pose parameters. In comparison, Xiong et al. [139] include the eyelid contours within the 3-dimensional face model by which they estimate the head pose. Following pose-normalisation of the eyelid contours and pupil centre coordinates according to the estimated head rotation angles, the gaze information is inferred from the relative distances between the pupil centre and the upper and lower eyelids.

Methods that exploit the photometric appearance of the eye and necessitate the collection of training data for gaze estimation also benefit from pose-normalisation since this permits the collection of training samples under a frontal head pose alone [134, 135, 140], hence alleviating the need for a large corpus of training data that captures the eye appearance under different head orientations [38, 40, 44]. In this regard, the methods of [134, 135, 140]

first estimate the head pose parameters by fitting a 3-dimensional morphable model to specific facial landmarks [135, 140] or depth information acquired by a Kinect sensor [134]. By compensating for changes in head pose, the method proposed by Egger et al. [135] permits the use of a single random forest regressor trained on frontal samples in order to estimate the gaze information from pose-normalised eye images. Similarly, Mora and Odobez [134] train an adaptive linear regression (ALR) algorithm on pairs of frontal eye images and corresponding gaze information, as previously described in Section 5.2.1, permitting the estimation of gaze based on eyeball rotation within a frontal head reference frame. The gaze direction under a changing head pose may be subsequently corrected by a rotation transformation according to the estimated head pose parameters [134]. Jeni and Cohn [140] also train a linear support vector regressor on binary features extracted around specific points on the eye contours and pupils, after a frontal view of the face had been synthesised according to the estimated head pose.

Alternatively, pose-normalisation of the eye images to a frontal view allows the estimation of eye movements via a direct comparison of the image space position of a specific feature, such as the iris, with respect to the corresponding feature position within a frontal reference image [141, 142]. Asteriadis et al. [141, 142] propose to warp the eye region images around a cylindrical shape and perform an inverse transformation of the head pose parameters, computed via an adaptive fusion of feature and appearance-based image information within a Bayesian framework. The proposed methods seek to infer the direction of eye movement with respect to a reference frontal pose and, hence, simplify the iris movement to a linear displacement inside the image space. The direction of eye movement is finally identified by comparing the image distance between the pose-normalised and reference frontal iris centre coordinates to pre-defined threshold values.

Most of the aforementioned eye-gaze tracking methods employ a head pose estimation approach that necessitates the identification and tracking of *specific* facial landmarks [26, 27, 135, 136, 141, 142]. This requires that the range of possible head rotation angles is constrained such that the entire set of facial landmarks remains visible during tracking, a condition which may limit the usability of these methods in wide spaces that often necessitate large head rotations. Facial landmarks which become occluded due to large head rotations may lead to further inaccuracies in the estimation of head pose, which may in turn increase the gaze estimation error since the latter is now contingent upon the accuracy of both the head and eye pose estimates

combined. Hence, under free head pose conditions, the reliability of the head pose estimates is an important factor to consider as one of the components that contributes to the estimation of gaze, since this may introduce a new source of error.

5.4.2. *Compensation for Head Rotation*

Rather than compensating for head movement by normalising the eye image appearance to a frontal pose prior to gaze estimation, a second subset of methods extract information relating to the eye-gaze from the eye region under the appearance changes directly [126, 143–146]. This information is then combined with head pose information in order to compute a joint gaze estimate. Similarly to the methods in Section 5.4.1, different approaches seek to exploit the photometric or geometric properties, or the local features of the eye regions in order to estimate the gaze direction under a changing head pose.

One approach towards the estimation of the eyeball rotation angles under a changing head pose follows the definition of a 3-dimensional geometric model of the eyeball [126, 143, 144]. For instance, Xiong et al. [143] first estimate the eyeball rotation angles by projecting the image space pupil centre coordinates to a 3-dimensional eyeball model, in the knowledge of the intrinsic camera parameters. The head pose parameters are subsequently applied on the eyeball model in order to rotate its optical axis, hence permitting the estimation of a gaze vector that combines the eye and head pose information together. Utsumi et al. [144] propose to estimate a gaze vector by minimising the image distance between the observed iris centre coordinates and the image re-projection of a 3-dimensional eyeball model under different rotation angles. Their method also requires knowledge regarding the intrinsic and extrinsic camera parameters. Wood et al. [147] make use of high-resolution face scans in order to construct a multi-part model of the eye region, which includes an anatomy-based model of the eyeball. The gaze information is estimated by iteratively synthesising an image of the multi-part model under different eyeball and head rotations, and identifying the best match to an observed eye region image by minimising an energy function. Cristina and Camilleri [126] also define a 3-dimensional eyeball model, onto which they project the image space displacement between the iris centre coordinates in a newly observed eye region image and a reference frontal image rotated on a cylindrical head rotation model by the head pose parameters. This permits the computation of a gaze vector that projects from the eyeball model

by joint eye and head pose estimation. The methods of Xiong et al. [143], Utsumi et al. [144], Wood et al. [147], and Cristina and Camilleri [126] require the identification of specific facial landmarks in order to estimate the head [126, 143, 144, 147] and eye [147] poses via model-fitting techniques. Occlusion of these features of interest may lead to increased head or eye pose estimation error as previously discussed in Section 5.4.1, and further corroborated by the experimental results of Cristina and Camilleri [126], which reports a reduction in the gaze estimation accuracy with the introduction of head pose estimation in comparison to the use of ground truth head pose data.

The problem of relying on specific facial landmarks for head pose estimation may be alleviated by appearance-based methods that exploit the photometric image information as high or low-dimensional input to a machine learning algorithm. Lu et al. [145] propose a method in this regard that initially estimates the eye-gaze from a newly observed eye region image via regression, based on the availability of a training data set collected during calibration under stationary head conditions. The initial gaze estimate is subsequently corrected to compensate for the estimation bias that arises due to a changing head pose and the resulting eye appearance distortion. This compensation process requires further calibration, which necessitates the user to hold a fixed gaze and perform head rotations, in order to collect the necessary training samples to perform Gaussian process regression. The prolonged user cooperation required by the method of [145] during calibration may, however, limit its applicability for pervasive eye-gaze tracking. In comparison, Sugano et al. [148] handle head movement without requiring extensive calibration, by synthesising the appearance of the eye regions under different head rotations from frontal images collected under a stationary head pose. The resulting dense data set is subsequently used to cross-train a random forest-based algorithm to estimate the gaze from newly observed eye images captured under head movement. Lai et al. [149] employ a random-forest based technique as well, trained on image data collected by requesting the user to hold a fixed gaze and perform different head rotations, in order to select neighbours on a low-dimensional manifold given the observed head pose and eye appearance features. The selected neighbouring features under similar eye and head rotation angles are then used as input to an ALR technique, as previously described in Section 5.2.1, in order to estimate the gaze direction. Liu et al. [150] estimate the gaze point of newly observed eye images, from a linear combination of gaze points corresponding to neigh-

bouring appearance vectors on an appearance manifold. If the left and right eye images are considered separately under natural head movement, the dissimilarity in eye appearance due to a rotated head pose might lead to an erroneous gaze estimation [150]. In order to address this problem, Liu et al. [150] unify the left and right appearance manifolds by concatenating the two appearance vectors, hence implicitly modelling the relation of both eye images in the same linear combination of gaze points.

In a different appearance-based approach, Cai and Lin [151] exploit the head pose information in order to extend the limited field-of-view of commercial eye-gaze trackers for applications that require a wider field-of-view, such as in virtual driving simulations. Rather than opting for joint eye and head pose estimation, Cai and Lin [151] employ a commercial tracker to estimate the gaze within a limited range of near-frontal head poses and switch to gaze estimation by the head pose alone outside this narrow range. An estimate of the head pose is computed as the mean value of multiple neural network outputs, each of which is trained on head images collected for every user during calibration. Similar to other appearance-based methods that employ machine learning techniques, the methods of Lu et al. [145], Lai et al. [149], Liu et al. [150] and Cai and Lin [151] necessitate the collection of training samples under the expected tracking conditions, which in this case include the presence of head movement. Hence, while appearance-based methods may potentially alleviate the issues associated with model-fitting techniques for head pose estimation, the need to perform calibration for data collection purposes may limit their suitability in pervasive settings where active user cooperation cannot be guaranteed.

Alternatively, the eyeball rotation angles may be estimated from local features around the eye regions. Weidenbacher et al. [146] propose to initially infer the gaze direction based on the eye movements alone, by computing the phase change of Gabor filter responses within sub-regions around the eyes. Subsequently, the same method is applied to estimate the head pose from filter responses at ten specific facial landmarks, which method requires the availability of prototype sets of face landmarks at known head poses together with the corresponding Gabor responses. The performance of this method was compared to a model-fitting technique for head pose estimation and both methods were reported to perform comparably. A combined gaze estimate is finally computed by mapping the initial gaze direction, computed according to the eye movements alone, and the head pose estimate on a lookup table that needs to be generated in advance.

The aforementioned methods within this section face similar issues to the methods discussed in Section 5.4.1, which mainly relate to a degradation in the gaze estimation accuracy with the introduction of head pose compensation. For instance, Cristina and Camillieri [126] report a degradation of approximately 0.5°-2° in the gaze estimation accuracy with the introduction of head pose estimation via a model-fitting technique, over the use of ground truth head pose information. One of the main sources of error in this regard potentially stems from inaccuracies in the detection and tracking of specific facial landmarks, which may become increasingly distorted or self occluded with increasing head rotation angles and subsequently lead to an increase in head pose estimation error [126]. Nonetheless, the capability to handle head movement at different head rotation angles is a desirable aspect of pervasive eye-gaze tracking that serves to lift any constraints on the natural movement of the user. Hence, in tandem with the challenges of pervasive eye-gaze tracking, these shortcomings call for an improvement in the achievable head pose estimation accuracy, especially at large head rotation angles typically expected within pervasive environments.

5.5. Challenge E: Eye-Gaze Tracking on Mobile Platforms

The challenge of estimating the eye-gaze direction on mobile devices, such as tablets [84, 89, 152–154] and smartphones [41, 89, 153], encompasses most of the previously discussed challenges that relate to the estimation of gaze from low-resolution eye images captured by the integrated camera on the device, handling of head movement during tracking and support for brief user calibration. Moreover, the use of mobile devices extends these challenges further due to the limited availability of computational resources on such platforms and the movement of the device itself.

Methods that seek to estimate the eye-gaze on mobile platforms with limited computational resources typically aim for lightweight algorithms that can run robustly without jeopardising the gaze estimation accuracy. This often requires a thoughtful choice of algorithms for gaze estimation in order to make effective use of the limited computational resources. Hence, for these particular methods, the challenges of achieving gaze estimation accuracy that is competitive with the state-of-the-art, together with a suitable temporal resolution that permits the method to run in real-time, are often inter-linked.

It has been discussed in Section 5.1 that appearance-based methods have been found effective in estimating the eye-gaze from low resolution images.

Indeed, Holland and Komogortsev [84] make use of machine learning in order to map the eye image information directly to a PoR on a tablet screen. Their method consists of three main steps performed at every time step; face and eye region detection is initially performed by means of a Haar classifier, followed by iris region localisation via template-matching, and finally PoR estimation by feeding a neural network with the pixel intensities of the eye region image. Since the proposed method is to run on a tablet, Holland and Komogortsev [84] decrease the required computational cost by converting the image to grayscale and reducing its pixel resolution by half. Furthermore, the structure of the neural network is limited to two layers in order to reduce the size of the training data set that needs to be collected via calibration. Nonetheless, the method reportedly achieves a low temporal resolution of 0.23 fps on a 1 GHz dual-core processor with 512 MB of memory. The temporal resolution may be improved to 0.70 fps with the removal of the face detection step, however this is reported to reduce the achievable PoR accuracy and constrains both the user and device to a fixed spatial configuration [84], hence counteracting the primary benefit of utilising a portable device. Pino and Kavasidis [41] alternatively exploit the availability of sensors integrated on a smartphone device in order to reduce the computational cost and hence improve the temporal resolution. At every image frame, gyroscope and accelerometer data are retrieved and compared to the corresponding data at the previous time step, permitting computation of the device displacement. Under the assumption that the movement of the handheld device exceeds any head movement performed by the user, eye detection by Haar classification is only carried out if this movement exceeds a certain threshold. Otherwise, the positions of the eye regions are tracked by a Camshift algorithm and the PoR is subsequently estimated according to the centroids of the bounding boxes enclosing the eyes. Pino and Kavasidis [41] do not report the achieved PoR accuracy and processing frame rate, nonetheless a significant reduction in the number of accurately detected fixation points with respect to ground truth was reported when the smartphone was undocked and held by the subjects. Krafka et al. [153] and Huang et al. [154] collect publicly available data sets consisting of images [153] or videos [154] captured under unconstrained conditions by the front-facing camera of a smartphone [153] or tablet device [153, 154]. Krafka et al. [153] aim for a large-scale data set that is characterised by substantial variability in pose, appearance and illumination, while Huang et al. [154] capture video data while participants, of diverse ethnic backgrounds, perform gaze shifts while standing, sitting, slouching or lying.

Huang et al. [154] investigate the use of several feature descriptors and regressors to estimate the gaze on the collected video data set. They report that a combination of histogram of oriented gradients descriptors together with random forest regression achieves a gaze estimation accuracy that is competitive with the state-of-the-art, without constraining the head movement or the user-to-screen distance. Krafka et al. [153], on the other hand, employ the use of a deep convolutional neural network (CNN), providing it with separate images of the eyes together with an image of the face and its location within the full image, at the input layer. Based on such input to the model, Krafka et al. [153] aim that the CNN infers meaningful information that relates the eye and head poses, and the head and camera poses. In order to compare with other state-of-the-art methods, support vector regression was applied on features extracted from one of the fully-connected layers of the CNN, which reportedly outperformed the results achieved by other methods under consideration, including those achieved by Huang et al. [154]. By reducing the complexity of the CNN model, Krafka et al. [153] report an expected temporal resolution of 10 to 15 fps on a typical mobile device.

While the appearance-based methods have been gaining increasing attention within the context of gaze estimation from low-resolution images, Wood and Bulling [152] propose a shape-based approach that exploits the circular limbus boundary to estimate a PoR on a tablet screen. At every image frame, the proposed method initially employs a Haar classifier to localise the eye regions coarsely and subsequently improves upon this coarse estimate by exploiting the change in intensity gradient across the limbus boundary. Following ellipse-fitting to the limbus boundary and the computation of a normal vector to the ellipse, the PoR is finally estimated as the average of the intersection points between the vectors projecting from both eyes and the tablet screen. The extraction of fine image detail required by the shape-based method is reported as the main source of error with the highest impact on the estimation accuracy of the PoR. Furthermore, the method of Wood and Bulling [152] is reported to achieve a temporal resolution of 12 fps on a 2 GHz quad-core processor with 8 GB of memory. While the discrepancy in processing power between the devices employed by Holland and Komogortsev [84] and Wood and Bulling [152] does not permit a direct comparison between their computational efficiency, a similar improvement in temporal resolution was also confirmed by Kunze et al. [89] who implemented a shape-based approach based on ellipse-fitting, over the appearance-based method of Holland

and Komogortsev [84] on the same tablet and smartphone devices. Hence, while the use of appearance-based methods appears to be advantageous for gaze estimation from low-resolution eye region images, such as those acquired by the integrated camera on mobile devices, the computational requirements of the method under consideration should complement the limited processing resources of the mobile platform in use.

6. Summary and Discussion of Reviewed Methods

Until recently, eye-gaze tracking has been mostly regarded as a desktop technology that finds its use within the controlled conditions of the classical office environment. As the host of applications for eye-gaze tracking broadens, the necessity to integrate this technology into everyday life settings has been gaining increasing attention. It has been argued herein that in absence of any dependence upon specialised IR illumination sources and imaging hardware, passive video-based eye-gaze tracking would be the natural option for pervasive applications. This critical review, therefore, focused on emerging passive and unobtrusive video-based eye-gaze tracking approaches in recent literature, with the aim to identify different research avenues that are being followed in response to the challenges of pervasive scenarios.

We have seen that the shift towards pervasive eye-gaze tracking has brought with it important progress in the field. The use of consumer-grade cameras, such as webcams, for eye-gaze tracking has been gaining prevalence in recent years, allowing the possibility to perform eye-gaze tracking on devices with integrated imaging hardware without requiring further hardware modification. This has been further reinforced with the widespread use of mobile platforms, such as tablets and smartphones. As a result, we have seen in Section 5.1 that appearance-based methods for gaze estimation have also been receiving increasing attention, since these do not rely upon the extraction of fine image detail and hence permit the estimation of gaze from wide field-of-view images in which the eye regions appear at low resolution [38, 40, 56, 57, 61, 81–84, 86].

This increased attention to apperance-based methods has given rise to a further challenge that relates to the estimation of gaze from sparse [87, 104, 105] and person-independent data [108, 110], as opposed to large data sets of training samples that would require considerable user cooperation in order to collect. User cooperation is also required in the calibration of methods that estimate the PoR by mapping specific eye image features to a

monitor screen, leading to the emergence of methods that aim for implicit calibration or calibration-free gaze estimation [10, 11, 114, 115, 120–122, 155]. We consider these research areas to be in their infancy with much room for further research work and improvements.

The ability to estimate the eye-gaze under different head orientations is a desirable characteristic of pervasive eye-gaze tracking, nonetheless head pose estimation is a challenging problem in itself that has been studied considerably over the years [133]. As methods for head pose estimation evolve and improve in accuracy [156], we expect to see an increasing number of eye-gaze tracking methods that compensate for head movement, together with an improvement in gaze estimation accuracy over the accuracy that has been so far achieved by present head pose invariant methods.

6.1. Use of Data Sets in Comparing Between Methods

It is common practice to report the accuracy of the proposed method after having evaluated this on a collection of data, and to compare the results with other state-of-the-art methods in order to position oneself with respect to the relevant literature. Nonetheless, the data set in use is an important factor to consider when comparing between results of different methods. These results often originate from experimental procedures that use different data sets, which may not have been generated under the same user and ambient conditions.

It may indeed be noticed from Table 1 that the reviewed methods are evaluated on a variety of data sets, where most of these methods are evaluated on test data collected by the authors themselves, while several others make use of different publicly available data sets. The brief descriptions of the typically used public data sets, in Table 2, reveal the diversity of conditions under which these data sets are collected. The conditions under which the self-collected data sets are generated tend to be similarly diverse. These data sets usually feature variations in the characteristics of the participants and their backdrops, the illumination conditions, and the eye and head movements, as well as the imaging hardware by which the images are captured, which in turn affects the image resolution and frame rate. Furthermore, while the public data sets often report the conditions under which the data has been collected, this information is not always provided with self-collected data sets, hence making it difficult for other researchers to reproduce.

One should, therefore, consider these variations in data set collection when comparing between results of eye-gaze tracking methods that make

use of different data sets for evaluation. Kar and Corcoran [157] investigate specifically this problem, identifying impacting factors such as changes in head pose, camera resolution, display properties, positioning of the user with respect to the display, platform motion and illumination changes. They report that most of the methods in the literature investigate the effect of only a few of these factors on the gaze estimation accuracy, such as the effect of head pose variations, while the impact of the remaining factors is rarely considered at all. This lack of a common benchmark for evaluation may indeed pose a serious difficulty in identifying the advantages and disadvantages of different eye-gaze tracking methods. This problem may be even more pronounced for pervasive eye-gaze tracking methods that aim to operate *in the wild*, where variations in the operating conditions are much less constrained. This presents an open problem for the development of pervasive eye-gaze tracking, as will be outlined in Section 7.

6.2. Comparison of Quantitative Results

Many of the reviewed eye-gaze tracking methods quantify their gaze estimation accuracy by measuring either the angular error in yaw and pitch separately, or combine the two into a single measurement, as reported in Table 1. This error is measured with respect to ground truth data, and is typically presented as the mean angular error of the left and right eyes. Most eye-gaze tracking methods follow one of two approaches in order to quantify their accuracy: by performing a validation procedure whereby the user is asked to gaze at physical or screen markers, where the marker positions serve as ground truth data, or by running the implemented algorithms on a data set of images, as indicated in Table 1. In both cases, the collection of data in use, whether self-collected or publicly available, serves as ground truth with respect to which the gaze estimation error may be measured.

The results in Table 1 suggest that free head movement tends to introduce an increase in the gaze estimation error, when compared with the results achieved by methods that consider a frontal and stationary head pose, or which compensate for small head movement alone. Several of the methods that have been tested under free head movement in Table 1 [106, 107, 110, 134, 135, 146, 147], indeed report some of the highest gaze estimation errors among all the state-of-the-art methods that have been considered. In particular, some of the highest gaze estimation errors have been reported by appearance-based methods that address Challenge B, which relates to

the estimation of gaze from sparse, synthesised or person-independent training samples, and which aim to compensate for free head movement as well [106, 107, 110]. The gaze estimation accuracy of such methods generally depends on the appearance variations of the training data, which should represent the expected tracking conditions well. Head movement offers a challenge in this regard by introducing additional variations in facial appearance under different head poses. A possible source of error for these appearance-based methods may, therefore, relate to insufficient representation of the appearance variations that arise due to head movement, by the data set in use. In the case of feature-based methods that rely on model-fitting for head pose estimation [134, 135, 146, 147], the gaze estimation error often originates from inaccurate detection or tracking of specific face landmarks, which become increasingly distorted and occluded with increasing head rotation angles.

Nonetheless, as has been previously discussed in Section 6.1, this increase in error may be due to multiple factors, one of which relates to the conditions under which the data set in use has been collected. In light of this difficulty in comparing across different methods, we alternatively compare between different performance measures of the same method when this has been tested under different tracking conditions. Table 3 presents the state-of-the-art methods for which such performance measures have been reported. The table reports the gaze estimation error achieved under stationary, natural or free head movement for the challenge under consideration, together with the gaze estimation error achieved in absence of the challenge being addressed. The values within this table indeed reveal that different levels of head movement tend to introduce an increase in the gaze estimation error, over the error that may otherwise be achieved by the same method under a stationary head pose [87, 110, 123]. This increase in error was also observed in the values of Table 1 and is consistent with the findings of Kar and Corcoran [157], who investigated the effect of head pose variations on the gaze estimation accuracy. It has been previously discussed that possible causes for this may be the insufficient representation of appearance variations in the training data [87], or inaccurate fitting of a face model to specific face landmarks for detection and tracking [110]. The values within the same table also reveal that methods which do not handle head movement but are tested in conditions where the user is allowed to move naturally, could benefit from head movement compensation [97, 98]. Indeed, Cristina and Camilleri [98] improve over previous research work [97] by introducing a mechanism that handles variations in head pose, by approximating these to planar movement,

and test both methods under natural head movement [97, 98]. A comparison of the achieved results reveals a reduction in the gaze estimation error when head movement compensation was included [98].

Furthermore, the gaze estimation error is also noticed to increase in absence of controlled conditions, which would otherwise constrain the user to perform calibration or permit the collection of user-specific training data [110, 123, 128]. For comparison purposes, Mora and Odobez [110] evaluate the gaze estimation results obtained through the use of person-independent training samples, against the use of user-specific training data. Similarly, Lu et al. [123] and Model and Eizenman [128] compare the results achieved through their calibration-free methods, with those obtained following the inclusion of a calibration procedure prior to the estimation of gaze. The reported results suggest that facilitating eye-gaze tracking in less controlled scenarios by reducing the burden of data collection and calibration on the user, tends to jeopardise the achievable gaze estimation accuracy. The level of gaze estimation error presently reported by methods that address the challenges of pervasive eye-gaze tracking [110, 123, 128] appears to be higher than the error reported for methods that are tested in the absence of the respective challenge. This suggests that the methods by which the pervasive eye-gaze tracking challenges are being addressed have, so far, not been able to compensate sufficiently for the variations brought about by such challenges. This highlights the importance of addressing the challenges of pervasive eye-gaze tracking, and to report such performance measures for the proposed method which allows the results to be interpreted in the correct context.

7. Future Directions for Pervasive Eye-Gaze Tracking

Several of the future directions that Hansen and Ji had outlined in their survey six years ago [15] have been increasingly taken up by the eye-gaze tracking community over recent years. Several of these directions point towards a limited use of IR illumination, limited need for calibration, and reduced costs by the use of consumer-grade cameras. This uptake is evidenced by the numerous methods discussed within this review, which aimed to address different challenges associated with pervasive eye-gaze tracking.

We propose additional directions to be considered in the development of future pervasive eye-gaze tracking methods, as follows:

- ***Robustness to non-rigid face deformations***: Non-rigid face deformations due to changes in facial expression are a common occurrence

during tracking, yet this challenge is not typically addressed by the eye-gaze tracking community. Head pose invariant methods that infer the head orientation from the relative configuration of specific facial features are particularly susceptible to non-rigid face deformations [158], since these often regard the face as a rigid body. In light of the challenges of pervasive eye-gaze tracking, a future direction is to develop methods that are robust to non-rigid face deformations, hence lifting further constraints on the natural user movement.

- ***Robustness to appearance variations***: Issues arising from appearance variations, such as different user ethnicities or the use of spectacles during tracking, have also received little attention in the literature. Ethnic variations of the face features, such as those relating to the appearance of the eyes and colour of the skin, may affect methods that exploit the photometric appearance if these are not catered for. Furthermore, the use of spectacles has also been reported to hinder participants from making effective use of an eye-controlled public display [11], with inaccuracies in the eye detection due to the presence of the spectacles frame around the eye regions being reported as the main cause. Future work on eye-gaze tracking should, therefore, address these issues as an important step towards gaze estimation in pervasive scenarios.

- ***Simultaneous tracking of multiple users***: Research on user interaction in the wild reportedly indicates that often multiple users will attempt to interact with a display screen simultaneously [10, 11], which requires further understanding of the interaction modes between multiple users in pervasive scenarios. Hence, this calls for methods that address the challenges associated with eye-gaze tracking in unconstrained environments, such as the detection of multiple eye image pairs and subsequent mapping of the image information to multiple PoRs on the same display screen, without generating conflicting commands.

- ***Illumination invariance***: Variability of the surrounding illumination is one of the main challenges of eye-gaze tracking in the wild, such as in open [10, 11] and outdoor spaces [127], yet this challenge has not been widely addressed in the literature. Methods that exploit the photometric appearance of the eye region may be particularly susceptible to illumination variations, since these variations may alter the perceived eye region appearance inside the image space. A few of such

methods that address this challenge typically perform histogram equalisation during a pre-processing stage [37, 95], in order to enhance the image contrast. Alternatively, other methods alleviate this problem by employing artificial illumination sources, such as lamplight [62] or IR illumination [57, 90], in order to stabilise the surrounding illumination. Nonetheless, the latter is not a desirable solution in real-life scenarios, especially if the eye-gaze had to be tracked in open and outdoor spaces that would not permit the use of artificial illumination. Hence, future work should aim for illumination invariance as one of the main challenges of pervasive eye-gaze tracking.

- **Platforms with limited resources**: Over recent years, we have witnessed an increasing uptake in the use of off-the-shelf mobile devices comprising integrated imaging hardware, such as tablets and smartphones. It is desirable to employ the eye-gaze as an additional modality for interaction with these devices, as we have seen within this review, despite their limited resources in terms of computing power and screen real estate. Most research work has so far implemented widely-used algorithms for eye-gaze tracking on such mobile devices, performing small algorithmic modifications in order to reduce the required computational cost. These methods have typically reported temporal resolutions lower than 12fps, depending on the available computational resources, and gaze estimation error of up to $7°$, which is considerably large with respect to the small screen size of such devices. Hence, future research may consider the development of methods that are specifically targeted to exploit the limited resources of such devices, aiming to improve the temporal resolution and gaze estimation accuracy.

- **Spontaneous user interaction**: Most research effort over the past years has been mainly dedicated to the development of methods that condition the user to a sitting position in front of a display device, such as a computer monitor. Recent years have, nonetheless, seen the deployment of an increasing number of display devices in public spaces [159], such as in shopping malls, airports and train stations, which present a good opportunity for user engagement and interaction [159]. Since passersby will tend to interact with the display spontaneously and over short periods of time, particular challenges of tracking on public displays relate to interaction while the user is on the move, at different

user-display distances and positions [159]. Methods that address these challenges appear to be gaining traction in the literature [159], which will potentially lead to compelling applications for eye-gaze tracking outside of the office environment in the near future.

- ***Improvement of gaze estimation accuracy***: The level of acceptable gaze estimation accuracy may be subject to the application at hand, such as for gaze interaction on large displays, where visual stimuli tend to be large in size and spaced apart, high accuracy may not necessarily be required. Nonetheless, a high level of gaze estimation accuracy would be desirable in many other applications facilitating, for instance, the analysis of gaze patterns on cluttered scenes or designs, or gaze interaction with dense user interfaces on a tablet or smartphone screen. It does not appear that present gaze estimation methods are yet achieving this level of accuracy, with many of the reported gaze estimation errors being higher than $5°$. We have also seen, in Section 6.2, that the gaze estimation error also tends to increase with the introduction of challenges associated with pervasive eye-gaze tracking, where at times this error was seen to exceed $10°$. Hence, future work should strive to improve the gaze estimation accuracy of passive methods, bringing this closer to that achieved by active eye-gaze tracking systems [15]. A possible direction for future work may be to explore new ways of exploiting the information in the image to improve the gaze estimation accuracy.

- ***Common benchmark for evaluation***: We have previously outlined in Section 6.1, that in absence of a common benchmark for evaluation, a reliable comparison between the results of different eye-gaze tracking methods becomes difficult to carry out. We have seen that, while there are various factors that may have a considerable impact on the gaze estimation accuracy, their effect is rarely evaluated by most of the methods in the literature [157]. In view of this, designing a benchmark, which permits a reliable assessment of different eye-gaze tracking methods, becomes an imperative task to shed light on key factors that may impact the field in such a way as to hamper its advancement if not addressed. The information on each error source that may be extracted through controlled experiments, may possibly steer the development of future methods into new directions.

8. Conclusion

An increasing interest in pervasive eye-gaze tracking has generated a growing effort toward the development of eye-gaze tracking methods that can robustly handle the challenges associated with uncontrolled everyday life scenarios, as evidenced by the literature that has been discussed within this review. Nonetheless, further open challenges that have not yet received substantial attention remain and, hence, these call for methods that can handle emerging challenges associated with the use of eye-gaze tracking under uncontrolled conditions in a robust manner. This paradigm shift in approaching the problem of eye-gaze tracking that progresses over the classical notion of eye-gaze tracking as a desktop technology, increases the applicability and uptake of eye-gaze tracking technology through a wider variety of applications.

Acknowledgement

References

[1] R. J. K. Jacob, What you look at is what you get: eye movementbased interaction techniques, in: Proceedings of the SIGCHI conference on Human factors in computing systems: Empowering people, 1990, pp. 11–18.

[2] J. L. Levine, An Eye-controlled Computer, Vol. Res. Rep. RC-8857, IBM Research Division, T.J. Watson Research Center, 1981.

[3] M. I. Khan, A. B. Mansoor, Real time eyes tracking and classification for driver fatigue detection, in: Proceedings of the 5th International Conference on Image Analysis and Recognition, 2008, pp. 729–738.

[4] H. Singh, J. S. Bhatia, J. Kaur, Eye tracking based driver fatigue monitoring and warning system, in: 2010 India International Conference on Power Electronics, 2011, pp. 1–6.

[5] S. J. Lee, J. Jo, H. G. Jung, K. R. Park, J. Kim, Real-time gaze estimator based on driver's head orientation for forward collision warning system, IEEE Transactions on Intelligent Transportation Systems 12 (1) (2011) 254–267.

[6] M. Shahid, T. Nawaz, H. A. Habib, Eye-gaze and augmented reality framework for driver assistance, Life Science Journal 10 (3) (2013) 1571–1578.

[7] S. Kohlbecher, E. Wiese, K. Bartl, J. Blume, A. Bannat, E. Schneider, Studying gaze-based human robot interaction: An experimental platform, in: Proceedings of the Gaze in Human-Robot Interaction Workshop, 2012.

[8] L. Gentry, Marketing and eye-tracking machines: Research opportunities for an affordable tool, Journal of International Management Studies (2007) 60–65.

[9] A. Bulling, A. Duchowski, P. Majaranta, The 1st international workshop on pervasive eye tracking and mobile eye-based interaction (petmei 2011), in: Proceedings of the 13th International Conference on Ubiquitous Computing, 2011, pp. 627–628.

[10] Y. Zhang, A. Bulling, H. Gellersen, Sideways: A gaze interface for spontaneous interaction with situated displays, in: SIGCHI Conference on Human Factors in Computing Systems, 2013, pp. 851–860.

[11] Y. Zhang, M. K. Chong, J. Muller, A. Bulling, H. Gellersen, Eye tracking for public displays in the wild, Personal and Ubiquitous Computing 19 (5) (2015) 967–981.

[12] R. Urano, R. Suzuki, T. Sasaki, Eye gaze estimation based on ellipse fitting and three-dimensional model of eye for "intelligent poster", in: IEEE/ASME International Conference on Advanced Intelligent Mechatronics, 2014, pp. 1157–1162.

[13] R. Remmel, An inexpensive eye movement monitor using the scleral search coil technique, IEEE Transactions on Biomedical Engineering (1984) 388–390.

[14] R. Barea, System for assisted mobility using eye movements based on electrooculography, IEEE Transactions on Neural Systems and Rehabilitation Engineering 10 (4) (2002) 209–218.

[15] D. W. Hansen, Q. Ji, In the eye of the beholder: A survey of models for eyes and gaze, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (3) (2010) 478–500.

[16] COGAIN, Eye trackers, http://wiki.cogain.org/index.php/Eye_Trackers.

[17] O. V. Komogortsev, I. Rigas, Current research in eye movement biometrics: An analysis based on bioeye 2015 competition, Image and Vision Computing 58 (2017) 129–141.

[18] C. Galdi, M. Nappi, D. Riccio, H. Wechsler, Eye movement analysis for human authentication: a critical survey, Pattern Recognition Letters 84 (2016) 272–283.

[19] C. Galdi, M. Nappi, D. Riccio, V. Cantoni, M. Porta, A new gaze analysis based soft-biometric, in: J. A. Carrasco-Ochoa, J. F. Martínez-Trinidad, J. S. Rodríguez, G. S. di Baja (Eds.), Pattern Recognition, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 136–144.

[20] I. Rigas, E. Abdulin, O. Komogortsev, Towards a multi-source fusion approach for eye movement-driven recognition, Information Fusion 32 (Part B) (2016) 13–25.

[21] K. D. Toennies, F. Behrens, M. Aurnhammer, Feasibility of hough-transform-based iris localisation for real-time-application, in: Proceedings of the 16th International Conference on Pattern Recognition, Vol. 2, 2002, pp. 1053–1056.

[22] A. Al-Rahayfeh, M. Faezipour, Classifiers comparison for a new eye gaze direction classification system, in: IEEE Long Island Systems, Applications and Technology Conference (LISAT), 2014, pp. 1–6.

[23] R. Kothari, J. L. Mitchell, Detection of eye locations in unconstrained visual images, in: Proceedings of the International Conference on Image Processing, Vol. 3, 1996, pp. 519–522.

[24] R. Valenti, T. Gevers, Accurate eye center location and tracking using isophote curvature, in: Proceedings of the IEEE Conference Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[25] R. Valenti, J. Staiano, N. Sebe, T. Gevers, Webcam-based visual gaze estimation, in: Proceedings of the 15th International Conference on Image Analysis and Processing, 2009, pp. 662–671.

[26] R. Valenti, A. Lablack, N. Sebe, C. Djeraba, T. Gevers, Visual gaze estimation by joint head and eye information, in: 20th International Conference on Pattern Recognition, 2010, pp. 3870–3873.

[27] R. Valenti, N. Sebe, T. Gevers, Combining head pose and eye location information for gaze estimation, IEEE Transactions on Image Processing 21 (2) (2012) 802–815.

[28] Z. Hammal, C. Massot, G. Bedoy, A. Caplier, Eyes segmentation applied to gaze direction and vigilance estimation, in: Third International Conference on Advances in Pattern Recognition, 2005, pp. 236–246.

[29] T. A. Camus, R. Wildes, Reliable and fast eye finding in close-up images, in: Proceedings of the 16th International Conference on Pattern Recognition,, Vol. 1, 2002, pp. 389–394.

[30] D. B. L. Bong, K. H. Lim, Application of fixed-radius hough transform in eye detection, International Journal of Intelligent Information Technology Application 2 (3) (2009) 121–127.

[31] D. Zhu, S. T. Moore, T. Raphan, Robust pupil center detection using a curvature algorithm, Computer Methods and Programs in Biomedicine, 59 (3) (1999) 145–157.

[32] D. Leimberg, M. Vester-Christensen, B. K. Ersboll, L. K. Hansen, Heuristics for speeding up gaze estimation, in: Proceedings Svenska Symposium i Bildanalys, 2005.

[33] V. Vezhnevets, A. Degtiareva, Robust and accurate eye contour extraction, in: International Conference GraphiCon, 2003, pp. 81–84.

[34] Z. Hammal, A. Caplier, Eyes and eyebrows parametric models for automatic segmentation, in: 6th IEEE Southwest Symposium on Image Analysis and Interpretation, 2004, pp. 138–141.

[35] W. Haiyuan, Y. Kitagawaa, T. Wada, Tracking iris contour with a 3d eye-model for gaze estimation, in: Proceedings of the 8th Asian Conference on Computer Vision, 2007, pp. 688–697.

[36] T. Moriyama, T. Kanade, J. Xiao, J. F. Cohn, Meticulously detailed eye region model and its application to analysis of facial images, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (5).

[37] T. H. Huynh, A new eye gaze detection algorithm using pca features and recurrent neural networks, in: 2013 IEEE Symposium on Computational Intelligence in Control and Automation, 2013, pp. 24–29.

[38] B. L. Nguyen, Y. Chahir, M. Molina, C. Tijus, F. Jouen, Eye gaze tracking with free head movements using a single camera, in: Symposium on Information and Communication Technology, 2010, pp. 108–113.

[39] B. L. Nguyen, Eye gaze tracking, in: International Conference on Computing and Communication Technologies, 2009, pp. 1–4.

[40] R. Dhyawala, J. Dhariwal, N. Nain, Eye gazing with low resolution web-cam images using artificial neural network, in: Proceedings of the International Conference on Advances in Computer, Electronics and Electrical Engineering, 2012, pp. 443–447.

[41] C. Pino, I. Kavasidis, Improving mobile device interaction by eye tracking analysis, in: Proceedings of the Federated Conference on Computer Science and Information Systems, 2012, pp. 1199–1202.

[42] R. C. Motwani, M. C. Motwani, F. C. Harris, Eye detection using wavelets and ann, in: Proceedings of the Global Signal Processing Expo and Conference (GSPx), 2004, pp. 27–30.

[43] B. Noris, K. Benmachiche, A. G. Billard, Calibration-free eye gaze direction detection with gaussian processes, in: Proceedings of the International Conference on Computer Vision Theory and Applications, Vol. 2, 2008, pp. 611–616.

[44] N. Markus, M. Frljak, I. S. Pandzic, J. Ahlberg, R. Forchheimer, Eye pupil localization with an ensemble of randomized trees, Pattern Recognition 47 (2) (2014) 578–587.

[45] E. Demjen, V. Abosi, Z. Tomori, Eye tracking using artificial neural networks for human computer interaction, Physiological Research 60 (2011) 841–844.

[46] D. Droege, D. Paulus, Pupil center deetection in low resolution images, in: Proceedings of the 2010 Symposium on Eye-Tracking Research and Applications, 2010, pp. 169–172.

[47] J. Heinzmann, A. Zelinsky, 3-d facial pose and gaze point estimation using a robust real-time tracking paradigm, in: Third IEEE International Conference on Automatic Face and Gesture Recognition, 1997, pp. 142–147.

[48] N. Bhoi, M. N. Mohanty, Template matching based eye detection in facial image, International Journal of Computer Applications 12 (5) (2010) 15–18.

[49] R. Ohtera, T. Horiuchi, S. Tominaga, Eye-gaze detection from monocular camera image using parametric template matching, in: Proceedings of the 8th Asian Conference on Computer Vision, 2007, pp. 708–717.

[50] T. Akashi, Y. Wakasa, K. Tanaka, S. Karungaru, M. Fukumi, Using genetic algorithm for eye detection and tracking in video sequence, Journal of Systemics, Cybernetics and Informatics 5 (2) (2007) 72–78.

[51] R. F. Abdel-Kader, R. Atta, S. El-Shakhabe, An efficient eye detection and tracking system based on particle swarm optimization and adaptive block-matching search algorithm, Engineering Applications of Artificial Intelligence 31 (2014) 90–100.

[52] K. Kim, R. S. Ramakrishna, Vision-based eye-gaze tracking for human computer interface, in: IEEE International Conference on Systems, Man, and Cybernetics, Vol. 2, 1999, pp. 324–329.

[53] F. Timm, E. Barth, Accurate eye centre localisation by means of gradients, in: Proceedings of the Sixth International Conference on Computer Vision Theory and Applications, 2011, pp. 125–130.

[54] E. G. Dehkordi, M. Mahlouji, H. E. Komleh, Human eye tracking using particle filters, International Journal of Computer Science Issues 10 (2) (2013) 107–115.

[55] T. D'Orazio, M. Leo, G. Cicirelli, A. Distante, An algorithm for real time eye detection in face images, in: Proceedings of the 17th International Conference on Pattern Recognition, Vol. 3, 2004, pp. 278–281.

[56] L. Q. Xu, D. Machin, P. Sheppard, A novel approach to realtime non-intrusive gaze finding, in: British Machine Vision Conference, 1998, pp. 428–437.

[57] K. H. Tan, D. J. Kriegman, N. Ahuja, Appearance-based eye gaze estimation, in: Proceedings of the Sixth IEEE Workshop on Applications of Computer Vision, 2002, pp. 191–195.

[58] T. Rajpathak, R. Kumar, E. Schwartz, Eye detection using morphological and color image processing, in: Florida Conference on Recent Advances in Robotics, 2009, pp. 1–6.

[59] R. Stiefelhagen, J. Yang, Gaze tracking for multimodal human- computer interaction, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 4, 1997, pp. 2617–2620.

[60] J. G. Ko, K. Kim, S. Choi, J. Y. Kim, K. Kim, J. Kim, Facial feature tracking and head orientation-based gaze tracking, in: Proceedings of the International Conference on Circuits/Systems, Computers and Communications, 2000, pp. 415–420.

[61] Y. T. Lin, R. Y. Lin, Y. C. Lin, G. C. Lee, Real-time eye-gaze estimation using a low-resolution webcam, Multimedia Tools and Applications 65 (2013) 543–568.

[62] N. M. Piratla, A. P. Jayasumana, A neural network based real-time gaze tracker, Journal of Network and Computer Applications (2002) 179–196.

[63] S. Wibirama, S. Tungjitkusolmun, C. Pintavirooj, K. Hamamoto, Real time eye tracking using initial centroid and gradient analysis technique, in: Proceedings of the 6th IEEE International Conference of Electrical Engineering/Electronics, Computer, Telecommunications, and Information Technology (ECTI), 2009, pp. 1054–1057.

[64] D. Merad, S. Metz, S. Miguet, Eye and gaze tracking algorithm for collaborative learning system, in: Proceedings of the Third International Conference on Informatics in Control, Automation and Robotics, Robotics and Automation, 2006, pp. 326–333.

[65] P. Jafari Moghadam Fard, M. H. Moradi, S. Parvaneh, Eye tracking using a novel approach, in: IFMBE Proceedings, Vol. 14, 2007, pp. 2407–2410.

[66] K. M. Evans, R. A. Jacobs, J. A. Tarduno, J. B. Pelz, Collecting and analyzing eye-tracking data in outdoor environments, Journal of Eye Movement Research 5 (2) (2012) 1–19.

[67] R. Szeliski, Computer Vision: Algorithms and Applications, Springer London, 2011.

[68] W. Haiyuan, Y. Kitagawaa, T. Wada, Tracking iris contour with a 3d eye-model for gaze estimation, in: Proceedings of the 8th Asian Conference on Computer Vision, 2007, pp. 688–697.

[69] T. Ishikawa, S. Baker, I. Matthews, T. Kanade, Passive driver gaze tracking with active appearance models, in: Proceedings of the 11th World Congress on Intelligent Transportation Systems, Vol. 9, 2004, pp. 100–109.

[70] E. Pogalin, A. Redert, I. Patras, E. A. Hendriks, Gaze tracking by using factorized likelihoods particle filtering and stereo vision, in: Third International Symposium on 3D Data Processing, Visualization, and Transmission, 2006, pp. 57–64.

[71] M. Reale, T. Hung, L. Yin, Viewing direction estimation based on 3d eyeball construction for hri, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2010, pp. 24–31.

[72] M. J. Reale, S. Canavan, L. Yin, K. Hu, T. Hung, A multi-gesture interaction system using a 3-d iris disk model for gaze estimation and an active appearance model for 3-d hand pointing, IEEE Transactions on Multimedia 13 (3) (2011) 474–486.

[73] J. Chen, Q. Ji, 3d gaze estimation with a single camera without ir illumination, in: 19th International Conference on Pattern Recognition, 2008, pp. 1–4.

[74] Y. Matsumoto, A. Zelinsky, An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement, in: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, 2000, pp. 499–504.

[75] J. Wang, E. Sung, R. Venkateswarlu, Estimating the eye gaze from one eye, Computer Vision and Image Understanding - Special Issue on Eye Detection and Tracking 98 (1) (2005) 83–103.

[76] H. Yamazoe, A. Utsumi, T. Yonezawa, S. Abe, Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions, in: Proceedings of the 2008 Symposium on Eye Tracking Research and Applications, 2008, pp. 245–250.

[77] R. Ronsse, O. White, P. Lefevre, Computation of gaze orientation under unrestrained head movements, Journal of Neuroscience Methods 159 (1) (2007) 158–169.

[78] N. H. Cuong, H. T. Hoang, Eye-gaze detection with a single webcam based on geometry features extraction, in: 2010 11th International Conference on Control, Automation, Robotics and Vision, 2010, pp. 2507–2512.

[79] J. Zhu, J. Yang, Subpixel eye gaze tracking, in: 5th IEEE International Conference on Automatic Face and Gesture Recognition, 2002, pp. 124–129.

[80] H. Wang, C. Pan, C. Chaillou, Tracking eye gaze under coordinated head rotations with an ordinary camera, in: 9th Asian Conference on Computer Vision, Vol. 5995, 2010, pp. 120–129.

[81] S. Kim, B. Hwang, M. Lee, Gaze tracking based on pupil estimation using multilayer perception, in: Proceedings of International Joint Conference on Neural Networks, 2011, pp. 2683–2689.

[82] W. Sewell, O. Komogortsev, Real-time eye gaze tracking with an unmodified commodity webcam employing a neural network, in: Proceedings of the 28th International Conference Extended Abstracts on Human Factors in Computing Systems, 2010, pp. 3739–3744.

[83] R. Stiefelhagen, J. Yang, A. Waibel, Tracking eyes and monitoring eye gaze, in: Proceedings of the Workshop on Perceptual User Interfaces, 1997, pp. 98–100.

[84] C. Holland, O. Komogortsev, Eye tracking on unmodified common tablets: Challenges and solutions, in: Proceedings of the Symposium on Eye Tracking Research and Applications, 2012, pp. 277–280.

[85] Y. Zhang, A. Bulling, H. Gellersen, Towards pervasive eye tracking using low-level image features, in: Proceedings of the Symposium on Eye Tracking Research and Applications, 2012, pp. 261–264.

[86] J. Mansanet, A. Albiol, R. Paredes, J. M. Mossi, A. Albiol, Estimating point of regard with a consumer camera at a distance, Pattern Recognition and Image Analysis (2013) 881–888.

[87] F. Lu, Y. Sugano, T. Okabe, Y. Sato, Adaptive linear regression for appearance-based gaze estimation, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (10) (2014) 2033–2046.

[88] J. Wang, E. Sung, Gaze determination via images of irises, Image and Vision Computing 19 (12) (2001) 891–911.

[89] K. Kunze, S. Ishimaru, Y. Utsumi, K. Kise, My reading life - towards utilizing eyetracking on unmodified tablets and phones, in: Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing, 2013, pp. 283–286.

[90] S. Wibirama, V. Mahasitthiwat, S. Tungjitkusolmun, C. Pintavirooj, K. Vichienchom, Eye-motion tracking using motion gradient, in: The 3rd International Symposium on Biomedical Engineering, 2008, pp. 84–87.

[91] P. Smith, M. Shah, N. da Vitoria Lobo, Monitoring head/eye motion for driver alertness with one camera, in: Proceedings of the 15th International Conference on Pattern Recognition, Vol. 4, 2000, pp. 636–642.

[92] T. Okabe, F. Lu, Y. Sugano, Y. Sato, Adaptive linear regression for appearance-based gaze estimation, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (10) (2014) 2033–2046.

[93] I. F. Ince, J. W. Kim, A 2d eye gaze estimation system with low-resolution webcam images, Journal on Advances in Signal Processing 2011 (2011) 1–11.

[94] Z. Janko, L. Hajder, Improving human-computer interaction by gaze tracking, in: IEEE International Conference on Cognitive Infocommunications, 2012, pp. 155–160.

[95] A. Wojciechowski, K. Fornalczyk, Single web camera robust interactive eye-gaze tracking method, Bulletin of the Polish Academy of Sciences, Technical Sciences 63 (4) (2015) 879–886.

[96] F. Lu, Y. Gao, X. Chen, Estimating 3d gaze directions using unlabeled eye images via synthetic iris appearance fitting, IEEE Transactions on Multimedia 18 (9) (2016) 1772–1782.

[97] S. Cristina, K. P. Camilleri, Cursor control by point-of-regard estimation for a computer with integrated webcam, in: The 8th International Conference on Advanced Engineering Computing and Applications in Sciences (ADVCOMP), 2014, pp. 126–131.

[98] S. Cristina, K. P. Camilleri, On-screen point-of-regard estimation under natural head movement for a computer with integrated webcam, Journal on Advances in Software 8 (3 and 4) (2015) 327–338.

[99] P. Koutras, P. Maragos, Estimation of eye gaze direction angles based on active appearance models, in: IEEE International Conference on Image Processing, 2015, pp. 2424–2428.

[100] F. Martinez, A. Carbone, E. Pissaloux, Gaze estimation using local features and non-linear regression, in: IEEE International Conference on Image Processing, 2012, pp. 1961–1964.

[101] B. C. Chen, P. C. Wu, S. Y. Chien, Real-time eye localization, blink detection, and gaze estimation system without infrared illumination, in: IEEE International Conference on Image Processing, 2015, pp. 715–719.

[102] A. George, A. Routray, Real-time eye gaze direction classification using convolutional neural network, in: IEEE International Conference on Signal Processing and Communication, 2016, pp. 1–5.

[103] N. Wojke, J. Hedrich, D. Droege, D. Paulus, Gaze-estimation for consumer-grade cameras using a gaussian process latent variable model, Pattern Recognition and Image Analysis 26 (1) (2016) 248–255.

[104] F. Lu, Y. Sugano, T. Okabe, Y. Sato, Inferring human gaze from appearance via adaptive linear regression, in: IEEE International Conference on Computer Vision, 2011, pp. 153–160.

[105] K. Liang, Y. Chahir, M. Molina, C. Tijus, F. Jouen, Appearance-based gaze tracking with spectral clustering and semi-supervised gaussian process regression, in: Proceedings of the 2013 Conference on Eye Tracking South Africa, 2013, pp. 17–23.

[106] E. Wood, T. Baltrusaitis, X. Zhang, Y. Sugano, P. Robinson, A. Bulling, Rendering of eyes for eye-shape registration and gaze estimation, in: IEEE International Conference on Computer Vision, 2015, pp. 3756–3764.

[107] E. Wood, T. Baltrusaitis, L. Morency, P. Robinson, A. Bulling, Learning an appearance-based gaze estimator from one million synthesised images, in: ACM Symposium on Eye Tracking Research and Applications, 2016, pp. 131–138.

[108] T. Schneider, S. B., R. Stiefelhagen, Manifold alignment for person independent appearance-based gaze estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1167–1172.

[109] F. Lu, X. Chen, Person-independent eye gaze prediction from eye images using patch-based features, Journal of Neurocomputing 182 (2016) 10–17.

[110] K. A. F. Mora, J. M. Odobez, Person independent 3d gaze estimation from remote rgb-d cameras, in: IEEE International Conference on Image Processing, 2013, pp. 2787–2791.

[111] N. J. Sasson, J. T. Elison, Eye tracking young children with autism, Journal of Visualized Experiments 61 (3675) (2012) 1–5.

[112] D. Moore, A. Gorra, M. Adams, J. Reaney, H. Smith, Disabled Students in Education: Technology, Transition, and Inclusivity, 1st Edition, IGI Global, 2011.

[113] D. R. Flatla, C. Gutwin, L. E. Nacke, S. Bateman, R. L. Mandryk, Calibration games: making calibration tasks enjoyable by adding motivating game elements, in: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, 2011, pp. 403–412.

[114] K. Pfeuffer, M. Vidal, J. Turner, A. Bulling, H. Gellersen, Pursuit calibration: Making gaze calibration less tedious and more flexible, in: ACM Symposium on User Interface Software and Technology, 2013, pp. 261–269.

[115] F. Alnajar, T. Gevers, R. Valenti, S. Ghebreab, Calibration-free gaze estimation using human gaze patterns, in: IEEE International Conference on Computer Vision, 2013, pp. 137–144.

[116] A. Papoutsaki, P. Sangkloy, J. Laskey, N. Daskalova, J. Huang, J. Hays, Webgazer: Scalable webcam eye tracking using user interactions, in: International Joint Conference on Artificial Intelligence, 2016, pp. 3839–3845.

[117] Y. Sugano, Y. Matsushita, Y. Sato, H. Koike, Appearance-based gaze estimation with online calibration from mouse operations, IEEE Transactions on Human-Machine Systems 45 (6) (2015) 750–760.

[118] M. X. Huang, T. C. K. Kwok, G. Ngai, S. C. F. Chan, H. V. Leong, Building a personalized, auto-calibrating eye tracker from user interactions, in: ACM Conference on Human Factors in Computing Systems, 2016, pp. 5169–5179.

[119] M. X. Huang, T. C. K. Kwok, H. V. Leong, S. C. F. Chan, Building a self-learning eye gaze model from user interaction data, in: ACM International Conference on Multimedia, 2014, pp. 1017–1020.

[120] Y. Sugano, Y. Matsushita, Y. Sato, Calibration-free gaze sensing using saliency maps, in: 23rd IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2667–2674.

[121] R. Valenti, N. Sebe, T. Gevers, What are you looking at? improving visual gaze estimation by saliency, International Journal of Computer Vision 98 (3) (2012) 324–334.

[122] Y. Zhang, A. Bulling, H. Gellersen, Pupil-canthi-ratio: A calibration-free method for tracking horizontal gaze direction, in: International Working Conference on Advanced Visual Interfaces, 2014, pp. 129–132.

[123] F. Lu, X. Chen, Y. Sato, Appearance-base gaze estimation via uncalibrated gaze pattern recovery, IEEE Transactions on Image Processing 26 (4) (2017) 1543–1553.

[124] S. Kohlbecher, S. Bardins, K. Bartl, E. Schneider, T. Poitschke, M. Ablassmeier, Calibration-free eye tracking by reconstruction of the pupil ellipse in 3d space, in: Symposium on Eye Tracking Research and Applications, 2008, pp. 135–138.

[125] F. Klefenz, P. Husar, D. Krenzer, A. Hess, Real-time calibration-free autonomous eye tracker, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, 2010, pp. 762–765.

[126] S. Cristina, K. P. Camilleri, Model-based head pose-free gaze estimation for assistive communication, Computer Vision and Image Understanding - Special Issue on Assistive Computer Vision and Robotics.

[127] T. Nagamatsu, T. Ueki, J. Kamahara, Automatic user calibration for gaze-tracking systems by looking into the distance, in: International Workshop on Pervasive Eye Tracking and Mobile Eye-based Interaction, 2013, pp. 125–130.

[128] D. Model, M. Eizenman, An automatic personal calibration procedure for advanced gaze estimation systems, IEEE Transactions on Biomedical Engineering 57 (5) (2010) 1031–1039.

[129] J. Chen, Q. Ji, Probabilistic gaze estimation without active personal calibration, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 609–616.

[130] J. Chen, Q. Ji, A probabilistic approach to online eye gaze tracking without explicit personal calibration, IEEE Transactions on Image Processing 24 (3) (2015) 1076–1086.

[131] C. Nitschke, A. Nakazawa, T. Nishida, I see what you see: Point of gaze estimation from corneal images, in: Asian Conference on Pattern Recognition, 2013, pp. 298–304.

[132] K. Takemura, S. Kimura, S. Suda, Estimating point-of-regard using corneal surface image, in: Symposium on Eye Tracking Research and Applications, 2014, pp. 251–254.

[133] M. Murphy-Chutorian, M. M. Trivedi, Head pose estimation in computer vision: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (4) (2009) 607–626.

[134] K. A. F. Mora, J. M. Odobez, Gaze estimation from multimodal kinect data, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 2012, pp. 25–30.

[135] B. Egger, S. Schonborn, A. Forster, T. Vetter, Pose normalization for eye gaze estimation and facial attribute description from still images, in: German Conference on Pattern Recognition, 2014, pp. 317–327.

[136] S. J. Baek, K. A. Choi, C. Ma, Y. H. Kim, S. J. Ko, Eyeball model-based iris center localization for visible image-based eye-gaze tracking systems, IEEE Transaction on Consumer Electronics 59 (2) (2013) 415–421.

[137] K. A. F. Mora, J. M. Odobez, Geometric generative gaze estimation (g^3e) for remote rgb-d cameras, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1773–1780.

[138] K. A. F. Mora, J. M. Odobez, 3d gaze tracking and automatic gaze coding from rgb-d cameras, in: IEEE Conference on Computer Vision and Pattern Recognition, Vision Meets Cognition Workshop, 2014, pp. 4321–4322.

[139] C. Xiong, L. Huang, C. Liu, Gaze estimation based on 3d face structure and pupil centers, in: International Conference on Pattern Recognition, 2014, pp. 1156–1161.

[140] L. A. Jeni, J. F. Cohn, Person-independent 3d gaze estimation using face frontalization, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 792–800.

[141] S. Asteriadis, D. Soufleros, K. Karpouzis, S. Kollias, A natural head pose and eye gaze dataset, in: Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots (AFFINE '09), 2009.

[142] S. Asteriadis, K. Karpouzis, S. Kollias, Visual focus of attention in non-calibrated environments using gaze estimation, International Journal of Computer Vision 107 (3) (2014) 293–316.

[143] X. Xiong, Q. Cai, Z. Liu, Z. Zhang, Eye gaze tracking using an rgbd camera: A comparison with a rgb solution, in: ACM International Joint Conference on Pervasive and Ubiquitous Computing, 2014, pp. 1113–1121.

[144] A. Utsumi, K. Okamoto, N. Hagita, K. Takahashi, Gaze tracking in wide area using multiple camera observations, in: ACM Symposium on Eye Tracking Research and Applications, 2012, pp. 273–276.

[145] F. Lu, T. Okabe, Y. Sugano, Y. Sato, A head pose-free approach for appearance-based gaze estimation, in: Proceedings of the British Machine Vision Conference, 2011, pp. 1–11.

[146] U. Weidenbacher, G. Layher, P. Bayerl, H. Neumann, Detection of head pose and gaze direction for human-computer interaction, in: International Tutorial and Research Workshop on Perception and Interactive Technologies, 2006, pp. 9–19.

[147] E. Wood, T. Baltrusaitis, L. Morency, P. Robinson, A. Bulling, A 3d morphable eye region model for gaze estimation, in: European Conference on Computer Vision, 2016, pp. 297–313.

[148] Y. Sugano, Y. Matsushita, Y. Sato, Learning-by-synthesis for appearance-based 3d gaze estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1821–1828.

[149] C. C. Lai, Y. T. Chen, K. W. Chen, S. C. Chen, S. W. Shih, Y. P. Hung, Appearance-based gaze tracking with free head movement, in: International Conference on Pattern Recognition, 2014, pp. 1869–1873.

[150] Y. Liu, B. Lee, M. McKeown, A new reconstruction method in gaze estimation with natural head movement, in: International Conference on Machine Vision Applications, 2017, pp. 219–222.

[151] H. Cai, Y. Lin, An integrated head pose and eye gaze tracking approach to non-intrusive visual attention measurement for wide fov simulators, Virtual Reality, SI: Manufacturing and Construction 16 (1) (2012) 25–32.

[152] E. Wood, A. Bulling, Eyetab: Model-based gaze estimation on unmodified tablet computers, in: Symposium on Eye Tracking Research and Applications, 2014, pp. 207–210.

[153] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, A. Torralba, Eye tracking for everyone, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2176–2184.

[154] Q. Huang, A. Veeraraghavan, A. Sabharwal, Tabletgaze: Unconstrained appearance-based gaze estimation in mobile tablets, arXiv:1508.01244v3 [cs.CV] (2016).

[155] J. W. Lee, H. Heo, K. R. Park, A novel gaze tracking method based on the generation of virtual calibration points, Sensors 13 (2013) 10802–10822.

[156] B. Czupryński, A. Strupczewski, High accuracy head pose tracking survey, in: International Conference on Active Media Technology, 2014, pp. 407–420.

[157] A. Kar, P. Corcoran, Towards the development of a standardized performance evaluation framework for eye gaze estimation systems in consumer platforms, in: IEEE International Conference on Systems, Man, and Cybernetics, 2016, pp. 2061–2066.

[158] S. Cristina, K. P. Camilleri, Model-free non-rigid head pose tracking by joint shape and pose estimation, Machine Vision and Applications 27 (2016) 1229–1242.

[159] M. Khamis, A. Hoesl, A. Klimczak, M. Reiss, F. Alt, A. Bulling, Eyescout: Active eye tracking for position and movement independent gaze

interaction with large public displays, in: 30th Annual ACM Symposium on User Interface Software and Technology, 2017, pp. 1–12.

Table 1: Reviewed state-of-the-art methods that report the gaze estimation error, where: *Experimental conditions* indicates the level of head movement under which the methods have been tested; *HP* indicates whether the methods compensate for head movement, where ✓ refers to head pose estimation in 3-dimensional space, ≈ refers to a planar approximation of head movement, and × means that head movement is not compensated for; *Error* reports the gaze estimation error separately as (yaw°, pitch°) or as a single measurement that combines the two, and may include the standard deviation as $\pm\sigma°$; while *Data set* refers to a description of the data set in use as provided in Table 2.

| | Experimental conditions | | | | | | | | | | | |
| | Frontal and stationary head pose | | | | Natural/slight head movement | | | | Free head movement | | | |
Challenge	Ref	HP	Error (°)	Data set	Ref	HP	Error (°)	Data set	Ref	HP	Error (°)	Data set
A	[96]	×	6.91 ± 4.46	(i)	[94]	×	1.33 ± 0.13	○	[84]	×	3.55 ± 0.42	○
					[95]	×	(1.56, 2.57)	○	[84][1]	×	4.42 ± 0.55	○
					[97]	×	(1.46, 0.71)	○	[93]	×	(2.07, 2.48)	○
					[98]	≈	(0.54 ± 0.24, 0.42 ± 0.20)	○				
B	[87][2]	≈	0.59 ± 0.38 to 0.97 ± 0.57	○	[87]	≈	2.37 ± 1.42	○	[106][4]	✓	13.55	(ii)
	[104][3]	×	0.62 to 0.69	○	[105]	×	0.90	○	[106][5]	✓	11.12	(ii)
	[109]	×	7.50	(i)					[106][6]	✓	7.90	(ii)
	[110]	✓	10.90	○					[107]	✓	9.95	(ii)
									[110]	✓	16.80	○
C	[120]	×	6.00 ± 3.50	○	[12]	×	6.00	○	[117]	✓	2.90 ±2.10	○
	[116]	×	4.17	○	[115]	×	4.30	○	[118]	✓	2.56	○
	[123][7]	≈	2.57 ± 0.38	○	[122]	×	3.90	○	[119]	✓	4.06	○
					[128]	×	1.30	○	[126]	✓	(5.87 ± 3.73, 5.49 ± 3.45)	○, (iii)
					[131]	×	0.796	○	[129]	✓	<3.00	○
									[130]	✓	<3.00	○
D									[27]	✓	(4.60, 4.70)	○
									[126]	✓	(5.87 ± 3.73, 5.49 ± 3.45)	○, (iii)
									[134]	✓	9.90	○
									[135]	✓	9.74	(iv)
									[136]	✓	2.42 ± 1.06	○
									[137]	✓	6.30	○
									[139]	✓	3.33	○
									[140]	✓	4.2867	(i)
									[140]	✓	3.8925	(iii)
									[140][8]	✓	5.6265	(i), (iii)
									[143]	✓	4.40	○
									[144]	✓	(5.40, 9.70)	○
									[145]	✓	2.38	○
									[146]	✓	<10.00	○
									[147]	✓	9.44	(iii), (v)
									[149]	✓	(3.40, 4.49)	○
									[150]	✓	3.00 to 6.00	○
									[151]	✓	(2.10 ± 9.70, 5.90 ± 11.10)	○
E									[84]	×	3.55 ± 0.42	○
									[84][1]	×	4.42 ± 0.55	○
									[152]	×	7.00	○
									[154][9]	×	2.86 to 4.76	(vi)
									[154][10]	×	3.63 to 6.03	(vi)

[1] Face detection removed to improve the temporal resolution on a tablet device.
[2] Using four different training sets, containing 33, 23, 18, and nine samples.
[3] Using three different training sets, containing 33, 23, and 18 samples.
[4] Training on synthesised data set.
[5] Synthesised data set comprising expected head and eye poses only.
[6] Synthesised and real data comprising expected head and eye poses only.
[7] Using maximum number of considered on-screen fixation markers, 140.
[8] Frontal low-resolution face images with interocular distance of 50 pixels.
[9] Person-dependent; User-tablet distance varied between 30 to 50 cm.
[10] Person-independent; User-tablet distance varied between 30 to 50 cm.

Challenge A: Gaze Estimation from Low-Resolution Eye Images.
Challenge B: Gaze Estimation from Sparse, Synthesised or Person-Independent Training Samples.
Challenge C: Reduced or Implicit Calibration.
Challenge D: Head Pose Invariant Gaze Estimation.
Challenge E: Eye-Gaze Tracking on Mobile Platforms.

Table 2: Description of the data sets that have been used in testing the reviewed state-of-the-art methods of Table 1.

#	Data set	Description
○	Self-collected Data set	Test data collected by the authors themselves. The conditions under which different self-collected data sets are generated tend to be diverse, and these data sets usually feature variations in the characteristics of the participants and their backdrops, the illumination conditions, and the eye and head movements, as well as the imaging hardware by which the images are captured. Information regarding the conditions under which these data sets have been collected is not always provided by the authors.
(i)	Multi-view Gaze Data set	This data set contains a total of 64,000 images, captured via eight synchronised cameras while each of the 50 subjects in the data set gazed at 160 visual targets displayed on a monitor screen. The images have been captured at SXGA resolution at a camera-to-subject distance of 60 cm, while the head pose was stabilised by a chin rest. The gaze directions spanned approximately $\pm$ 25° horizontally and $\pm$ 15° vertically. The data set also includes the 3-dimensional positions of the visual targets and the monitor plane, together with the intrinsic and extrinsic camera parameters.
(ii)	MPIIGaze Data set	This data set contains 213,659 images collected from 15 participants during natural everyday laptop use. No constraints have been imposed on the environment and illumination conditions under which the images have been captured.
(iii)	Columbia Gaze Data set	This data set contains 5,880 images of 56 people, captured at a resolution of 5,184×3,456 pixels. Each subject features five head poses and 21 gaze directions per head pose, in front of a uniform background under controlled illumination conditions. The subjects are ethnically diverse and 21 of them wear glasses.
(iv)	UUlm HPG Data set	This data set contains 2,220 colour images of 20 subjects, captured at a resolution of up to 1600×1200 pixels. The participants perform various combinations of horizontal and vertical head poses and gaze directions, in front of a uniform background under controlled illumination conditions.
(v)	EYEDIAP Data set	This data set is composed of colour and depth images captured at a VGA resolution of 640×480 pixels and a frame rate of 30fps, together with images recorded at a full HD resolution of 1920×1080 pixels and a frame rate of 25fps. The data has been recorded from 16 participants, positioned in front of a uniform background under controlled illumination conditions. For three of these participants, two separate recording sessions were held on different days, and under different illumination and distance to the camera. The participants were instructed to maintain a stationary head pose while fixating at a moving visual target, or perform head movements while fixating at the same visual target.
(vi)	Rice TabletGaze Data set	The video sequences in this data set were captured by the front-facing camera on a Samsung Galaxy Tab S, at a resolution of 1280×720 pixels. Each of 51 participants recorded 16 video sequences while standing, sitting, slouching or lying, without being restricted on how they held the tablet, for a total of 816 videos. The group of participants was ethnically diverse and 26 of them wore prescription glasses. The videos were captured in a naturally lit office environment.

Table 3: Performance measures of the same method under different tracking conditions. *Challenge error* refers to the gaze estimation error under stationary, natural/slight or free head movement for the challenge under consideration, while $\overline{Challenge\ error}$ refers to the gaze estimation error achieved in absence of the challenge being addressed.

Challenge	Ref	Challenge error (°)			Challenge error (°)		
		Stationary	Natural/slight	Free	Stationary	Natural/slight	Free
A	[98][1]		(0.54 ± 0.24, 0.42 ± 0.20)				
A	[97][2]		(1.46, 0.71)				
B	[87]	0.59 ± 0.38 to 0.97 ± 0.57	2.37 ± 1.42				
B	[110]	10.90		16.80	7.60		12.90
C	[123]	2.57 ± 0.38	12.77 ± 5.71			2.37 ± 1.25	
C	[128]		1.30			0.80	

[1] With head movement compensation.
[2] Without head movement compensation.

Challenge A: Gaze Estimation from Low-Resolution Eye Images.
Challenge B: Gaze Estimation from Sparse, Synthesised or Person-Independent Training Samples.
Challenge C: Reduced or Implicit Calibration.