An End-to-End Review of Gaze Estimation and its Interactive Applications on Handheld Mobile Devices

YAXIONG LEI, University of St Andrews, UK
SHIJING HE, King's College London, UK
MOHAMED KHAMIS, University of Glasgow, UK
JUAN YE, University of St Andrews, UK

In recent years we have witnessed an increasing number of interactive systems on handheld mobile devices which utilise gaze as a single or complementary interaction modality. This trend is driven by the enhanced computational power of these devices, higher resolution and capacity of their cameras, and improved gaze estimation accuracy obtained from advanced machine learning techniques, especially in deep learning. As the literature is fast progressing, there is a pressing need to review the state of the art, delineate the boundary, and identify the key research challenges and opportunities in gaze estimation and interaction. This paper aims to serve this purpose by presenting an end-to-end holistic view in this area, from gaze capturing sensors, to gaze estimation workflows, to deep learning techniques, and to gaze interactive applications.

CCS Concepts: • **General and reference → Surveys and overviews**; • **Computing methodologies → Computer vision tasks**; • **Human-centered computing → Ubiquitous and mobile computing systems and tools**; **Interactive systems and tools**; **Interaction techniques**.

Additional Key Words and Phrases: Gaze Estimation, Domain Adaptation, Datasets, Eye Tracking, Eye Movement, Gaze Data Process, Gaze Interaction Methods, Gaze-based Interaction, Smartphones, Handheld Mobile Devices

ACM Reference Format:
Yaxiong Lei, Shijing He, Mohamed Khamis, and Juan Ye. 2023. An End-to-End Review of Gaze Estimation and its Interactive Applications on Handheld Mobile Devices. *ACM Comput. Surv.* 1, 1, Article 1 (January 2023), 38 pages. https://doi.org/10.1145/3606947

1 INTRODUCTION

Gaze interaction is to make use of gaze to facilitate interactions with computing devices, including virtual reality (VR) headsets, mobile or wearable devices, desktops, and robots. Gaze refers to a point on a screen or a direction in space and can be inferred from pupil positions, facial structure, and head movements. Through gaze, a system can sense users' attention [233] and enable touch-free interaction, which has driven a wide range of applications.

In a simulated surgical training task, gaze has been used to aid the identification of a target on a subject's laparoscopic screen in order to reduce mistakes and overcome language barriers [32]. In a driving situation, real-time gaze coding allows to track a driver's attention and detecting whether they are distracted or fatigue [106, 255], and support automated driving [170]. Gaze can also help diagnose mental health or autism by analysing the scan path [236]. Moving beyond, the gaze is also presented as a human-computer interface, facilitating controlling Internet of Things (IoT) devices in a smart home system [127], and empowering people with a physical impairment to interact with applications such as creative art [49, 140, 263].

Authors' addresses: Yaxiong Lei, yl212@st-andrews.ac.uk, University of St Andrews, St Andrews, UK, KY16 9SX; Shijing He, shijing.he@kcl.ac.uk, King's College London, London, UK; Mohamed Khamis, mohamed.khamis@glasgow.ac.uk, University of Glasgow, Glasgow, UK; Juan Ye, jy31@st-andrews.ac.uk, University of St Andrews, St Andrews, Fife, UK, KY16 9SX.

The advanced research in gaze interaction has inspired an increasing number of industrial applications, including optimising the system's scheduling of processing resources [166], facilitating the presentation of content [59, 126], marketing, and accessibility for people with disabilities [34, 58, 88, 267]. Smart Eye brings an AI-integrated eye-tracking technology to detect whether a driver is distracted [221], and Eyetech Digital Systems launched EyeOn to enable users to type and communicate using eye movement [55]. There are growing market players, including Tobii AB, LC Technologies, Eyetech Digital Systems, Ergoneers GmbH, Smart Eye AB., Mirametrix Inc., Pupil Labs GmbH, Seeing Machines, SR Research Ltd., and Gazepoint. The worldwide market for eye tracking technology is valued at USD 638.8 million and is anticipated to grow at a compound annual growth rate (CAGR) of 33.4% between 2022 and 2030 [76].

In recent years, mobile devices have advanced with significantly improved camera quality and computational power. These capabilities make it possible to explore gaze interaction on mobile devices. The advantage over traditional eye tracking is that gaze on mobile devices is derived from images or videos on cameras, which does not require extra devices. Industries are starting to explore novel gaze interactions on mobile devices. For example, Apple has paired facial recognition and eye movements to enhance the unlocking experience of FaceID and the Huawei Mate 40 Pro keeps the screen on when being gazed at.

What currently needs to be added to the literature is a comprehensive overview of gaze estimation to interaction on handheld mobile devices and a roadmap for advancing from low-level gaze points and directions to high-level gaze patterns, which play a significant role in many gaze applications. To close the gap, we describe the current technological advances in capturing gaze and a workflow along with deep learning algorithms to estimate gaze. We identify the challenges in enhancing the diversity of data collection and robustness of gaze estimation, especially reflecting the complexities of dynamic environments such as partial faces, varying lighting conditions, and constant changes of holding postures. These challenges come from the inherent characteristics of mobile devices: their size and mobility, which fundamentally differ from eye tracking on other platforms such as VR, desktops, and large displays. We also broadly review gaze-based applications, introduce eye physiology, and point to new types of gaze for future interaction design. Different from the existing surveys on gaze estimation, this paper makes the following key contributions.

- We present an end-to-end holistic review from sensors, to algorithms, and to application.
- We focus on appearance-based gaze estimation on handheld devices, and identify the unique challenges different from other platforms.
- We bridge the gap between gaze estimation, gaze data processes, and gaze interaction and present a pipeline for processing, analysing, and deriving high-level gaze events from raw gaze data.

2 METHODOLOGY

Our paper takes recent reviews and survey studies as an initial research methodology [31, 91, 111, 116, 118], with adaptation to suit our research aims; that is, understanding the workflow of gaze interaction from camera input to applications on handheld mobile devices. Our research methodology comprises four stages: defining keywords, paper collection, paper categorisation and collation, and paper analysis. We extract keywords based on the content in the papers mentioned in the related work, and eventually identify ("Eye" or "Gaze" or "Eye Tracking") and ("Gaze Estimation" or "Eye Tracking Technique" or "Eye Tracking algorithm" or "Algorithm" or "Dataset") for Section 4 and 5, ("Eye" or "Gaze" or "Eye Tracking" or "Eye Movement") and ("Data Analysis" or "Data Process" or "Application" or " Interaction") for Section 6 and 7. We select the conferences and journals in the areas of HCI, UbiComp, computer vision, and eye-tracking as our target sources, including ETRA, CHI, IJHCS, HCI, UbiComp/IMWUT, MobileHCI, IJHCI, CVPR, EMR, BRM, etc. We also search terms on Google Scholar to catch more publications.

We categorise papers based on research topics: gaze estimation algorithms, gaze data analysis, gaze interaction and gaze applications. We then sub-categorise them according to the platforms in which these items are applied, e.g., handheld mobile devices (mobile phones or tablets), desktop devices (desktop computers or laptops), and head-mounted devices (VR or glasses). Algorithms, interactions and applications that are currently used in various devices will be discussed in terms of their applicability to handheld mobile devices. Our discussion focuses on the characteristics of handheld mobile devices, i.e. the mobility of the device, the screen size, the application characteristics, and the usage scenario.

Based on the classification and year, we summarise the content of the collected papers, extract their motivation, methodology, subsequent research directions and their contribution to gaze estimation algorithms and applications for handheld mobile devices.

3 RELATED WORK

In recent years, more and more research projects on gaze are emerging, addressing the quest for performance and stability in gaze estimation, gaze interaction and gaze applications using the capabilities of commercial eye trackers or gaze estimation algorithms. The existing surveys and review studies have different perspectives and areas: algorithms and applications of gaze estimation. Kar and Corcoran [111] survey nearly 20 years of gaze estimation research on a variety of devices, including desktop, television, head-mounted, automotive and handheld, and also focus on methods for evaluating the performance of gaze tracking systems. Cheng et al. [31] provide a detailed review of appearance-based gaze estimation approaches. They build a pipeline of gaze estimation of deep learning. In terms of applications, Khamis et al. [118] summarise eye-tracking related studies on handheld mobile devices from 2002 to 2018. They list three main topics of gaze applications: gaze behaviour analysis, implicit gaze interaction and explicit gaze interaction. Katsini et al. [116] focus on the application of gaze in security and privacy, which include authentication, privacy protection and gaze monitoring during security critical tasks. They summarise the usage scenarios, devices and evaluation methods for these tasks. Hirzle et al. [91] conducted a study on existing gaze interaction projects around the Digital Eye Strain (DES), which has surveyed over 400 papers published in the last 46 years. Ghosh et al. [74] focus on the algorithms in AR/VR devices and the applications in healthcare and driver engagement. Nishan et al. [77] provides a detailed analysis of eye-tracking technology on mobile devices and present an edge computing solution for real-time eye-tracking experience.

Different from the above reviews, our paper aims to present an end-to-end overview of gaze estimation, data analysis and interaction, illustrating its workflow, mainstream deep learning techniques for estimating gaze, and a broad range of gaze interaction applications on handheld mobile devices.

4 GAZE ESTIMATION ON HANDHELD MOBILE DEVICES

This section presents an overview of gaze estimation process, including the task of gaze estimation (in Section 4.1), sensors common in handheld mobile devices for gaze estimation (in Section 4.2), and the workflow (in Section 4.3) including *pre-processing*, *learning*, *post-processing*, and *calibration*.

4.1 Task and Evaluation of Gaze Estimation

Gaze estimation refers to predicting a point of gaze (PoG) [31] from images or videos captured on the front camera of handheld mobile devices, as presented in Figure 1; i.e., a coordinate point $p = (u, v)$ on a screen; or gaze direction; i.e., a vector $\mathbf{g} = (g_x, g_y, g_z)$ in a 3D coordinate system [171]. The former is often referred to as 2D gaze while the latter as 3D gaze. In both cases, it is treated as a regression task, while it can also be treated as a classification problem [240]; that is, predicting which grid or area the user is looking at.

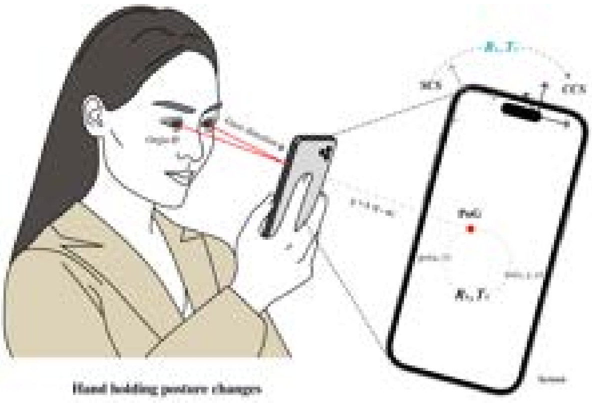


Fig. 1. The gaze estimation task involves predicting either a) a point of gaze (PoG) as a coordinate point on a screen (called 2D gaze) or b) a gaze direction in a 3D space (called 3D gaze) [31].

The accuracy of gaze estimation is defined as the average difference between the estimated gaze location and the location of the fixation target. In 2D, the difference is measured as the Euclidean distance between the true point $\boldsymbol{p}$ and estimated point $\hat{\boldsymbol{p}}$, where a point is commonly represented as a 2D coordinate in pixels with units of cm or mm [14, 25, 138, 184].

$$acc_{2D} = d = \|\boldsymbol{p} - \hat{\boldsymbol{p}}\| \tag{1}$$

The true points can be obtained via an extra high-resolution eye-tracking device or self-reported by participants; that is, indicate their gaze points using a cursor or touch [98, 108, 114, 156, 246]. For example, GazeCapture [138] designs an application to display a red dot on a screen of a mobile phone and requires participants to gaze at the dot and follow its movement. Before the dot moves, a small letter L or R is displayed for 0.05 seconds, which requires a participant to tap either the left (L) or right (R) side of the screen. This serves as a way to monitor participants' attention and thus validate the data collection. The ground truth points are the locations of the red dot.

In 3D, the difference in gaze direction refers to an angular distance between the true direction vector $\boldsymbol{g}$ and the estimated direction $\hat{\boldsymbol{g}}$, which is often represented in degrees.

$$acc_{3D} = \alpha^{\circ} = \frac{\boldsymbol{g} \cdot \hat{\boldsymbol{g}}}{\|\boldsymbol{g}\|\|\hat{\boldsymbol{g}}\|} \tag{2}$$

The collection of gaze direction is more complex, as the ground truth is a vector in a 3D space. Gaze360 [117] uses a 360° panoramic camera placed on a tripod in the centre of the scene, and a large moving rigid target board marked with an AprilTag and a cross on which participants are instructed to continuously fixate. The true gaze direction is derived from the camera coordinator system based on the distance between the eye, the camera, and the target board. Differently, ETH-XGaze [268] utilises 18 single-lens reflex cameras in a 3D space to gather ground truth while stimuli are displayed on a large screen in front of participants.

Commercial eye-tracking technologies like Tobii [229] also provide the other two measures: precision and data loss. Precision refers to the system's ability to produce the same gaze point or direction measurement reliably. It is calculated as the root mean square of accuracy on a sequence of consecutive data pairs of true and estimated gazes.

$$precision = \sqrt{\frac{\sum_{i=1}^{n} acc_i^2}{n}}, \tag{3}$$

where acc_i is the accuracy (2D or 3D) on the ith pair between the target location and the gaze (point or direction), referring to the above Equation (1) and (2).

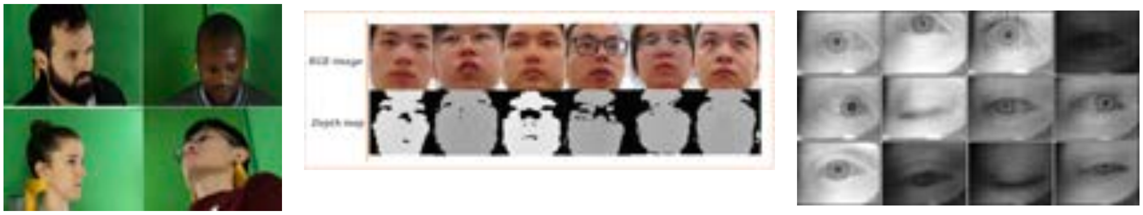
Data loss is defined as the ratio of gaze samples captured by the eye tracker during the fixation on a target. It is calculated by excluding invalid samples; that is, no gaze is detected.

$$data\_loss = \frac{\text{No. of invalid samples}}{\text{No. of total samples}} \tag{4}$$

Current gaze estimation algorithms typically focus on providing accuracy metrics and do not include precision and data loss measurements. However, for practical applications, it is important to consider all the above three metrics and provide them in their test report to ensure reliable and stable performance in real-world applications.

4.2 Sensors

Eye-tracking technologies have advanced significantly over the past 60 years [215], from electro-oculoGraphy (EOG) signals that are based on muscle action signals to visual signals from cameras. The recent examples include head-mounted devices [17] like Oculus VR or HoloLens 2, wearable eye trackers like Tobii Pro Glasses 3, and various types of screen-based eye trackers [185]. These devices come in a variety of shapes but generally tightly integrate the camera above or below a screen; for example, a camera at the top of a mobile phone. They have different eye-to-screen distances; for example, head-mounted eye trackers are close to the eye (e.g., within the distance of 0.5-3cm), while in the screen-based eye trackers (aka remote eye trackers), the distance is usually 50-80cm for desktop settings [195], and 15-50cm for handheld devices [100, 232]. The distance and the screen size of handheld devices may constrain what types of gaze interaction applications are more acceptable to users. In this survey, we focus on sensors on handheld mobile devices, which are RGB, RGB-D(epth), and Infrared (IR) cameras (see Figure 2).



(A) RGB from ETH-XGaze

(B) RGB-D from ShanghaiTechGaze

(C) NVGaze

Fig. 2. Examples of images from cameras that are readily integrated into handheld mobile devices, and can be used for gaze estimation: (A) RGB [268], (B) RGB-D [155], and (C) IR [124]

4.2.1 *RGB Camera.*

Currently, most mobile device-based gaze estimation methods are using RGB Cameras, which are widely supported. Over years, the quality of the cameras has improved, in terms of aperture, resolution, and frame rate.

4.2.2 RGB-D Camera. RGB-D camera adds depth information to RGB images; that is, each pixel relates to a distance between the object in the image and the image plane. RGB-D cameras are starting to be supported in modern smartphones, especially high-end phones. Depth information can assist in constructing a 3D model of head pose and provide information on the eye position, further improving gaze estimation [171, 253].

4.2.3 IR Camera. IR camera refers to a Near Infrared (NIR) camera [196] that uses an artificial IR light source aimed on- and off-axis at the eye, which introduces *glint*, called corneal reflection [161]. The glint acts as a reference point and the gaze direction is calculated by measuring the changing relationship between the glint and the moving pupil centre of the eye. IR camera is sensitive to wavelengths from 700 nanometers (nm) to 1,400 nm, and this band does not affect human vision beyond the range of visible spectrum [225], so it is often used as near-eye cameras; e.g., head-mounted augmented and virtual reality (AR/VR) devices [252]. Recently, there are more and more mobile phones that are configured with NIR cameras, including iPhone, Huawei, Samsung, Xiaomi and OPPO.

4.3 Workflow

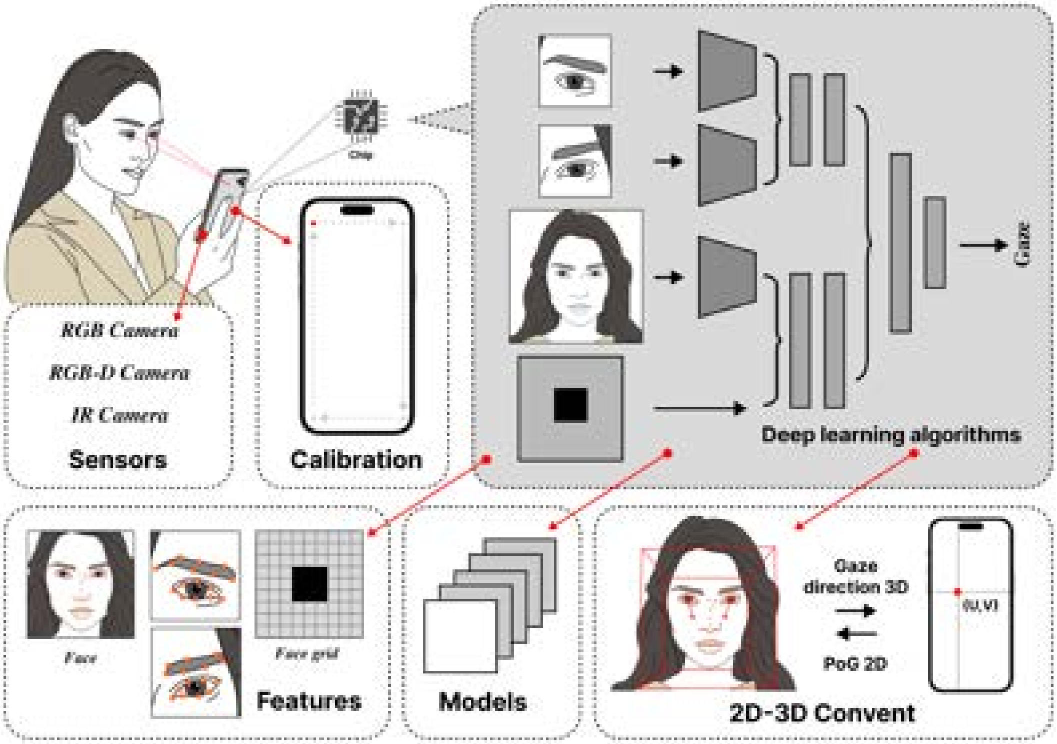


Fig. 3. Workflow of gaze estimation based on camera

As presented in Figure 3, the workflow of gaze estimation on handheld mobile devices consists of four main stages: (1) *pre-processing* – rectifying and processing images or videos acquired from a camera to identify a face and extract features; (2) *learning* – applying a machine learning technique to estimate a gaze point or vector;

(3) *post-processing* – converting the gaze output to suit the requirements of applications; and (4) *calibration* – calibrating the gaze estimation model to cater for the characteristics and context of a new environment, device, and user. In the following, we will illustrate each of the above stages.

4.3.1 Pre-processing and Feature Extraction. At the pre-processing stage, images are collected from a camera, from which a face can be identified and facial landmarks are extracted. Facial landmarks, referred to as a set of coordinates on an image, are used to locate and represent important regions of the face, including the chin, mouth, nose, eyes, and eyebrows. OpenFace [12] and Dlib [128] are the most commonly used libraries for face identification and landmark extraction [154, 271, 272]. Recently, researchers are starting to design customised networks for more accurate and richer facial feature extraction, including Multitask Cascaded Convolutional Network (MTCNN) [264], Deep Alignment Network (DAN) [136], Position Map Regression Network (PRNet) [60], and 3D Dense Face Alignment (3DDFA) [279]. Significant effort has also been dedicated to pupil segmentation and detection, including PupilNet [67], Circular Binary Features (CBF) [64] and Boosted-Oriented Edge optimisation (BORE) [63] for real-time pupil detection.

Once facial landmarks are extracted, face and/or eye images can be cropped. A common way to do so is to crop a square region with the centre of the face and a width. The centre of the face is the averaged coordinate positions of all the facial landmarks and the width can be set as a ratio to the maximum distance between the landmarks; for example, Zhang et al. [272] set the ratio as 1.5. As head pose has a significant impact on gaze estimation, often the images need to be rectified, including rotating and shifting to align with a reference point [31]. With cropped face and eye images, various features can be extracted, including eye, face, face and eye, and temporal.

Eyes. The eye is intrinsically connected to gaze estimation, as any change in the gaze direction leads to a corresponding alteration in the eye's appearance. For instance, eye rotation affects the iris's position and the eyelid's shape, resulting in a shift in the gaze direction. This connection enables gaze estimation based on the eye's appearance. Many methods leverage both eyes as cascading features for gaze estimates, with examples including, Minst [271], SAGE [82], GoogleGaze [232], EyeNet [185], and EVE-SCPT [13]. However, for gaze estimation from handheld mobile devices, the visibility of both eyes only has 68.18% [119], so monocular features remain valuable and are used by DPGE [187], Deng et al. [278] and OneEye-Net [5].

Face. Facial images provide valuable information on the head pose, which is beneficial for gaze estimation. Numerous methods utilise facial features as input for their models, such as FullFace [272], GazeTR [29], L2CS-Net [1], GazeCLR [109], SE-Gaze [178], FreeGaze [44], GazeNAS [174], and Gaze360 [117]. However, facial images can also contain redundant information. Researchers, as seen in PureGaze [28] and other models [179, 270, 272], have attempted to filter out irrelevant features in facial images, enabling the model to focus on the core facial features that are common to all. This optimisation mechanism implicitly eliminates gaze-irrelevant features, thus enhancing the gaze estimation network's robustness. Additionally, facial landmarks are employed as supplementary features to model head pose and eye position.

Face and Eyes. In order to have robust features, some studies have employed both eye and face features as input to have robust features, with these networks typically being multi-streamed to obtain information from the face and eyes simultaneously. Facial landmarks and face grid are also utilised as additional features to model for head pose, eye position and spatial information, often used in conjunction with eye and face features. Examples of such methods include iTracker [138], GazeAttentionNet [96], Dilated-Net [26], TAT [78], AFF-Net [14], iMon [102], RecurrentGaze [181],GAZEL [184], RT-Gene [62], and HAZE-Net [258].

Static and Temporal Features. Features obtained from images are static; however, the information captured by a camera is dynamic, with a correlation between frames. To improve the robustness on prediction, some studies [102, 117, 181, 185] investigate the temporal information by extracting features from video and learning their temporal correlations between frames. Temporal features, such as optical flow [102] and eye movement dynamics [241], have been used to improve gaze estimation accuracy. Optical flow provides information about motion between frames. Gaze360 [117], RecurrentGaze [181], EyeNet [185] and EVE-SCPT [13] directly apply models such as LSTM to obtain correlation information between video frames.

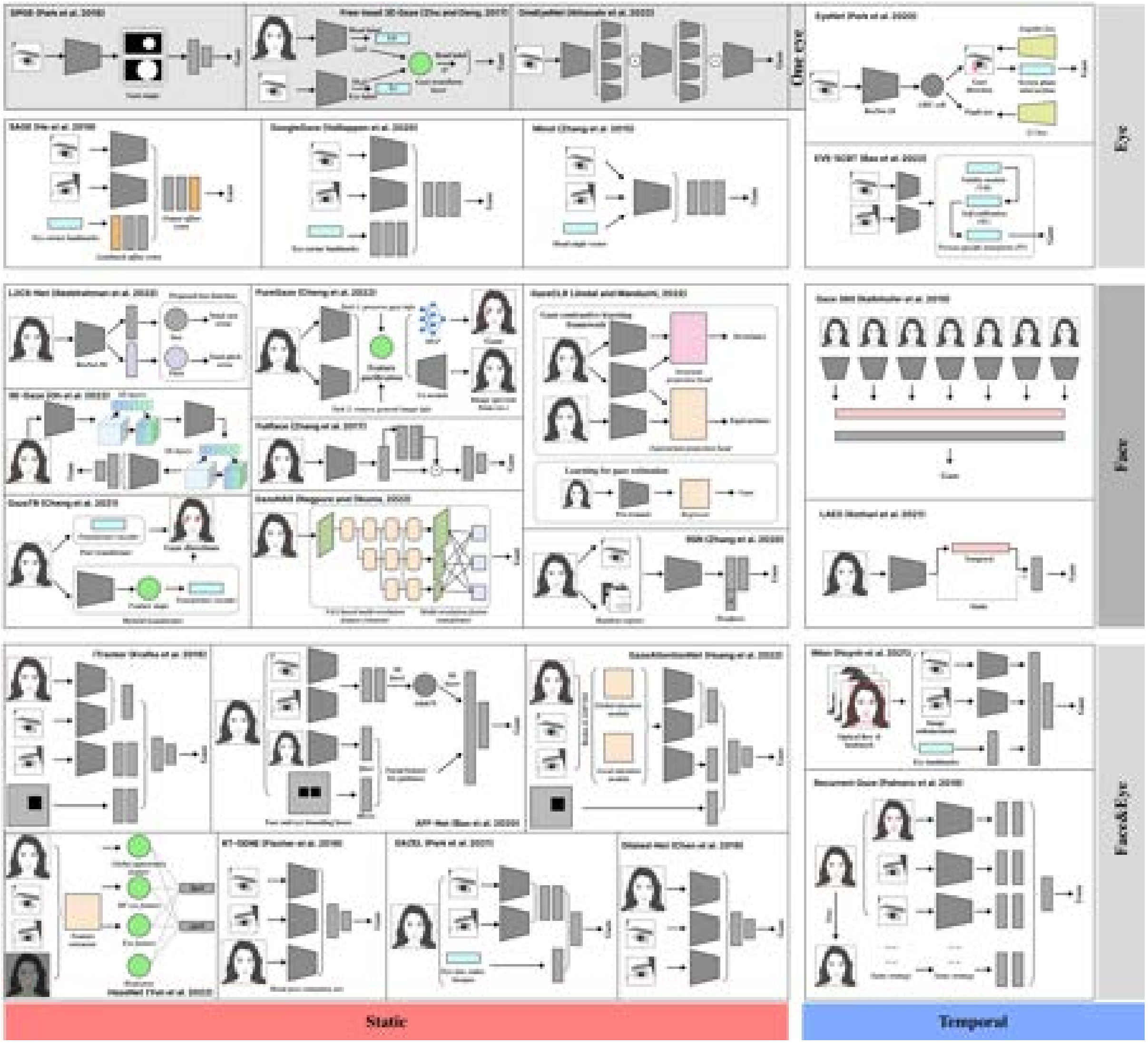


Fig. 4. Overview of the deep learning architectures for Gaze Estimation 2D & 3D tasks

4.3.2 Learning. Table 1 and Figure 4 present the mainstream deep learning models applied to gaze estimation, which are built on Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and more recently Transformers.

Convolutional Neural Networks. CNN is one of the most commonly used techniques for extracting facial and eye features. In gaze estimation, VGG [62, 218, 271, 273], ResNet [83, 117, 268], LeNet [146], AlexNet [141], MobileNet-V2 [25] and DenseNet [94] are the mostly adopted state-of-the-art architectures. Researchers also design novel architectures for face and eye feature extraction. Zhang et al. [272] have proposed a customised CNN with spatial weights for full-face 2D and 3D gaze estimation. The first few layers of the CNN extract image features, and then the last few layers are dedicated to learning spatial weights for the activation of the last convolutional layer. The purpose of the spatial weights is to indicate the importance of different regions of the face for gaze estimation.

Features from eye images can be combined with the head pose, and the eye features can be explored independently (with each of the two eyes) or jointly (with two eyes together). Fischer et al. [62] have applied two VGG-16 networks [218] to learn individual features from two eye images and then concatenate these features to predict the yaw and pitch gaze angles. Cheng et al. [30] have proposed an asymmetric regression evaluation network where a four-stream CNN is applied to extract and combine features on individual eye images, which are used in linear regression to estimate the 3D gaze direction.

Recurrent Neural Networks. RNN [208] and Long short-term memory (LSTM) [92] have been employed to explore and leverage the temporal relationships between consecutive frames to improve the estimation accuracy. Palmero et al. [181] have proposed a Recurrent CNN Regression Network for 3D gaze estimation. The network is composed of 3 modules: individual, fusion, and temporal. The individual module learns features from each appearance cue independently and consists of a two-stream CNN: one for processing face images and the other for jointly learning eye images. The fusion module combines the extracted features of each appearance stream in a single vector along with the normalised landmark coordinates. Then, it learns a joint representation between modalities in a late-fusion fashion. Both individual and fusion modules are applied to each frame, and the resulting feature vectors of each frame are input to the temporal module based on a many-to-one recurrent network. This module leverages sequential information to predict the normalised 3D gaze angles of the last frame of the sequence using a linear regression layer.

Bidirectional LSTMs are another popular choice for modelling temporal relationships between successive frames [117, 277]. Kellnhofer et al. [117] have utilised 7 consecutive sequence frames to obtain continuous head and eye movements in the video stream to improve the accuracy of predicting the gaze of the central frame in the input sequence. Specifically, a head crop from each frame is individually processed by ResNet-18, and the extracted features are fed to a bidirectional LSTM with two layers that take the sequence of forward and backward vectors. Finally, these vectors are concatenated and passed through a fully connected layer to produce two outputs: gaze prediction and error quantile estimation.

Combining the gaze direction with the optical flow in the eye region is another way of exploiting temporal information in motion. Bace et al. [8] have combined gaze direction with the optical flow in the eye region to identify the target that a user is following. The system consists of two components: one for taking a facial image as input and predicting 2D gaze points, and the other for estimating the motion pattern between consecutive frames. Aggregating the outputs from these two components can improve the robustness of estimation.

Transformers. Transformer, originating from natural language processing [234], contains self-attention layers, layer normalisation and multi-layer perception layers. Compared to RNN, self-attention layers have global computations and better memory to process long-sequence tasks. Transformers also perform well in computer vision tasks by replacing words in NLP tasks with non-overlapping image patches, called vision transformer (ViT) [39]. GazeTR [29] is an early ViT adopter in the field of gaze estimation. They apply the original ViT architecture to gaze estimation and also propose a hybrid ViT architecture combined with a CNN feature extractor. The hybrid ViT architecture achieved better results than ViT alone on various datasets. GazeNAS [174] is another early adaptor of ViT, where they propose a light-weighted ViT that only has 1 million parameters and uses 0.28 GFLOPs. It involves neural architecture search (NAS) from reinforcement learning as a multi-resolution feature extractor. GazeNAS achieves state-of-the-art results on various benchmark datasets.

Semi-supervised and Unsupervised Learning. Most deep learning algorithms require a large number of training data and collecting high-quality gaze datasets is a time- and effort-consuming task. Therefore, in recent years, we have witnessed more and more semi-supervised and unsupervised learning techniques being developed to reduce the reliance on labelled data. Kothari et al. [134] have proposed to curate videos from the Web where people are “looking at each other" (LAEO), and annotate each frame with whether LAEO is present. They design a weakly supervised algorithm for learning 3D gaze information by enforcing scene-level geometric 3D and 2D LAEO constraints between pairs of faces. Dubey et al. [45] propose an unsupervised learning technique based on a large in-the-wild dataset that contains many facial images from the Web. They localise the pupil-centre of each eye and use them to determine the region in which the subject is looking. Similarly, Yu et al. [257] also propose to learn gaze representation from unannotated eye images and then use a few labelled calibration samples for gaze estimation.

4.3.3 Post-processing. Depending on the available data for training gaze estimation algorithms and the requirements of the applications, there is a need to convert between 2D and 3D gaze. The conversion is performed by rotating and translating the camera coordinate system (CCS) and screen coordinate system (SCS) [31]. To convert a 2D gaze point to a 3D gaze direction, we first obtain the rotation R_s and translation T_s matrices of SCS with respect to CCS by geometric calibration. With these two matrices, we can compute the 3D gaze target with $t = R_s[u, v, 0]^T + T_s$, which is the intersection of gaze direction and the screen. The target t will be used to derive the 3D gaze direction $g = k(t - o)$, where k is the factor $\frac{1}{||t-o||}$ and o is the gaze origin; for example, the face or eye centre.

To convert a 3D gaze direction to a 2D point, we will still require R_s and T_s and the gaze origin and then revert the calculation process. First, we need to calculate the 3D gaze target vector $t = (x, y, z)$ which is the interaction of gaze direction to the screen. To do so, we use two equations: the line of sight and the screen plan. Given the origin $o = (o_x, o_y, o_z)$ and a 3D gaze direction $g = (g_x, g_y, g_z)$, we can get the equation of the line of sight as

$$\frac{x - o_x}{g_x} = \frac{y - o_y}{g_y} = \frac{z - o_z}{g_z}. \tag{5}$$

The equation of the screen plane is the relation between the target vector t and the rotation matrix R_s and the translation matrix T_s. From R_s, we can derive a normal vector of screen plane $n = R_s[:, 2] = (n_x, n_y, n_z)$. Given $T_s = [t_x, t_y, t_z]^T$, we can deduce the equation of screen plane as

$$n_x x + n_y y + n_z z = n_x t_x + n_y t_y + n_z t_z. \tag{6}$$

Solving the above equations 5 and 6 gives us the target vector t. Then we can compute the corresponding 2D gaze point as $p = (u, v, 0) = R_s^{-1}(t - T_s)$.

With handheld mobile devices, most of the existing interactive systems use 2D gaze points. The reason is that the holding posture of devices often changes, and the relation between the eye and the screen is not stable, which either leads to inaccurate calculation of the direction or requires constant recalibration of the rotation and translation matrices. Nowadays, research effort is increasingly devoted to 3D gaze direction; for example, Zhang et al. have set up a multi-camera system to capture over 500 gaze directions with various illumination conditions [268]. 3D gaze directions depending on the coordinate system of the head and face may improve robustness for mobile devices in complex environments.

4.3.4 Calibration. Eye-tracking systems often need to be calibrated; otherwise, the gaze output might incur too large errors to be useful. Calibration is the process of adjusting and customising the gaze output to reflect

Table 1. A list of representative appearance-based gaze estimation models

Attributes Methods	Year	Feature	Model	Dataset	PoG(cm)		Direction(°)
					Tablet	Phone	
Minst [271]	2015	Eyes, Lmks	CNN	MPIIGaze	-	-	6.27°
iTracker [138]	2016	Face, Eyes, Grid	CNN	GazeCapture	2.81	1.86	-
FullFace [272]	2017	Face	CNN	MPIIFaceGaze	-	-	4.80°
Dilated-Net [26]	2018	Face, Eyes	CNN	MPIIGaze	-	-	5.12°
RT-Gene [62]	2018	Face, Eyes	CNN	MPIIGaze	-	-	4.66°
DPGE [187]	2018	Eye	CNN	MPIIGaze	-	-	4.50°
RecurrentGaze [181]	2018	Face, Eyes, Lmks	CNN+LSTM	EyeDiap	-	-	3.38°
Gaze360 [117]	2019	Face	CNN+LSTM	Gaze360	-	-	11.1°
SAGE [82]	2019	Eyes, Lmks	CNN	GazeCapture	2.72	1.78	-
TAT [78]	2019	Face, Eyes	CNN	GazeCapture	2.66	1.77	-
RSN [270]	2020	Face	CNN	MPIIGaze	-	-	4.50°
GoogleGaze [232]	2020	Eyes, Lmks	CNN	GazeCapture	-	1.92	-
EyeNet [185]	2020	Eyes	CNN+GRU	EVE	3.85		3.48°
AFF-Net [14]	2021	Face, Eyes	CNN	GazeCapture	2.30	1.62	-
iMon [102]	2021	Face, Eyes	CNN	GazeCapture	1.94	1.49	-
GAZEL [184]	2021	Face, Eyes, Lmks	CNN	Private dataset	2.91	-	-
PureGaze [28]	2021	Face	CNN+SA+MLP	ETH-XGaze	-	-	4.50°
GazeTR [29]	2021	Face	ViT	MPIIFaceGaze	-	-	4.00°
EVE-SCPT [13]	2022	Eyes	CNN+GRU	EVE	2.75		2.49°
GazeAttentionNet [96]	2022	Face, Eyes, Grid	CNN+MLP	ETH-XGaze	-	-	4.5°
L2CS-Net [1]	2022	Face	CNN	Gaze360	-	-	9.02°
GazeCLR [109]	2022	Face	CNN	EVE	-	-	4.15°
OneEye-Net [5]	2022	Eye	CNN	GazeCapture	2.31		-
SE-Gaze [178]	2022	Face	CNN+SE+MLP	MPIIFaceGaze	-	-	4.04°
FreeGaze [44]	2022	Face	CNN	ETH-XGaze	-	-	2.95°
HAZE-Net [258]	2022	Face, Eyes, Lmks	CNN	EyeDiap	-	-	4.12°
GazeNAS [174]	2023	Face	ViT	MPIIFaceGaze	-	-	3.96°

the spatial geometry of the camera, the screen, and personal difference [42] in order to improve the estimation accuracy. The calibration process consists of *data collection* and *training*.

Data collection. The common way to collect ground truth data for calibration is to design an interactive interface to guide a user's gaze attention. The point-based method is the most used example, where users are asked to fixate at a point for a few seconds. There are often between 5 and 16 points being displayed consecutively at various locations on a screen [42]. Another popular method is pursuit-based, where users are asked to follow their gaze on a moving target, and the trajectory of the movement can be any shape such as a circle [42] or a rectangle [147]. Compared to the point-based method, the pursuit-based method requires less time to collect the same number of data points and provides a better user experience.

The above data collection is considered as *explicit*, requiring the users' attention and voluntary actions. In contrast, *implicit* calibration collects ground-truth data from the background by estimating user attention in possible fixation locations; for example, mouse and keyboard events [98, 114, 156], typing on the on-screen keyboard and screen touch events [108, 246]. A recent study [256] employs visual saliency, which has distinctive

Table 2. A List of Calibration Methods on Handheld Mobile Devices

Attributes Project	Year	Data Collection		Calibrator	PoG(cm)		Direction(°)
		Explicit	Implicit		Tablet	Phone	
iTracker [138]	2016	13-Point	-	SVR	2.12	1.34	-
FAZE [186]	2019	-	-	MAML	-	-	3.08°
GoogleGaze [232]	2020	Points	-	SVR	-	0.46	-
GazeRefineNet [185]	2020	-	Visual Saliency	-*	2.75		2.49°
GazeL [184]	2021	-	Touch Event	SVR	-	1.58	-
vGaze [256]	2021	-	Visual Saliency	Clustering + LR	-	1.51	-
DAGE [149]	2021	9-Point	-	MLP	2.43	1.58	-
iMon [102]	2021	5-Point	-	Kappa angle+LR	1.59	1.11	-
EasyGaze [27]	2022	9-Point	-	LR	-	-	1.93°
DynamicRead [147]	2023	Pursuit	-	SVR	-	0.95	-

Note: SVR - Support Vector Regression; MAML - Model-Agnostic Meta-Learning; MLP - Multi Layer Perceptron; LR - Liner Regression without a specific model; * - GazeRefineNet is a label-free PoG refinement model that employed visual saliency.

perceptual properties from their surroundings that attract users' gaze attention [4, 130]. For example, independent continuous objects such as a sailing speedboat and a bright moon in the dark in the video are used as visually salient locations for collecting calibration points implicitly [256].

Calibrator Training. Table 2 presents a list of calibration techniques. Earlier calibrators are often adaptive linear regression techniques [159]. Recently, domain adaptation techniques have been applied to tackle the personalised calibration problem. For example, Cui et al. [35] have applied Geodesic Flow Kernel (GFK) to adapt the gaze estimator trained on adult data (the source domain) to predict gaze on children (the target domain).

With deep learning, transfer learning has been widely attempted; that is, either fine-tuning the fully connected layer from a pre-trained CNN model [265] or taking the features extracted on a CNN to train a Support Vector Regression (SVR) [138, 147, 151, 232]. These techniques are sensitive to environmental changes [138, 159]. To overcome this problem, Wang et al. [242] introduce a Bayesian adversarial learning technique in which an adversarial learning block is employed to learn generalisable gaze features to various appearances and head poses.

The FAZE project [186] applies a few-shot learning technique, Model-Agnostic Meta-Learning (MAML), to tailor a gaze estimation model to an individual only with a few calibration samples. It first learns robust features via an encoder-decoder architecture, which captures latent features on head orientation, gaze direction, and the appearance around the eye regions. Then it fine-tunes the model with a few calibration samples on individuals. Often the size of these samples is small and thus leads to an overfitting problem. To resolve this problem, MAML is adopted, which employs 2-step gradient updates to learn the optimal weights on the person-specific models. It has the advantage of minimising the generalisation loss of the network [186].

To calibrate the estimation model for individual users, there are efforts from the interaction perspective; that is, using the relevant eye movements generated during the interaction for calibration [114] and accuracy maintenance [97]. We will discuss this practice in Section 8.

4.4 Summary

From the analysis in Section 4.2, we observe that mobile devices are increasingly integrating advanced cameras, such as RGB-D and IR cameras, facilitating the development of appearance-based gaze estimation on mobile devices. As shown in Table 1, numerous deep learning techniques have been employed to enhance gaze estimation precision, with the input for gaze estimation evolving from facial landmarks to eye landmarks, and eventually to original face and eye images. The evolution is driven by the challenge of changing holding posture of devices and rotation of heads. For example, full-face images have been used to detect head orientation, eye position, eyelid openness, and eyebrow movement. Such information can be used in conjunction with features extracted from face images. In addition, gaze estimation is highly correlated with eye appearance. Subtle muscle changes in the eye area can lead to a change in gaze direction. Intuitively, the up-and-down movement of the eye drives the interplay of surrounding muscles, such as the eyelids and iris. Therefore, eye appearance features from the whole frame are increasingly employed in gaze estimation.

Most of deep learning models are built on state-of-the-art CNNs or custom CNNs, while recurrent models and ViT have not demonstrated significant improvements over CNNs. As detailed in Tables 1 and 2, the accuracy of gaze estimation can reach 1.49 cm on mobile phones, this level can support coarse-grained gaze interactions on handheld mobile devices, considering that widget sizes typically range from 0.9 cm to 1.2 cm. With calibration, the precision can be further improved; for example, Lei et al. have applied SVR for pursuit calibration and reduced their errors to 0.95cm, which enables various types of gaze interfaces to support scrolling actions in a reading application [147]. Much of the current research prioritises enhancing gaze estimation accuracy in static conditions and fixed postures. Future research should focus on achieving high-accuracy eye-tracking in dynamic real-world scenarios, where users interact with their devices in various natural postures [147, 175].

5 DATASETS

Datasets are an important research aspect in gaze estimation, and we have listed and reviewed all the publicly available gaze datasets in Table 3. Most of these datasets are collected on the RGB cameras alone (17 out of 24), 3 of them are on RGB-D, and 4 are on IR. There is a trend of moving the platform and collection conditions from more controlled, desktop-based environments towards unconstrained, mobile device-based settings. In the early stage, gaze data is often collected in the laboratory environment and subjects are required to fix their head on a chin rest [70, 268]. Figure 5 has shown the collection settings and samples of the gaze datasets. More recently, research attention is gradually shifting to real-world settings with a variety of distances to screens, head poses, and illumination conditions. For example, MPIIGaze [273] consists of over 213,000 images from 15 people looking at different gaze positions. The dataset is collected over three months during daily laptop usage. GazeCapture [138] is one of the pioneers in collecting gaze data in in-the-wild environments, containing over 1.4 million images from over 1400 users. TEyeD [65] is the currently largest public dataset of eye images, which contains over 20 million real-world eye images with gaze vector, eye movement types, and pupil and eyelid. The data is collected when users wearing head-mounted eye trackers are performing various tasks, including outdoor sports, daily indoor activities, and car riding.

We have included the datasets collected on both desktop and mobile devices, even though our focus is on mobile device-based gaze estimation. The reason is that the desktop-based datasets can be used to train a deep learning model for extracting face and eye features. In addition, other benchmark datasets can also be used to train a deep-learning model. For example, the pupil and eyelid datasets [66, 68], face image datasets including VGG-Face [188], Labeled Faces in the Wild dataset (LFW) [95] and YouTube Faces (YTF) [248], Celebfaces [157], BayesianFace [158] can be used to extract facial features, and thus for gaze estimation [22].

6 GAZE ANALYTICS

To design gaze-based applications, we often need to process gaze and analyse them into high-level eye movement patterns. This is currently under investigated in the handheld gaze estimation research area. To bridge the gap, this section will first introduce eye physiology to provide a background on the meaning and function of eye

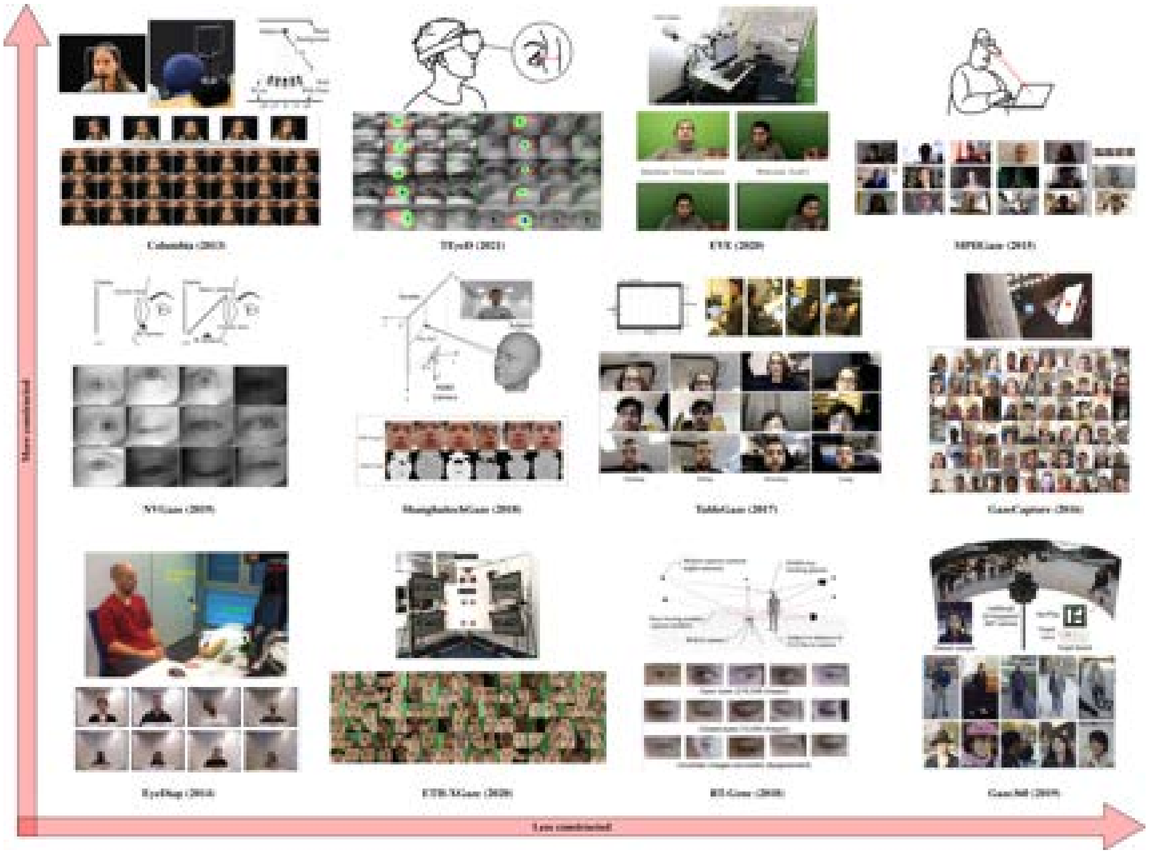


Fig. 5. Gaze dataset examples and their collection settings, including Columbia [222], EyeDiap [70], MPIIGaze [271], GazeCapture [138], TableGaze [101], ShanghaiTechGaze+ [155], RT-GENE [33], Gaze360 [117], NVGaze [124], EVE [185], and ETH-XGaze [268]

movements, and then illustrate how to process gaze data and define high-level gaze events and eye movement patterns.

6.1 Understanding of Eye Movement

Eye physiology plays a crucial role in gaze estimation and guides the development of gaze-interactive applications. Grasping the principles of eye movements and their relation to human consciousness levels is essential when designing ergonomic applications. Humans gather information about the external environment through their eyes, which involves continuous voluntary or involuntary movements, enabling the eyes to obtain a steady and continuous visual stimulus. These eye movements can be categorised into various types, such as fixation and saccades, as illustrated in Figure 6

(1) *Microscopic eye movements* encompass tremor, microsaccade, and drift, which underlie eye movements like fixation, saccade, and smooth pursuit [164]. Tremor is a periodic, wave-like eye movement with a frequency of 90-105Hz and is the smallest eye movement. It has potential for visual perception [2, 164, 165]. Microsaccades are small, fast, jerk-like eye movements that occur during voluntary fixational eye

Table 3. A summary of publicly available gaze datasets

Dataset	Year	Camera	Gaze	Head Move	IC	Distance	Sub	Resolution	Images
Columbia [222]	2013	RGB	2D, 3D	5	1	200cm	56	5184×3456	5,880
UT Multiview [228]	2014	RGB	2D, 3D	8	1	60cm	50	1280×1024	64,000
EyeDiap [70]	2014	RGB-D	2D, 3D	C	2	80-120cm	16	1920×1080	62,500
OMEG [84]	2015	RGB	3D	3 + C	10	varying	50	1280×1024	44,827
SynthesEyes [249]	2015	RGB	3D	C	4	varying	10	120×80	11,382
MPIIGaze [271]	2015	RGB	2D, 3D	C	D	40-60cm	15	1280×720	213,659
GazeFollow [200]	2015	RGB	3D	C	D	varying	130,339	variable	122,143
GazeCapture [138]	2016	RGB	2D	C	D	varying	1474	640×480	2,445,504
UnitEyes [250]	2016	RGB	3D	C	D	0.5-3cm	N/A	400 × 300	1,000,000
MPIIGazeFace [272]	2017	RGB	2D, 3D	C	D	varying	15	1280×720	37,639
TabletGaze [101]	2017	RGB	2D	C	1	30-50cm	51	1280 × 720	1,785
InvisibleEye [231]	2017	RGB	2D	N/A	1	0.5-2cm	17	5 × 5	280,000
RT-GENE [33]	2018	RGB-D	3D	C	1	80-280cm	15	1920×1080	122,531
Gaze360 [117]	2019	RGB	3D	C	D	varying	238	4096×3382	172,000
NVGaze [124]	2019	IR	2D	1	1	0.5-3cm	30	1280*960	4,500,000
SHTechGaze [154]	2018	RGB	2D	C	D	varying	137	1920×1080	233,796
SHTechGaze+ [155]	2019	RGB-D	2D	C	D	varying	218	1920×1080	165,231
EVE [185]	2020	RGB	2D, 3D	C	1	varying	54	1920×1080	12,308,334
ETH-XGaze [268]	2020	RGB	3D	C	16	100cm	110	6000×4000	1,083,492
GW [135]	2020	IR	3D	C	D	0.5-3cm	19	1920×1080	5,800,000
LAEO [134]	2021	RGB	3D	C	D	varying	485	variable	800,000
GOO [230]	2021	RGB	3D	C	D	varying	100	variable	201,552
OpenNEEDS [52]	2021	IR	3D	C	1	0.5-3cm	54	128×71	2,086,507
TEyeD [65]	2021	IR	2D, 3D	C	1	0.5-3cm	132	variable	20,867,073

Note: The columns of the above table are: (1) the publication *year*; (2) the type of *camera*; (3) type of *gaze* (2D PoG or 3D direction); (4) the number of variations in *Head Move* in reference to the screen, and *C* refers to continuous head movements; (5) illumination conditions (*IC*) where a number in this column refers to the number of illumination conditions being considered, *D* refers to the ambient light in normal daytime conditions; (6) the *distance* to the camera(s); (7) the number of *Sub*jects; (8) the resolution of each image; and (9) the number of *images*.

movements [164, 204, 209]. They can serve as a signal for fixation detection algorithms [53] and rotate around the point of fixation with small amplitudes [53]. Drift and tremor typically occur simultaneously and between microsaccades.

(2) *Fixation* stabilises the retina on a stationary object of interest to precept and process detailed information from the focused area. This eye movement is essential for tasks that require high visual acuity, such as reading or observing fine details. During fixation, microsaccades occur 1-2 times per second, it help to prevent the fading of the retinal image and maintain visual perception during fixation. [53, 165].

(3) *Saccades* are rapid, ballistic eye movements that occur between fixations, enabling the eye's fovea to continuously locate and track new objects in the visual filed [47]. As the fastest eye movements, saccades have speeds ranging from 30 to 900 degrees per second, influenced by factors such as target distance, amplitude, and individual differences. Saccades play a crucial role in tasks like reading, environmental scanning, and searching objects of interest [47, 209].

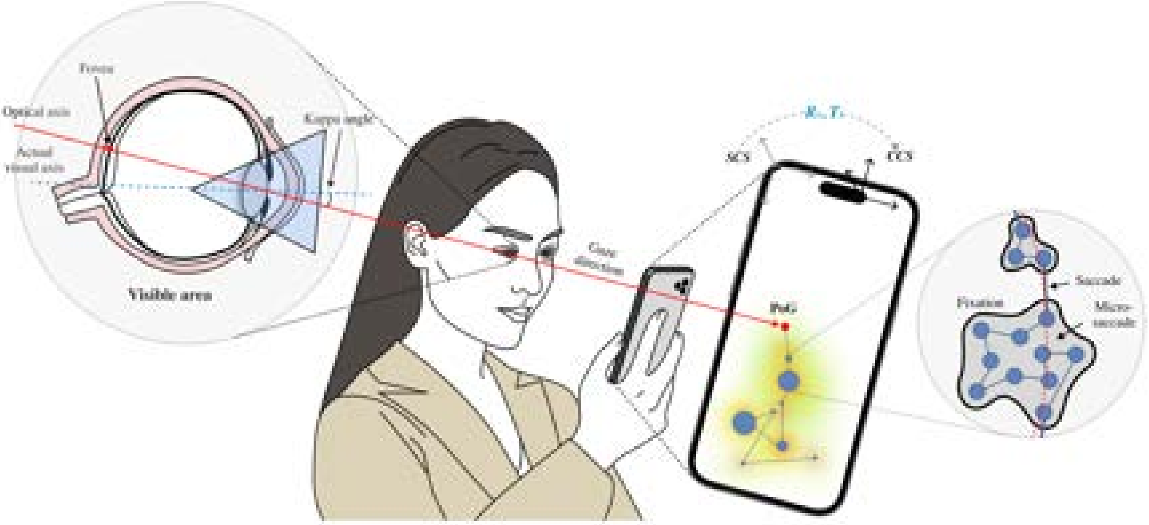


Fig. 6. From eye physiology to basic gaze events

(4) *Smooth Pursuit* is the action of stabilising the gaze on a moving visual target. Smooth pursuit has three phases: initiation, maintenance, and termination [69]. In the initiation phase, a delay of 100-130ms occurs between the target's movement and the start of smooth pursuit, with the first 100ms of smooth pursuit being in the open-loop phase [20]. It generally reaches peak eye velocity of 30° per second within 220-330ms after the response onset at the target. During maintenance, the eye may exhibit 3-4Hz oscillations for corrective shifts to realign the target image at the fovea. When the target stops, the smooth pursuit movement typically ends within 100ms [69, 203]. The human eye can use numerous available signals to provide cues and predictions for future smooth pursuit movements; for example, eyes continue to pursue the target after it disappears or during occlusion [137]. Intuitively, the velocity of smooth pursuit lies between that of saccades and fixation.

(5) *Blinking* protects the eyes by spreading tears to the corneal surface and blinking periodically keeps the cornea moist. Fixation and blink frequency can be affected by external factors such as humidity and illumination condition as well as internal factors such as cognitive load and fatigue [164].

In human visual systems, these eye movements are driven by the attention mechanisms: bottom-up and top-down, which have formed the basis for different gaze interactive applications. The former mechanism refers to the fact that the fixation point of the eye may be changed by external stimuli that direct attention to discriminative areas of the scene. The latter mechanism is driven by internal stimuli of cognition, which often reflect users' intentions. It makes information available in working memory, and people will consciously pay attention to the scene area that is important to the current behavioural target or task [6, 16]. Sattar et al. [212] learn the compatibility between user fixation scan path and potential targets to predict the correct target image. Wang et al. [243] design a model that learns visual saliency and information visualisation of scan paths based on a sequence of eye movements. These explorations based on gaze and eye movements have great potential for predictive reasoning about users' intentions and interests. The visual attention mechanisms have formed the foundation of these gaze interactions.

6.2 Gaze Data Analytics

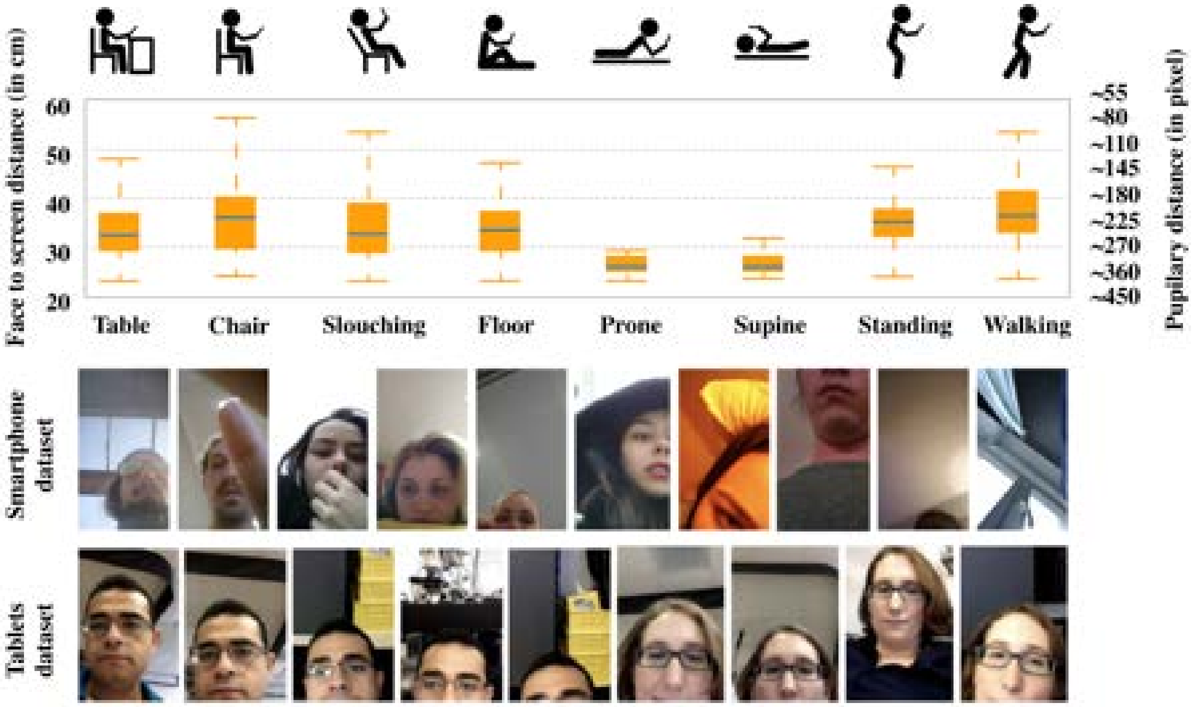


Fig. 7. Various holding postures of handheld mobile devices and their impact on the distance between the face/pupils and the screen, and challenging examples for gaze estimation, images adapted from [100, 101, 118, 119]

Commercial eye-tracking instruments, which generally come with data processing software such as Tobii Studio, process and transform raw eye-movement data into basic gaze events such as fixation, saccade, and saccade trajectory, and form scan path representation. These events are often defined by velocity, acceleration, and amplitude of gaze points and can be inferred via the following algorithms: velocity-threshold fixation identification (I-VT) [75, 113, 210] and identification by dispersion (I-DT) [129], or spatio-temporal dispersion, the identification by dispersion and duration thresholds (I-DDT) [139, 163, 210].

I-VT aims to separate fixation and saccade from raw gaze points by calculating point-to-point velocity and acceleration, while I-DT and I-DDT achieve so by spatial dispersion; for example, fixation is identified when the speed of gaze points is low, and these points form in a dense cluster. The task of distinguishing between saccade, stationary and smooth pursuit is called a ternary classification task. Threshold-based algorithms for this task require a combination of several basic algorithms, such as Velocity Velocity Threshold Identification (I-VVT) [132], Velocity Dispersion Threshold Identification (I-VDT) [132] and Velocity Movement Pattern Identification (I-VMP) [131]. These algorithms have worked well to distinguish saccade and fixation for commercial eye trackers in the lab setting, but due to their rigid threshold setting, they can result in a large error in detecting smooth pursuit [280] when there is a variety in pupil sizes and viewing directions. The above algorithms supported in the open-source libraries including PyGaze [36], EyetrackingR [38], PyTrack [73], Pupil [115], GazeParser [223], and

Table 4. Eye movement data analysis

Pipeline	Description
Data Clean	Noise Reduction, De-Nulling, Eliminate Outlier and Blink, Smoothing
Segmentation for Basic Gaze Events	Fixation, Saccade, Smooth Persuit; Fixation Statistics, Saccade Velocity, Event Duration
Data Processing	Removal of Implausible Movements, Merging Intra-Threshold Movements Data Visualisation
Segmentation for Further Purposes	Area of Interest (AOI), Scan Path Representationl; Gaze Pattern Modelling: Dwell, Pursuit, Gaze Gesture, etc
Possible Inferences	Modelling: Top-Down or Bottom-Up, Stochastic Processes; Inferences: Personal Attributes, Cognitive Process, Attention & Intention
Possible Applications	UI Control & Adaption, Medicine & Healthcare, User Security & Privacy, etc.

GazeR [71]. Currently, a wide range of machine learning-based approaches is proposed to tackle this problem, including Bayesian [211], Random Forest [260], Hidden Markov Models [280], DNN [259], CNN and LSTM [226].

However, for appearance-based gaze estimation on handheld devices, the applications only get the raw gaze points or directions, and there is no off-the-shelf software to process them into high-level patterns. The above algorithms might not be immediately applicable, either. The distance and angle between the head and handheld devices can constantly change, which can compromise the performance of these algorithms. Also, for many desktop environments, the user's gaze point is always within the screen, while for handheld devices, the cameras may only capture partial or occluded faces, or no faces at all (see Figure7 adapted from [100, 101, 118, 119]). Therefore, an interesting research direction is to design and develop robust gaze pattern detection algorithms, resistant to fluctuated and erroneous gaze estimations. The analysis and processing of eye movement data need to combine segmentation and organisation at different scales depending on the purpose of the task; for example, from the original gaze path to the gaze event segmentation, and then to higher dimensional segmentation and organisation incorporating physiological and cognitive factors.

6.3 Gaze Data Processing Pipeline

Table 4 presents a general pipeline of processing gaze points from the area of eye tracking community [47, 87]. It starts with data cleansing to improve the data quality, including noise reduction, de-nulling and dealing with outliers. Here, *null* refers to missing data; for example, an eye-tracking device does not report position coordinates or a participant is not looking at the screen. This is often a necessary step for offline eye-tracking data analysis in desktop environments [89, 244]. Gaze event segmentation refers to detecting gaze events such as fixation or saccade from continuous gaze points using the algorithms in Section 6.2.

The intermediate data processing step is to visualise eye-movement data and remove the data that are beyond a reasonable range and merge data; for example, combining immediately adjacent fixation points. Further segmentation could be used to identify the area of interest (AOI) [90], inferring scan trajectories and heatmaps by saccade path and dwell time, as presented in Figure 6. Such information can serve as advanced features for applications.

6.4 Summary

In the process of investigating gaze analytics and the development of gaze-based applications for handheld devices, we have uncovered several key findings that can guide future research and application design in this area.

Understanding the nuances of eye physiology and human visual attention mechanisms is crucial for the effective design of gaze interactive applications. By comprehending how different types of eye movements, such as saccades, fixations, and smooth pursuits, are related to human consciousness levels and attention mechanisms (bottom-up and top-down), researchers and designers can develop more ergonomic and intuitive gaze-based interaction. This understanding can also facilitate predictive reasoning about users' intentions and interests, which has a significant potential for creating more personalised and effective gaze interaction.

The development of robust gaze pattern detection algorithms is a critical research direction. Current algorithms are faced with the challenges such as changing distance and angle between the head and the device, partial or occluded face captures, and fluctuating gaze estimations. These factors can cause instabilities in eye movement data, leading to issues such as reduced frequency, incoherent absence, and other inconsistencies in gaze pattern detection. Therefore, creating algorithms that can withstand these variations and provide accurate gaze event segmentation and organisation is essential for the wider adoption of gaze-based applications on handheld devices.

7 GAZE INTERACTION

This section will review the existing gaze-based interactive applications. We first introduce the categorisation of gaze interactions and then describe applications across different platforms, including handheld mobile devices, desktops, Virtual Reality (VR) and Augmented Reality (AR). These applications help to uncover new opportunities for interactive applications on handheld mobile devices. Figure 8 illustrates the overall design flow of a gaze-engaged interaction from eye movement to final testing and optimisation.

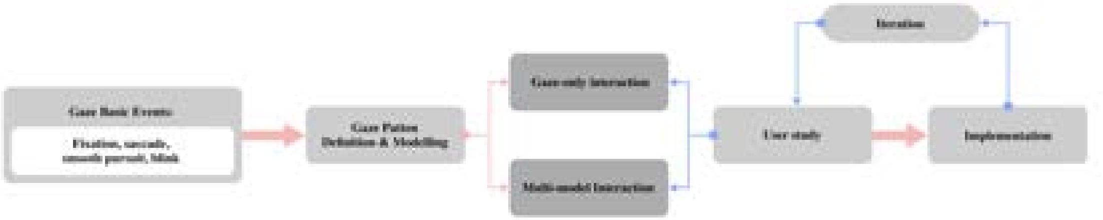


Fig. 8. The workflow of gaze interaction

7.1 Gaze Interaction

Gaze-only interaction can be broadly classified into two main types: implicit and explicit. Implicit interactions involve the interface adapting to the user's passive gaze behaviour, while explicit interactions require the user to intentionally move their eyes to provide direct input. Implicit interaction is usually tailored to specific tasks, devices, and environmental characteristics. For instance, a reading application might predict users' reading speed and automatically turn pages by implicitly analysing users' gaze trajectories [143, 147], or a display system could alter content rendering based on users' intentions [247]. Such predictions are often achieved by applying machine learning techniques to estimate gaze behaviours and infer human intentions.

Explicit interaction, on the other hand, relies on users' voluntary and intentional gaze movements for manipulation. This type of interaction typically employs dwell time, pursuit, gaze gestures, or a combination of these techniques, which will be detailed below.

7.1.1 Dwell-time. Dwell time involves a brief fixating on a target for a period of time [48, 160] in order to differentiate between casual viewing and gaze input. This addresses the Midas touch problem [104] where users unintentionally gaze over the potential target and make selections. This technique has been widely used in gaze-only interactions and is useful for interaction techniques that do not require precise gaze estimates, e.g., gaze typing [160, 172]. The recommended threshold for the dwell time is between 200 and 1500 ms for remote eye-tracking system [160, 172], and the exact threshold often requires trials-and-errors on the threshold for a specific task.

7.1.2 Pursuit. Pursuit refers to smooth pursuit eye movements where the eyes follow a moving object [137, 203, 238]. Pursuit measures the match between users' eye movement and the object's movement via Pearson correlation coefficient [41, 54, 235, 238], or machine learning techniques like CNN [226] and Bayesian [211]. It is often used as a calibration technique [24, 276] as mentioned in Section 4.3.4.

7.1.3 Gaze Gesture. Gaze gesture is a sequence of predefined eye movements (or called strokes) [43, 86, 198, 206]. It is a promising alternative to other gaze interaction techniques, especially when the screen is too small to support other techniques [9]. One advantage of gaze gesture is that it can support a large number of commands using a small number of gesture combinations [43]. However, the use of gaze gestures can introduce complexity, as users may have difficulty recalling complex gestures and initiating them physically [162]. There are machine learning-based methods such as Hierarchical Temporal Memory (HTM) [205, 207] and Graph Neural Networks (GNN) [216] to detect and separate gaze gestures from noisy eye movement signals.

7.2 Types of Gaze Interactions

Gaze-based applications are categorised into three groups based on the interaction and how gaze is acquired: explicit gaze interaction, implicit gaze interaction and multi-model gaze interaction [118, 161]. The two main types of explicit gaze interaction applications are gaze typing and gaze interface control. The former allows users to input text via dwell time or gaze gesture [160, 172, 198, 262], while the latter utilises people's voluntary eye movements and conscious gaze direction to control or communicate with a computer; for example, a user may perform simple horizontal or vertical eye movements to indicate disagreement or agreement [161].

In implicit gaze interaction, there are three main application scenarios: attentive user interfaces, passive eye monitoring, and gaze-based user modelling. The attentive user interface applies users' natural eye movements rather than expecting particular gaze behaviours for explicit commands; for example, changing the movie plot based on the viewer's visual interest [237]. Gaze-based user modelling monitors and analyses the dynamics of gaze behaviour over time to understand users' behaviour, intention, and cognitive processes [269]. For example, eye movement patterns have been used to recognise human activities such as reading or common office activities [18]. The implicit gaze interaction signals can be used to recognise and speculate humans' latent behaviours, including measurement of users' preferences [145], attention [56, 57, 169, 177], interests [150], individual stress [99], emotional states [173], and mental disorders such as schizophrenia and autism spectrum disorder [7, 50, 217]. For example, Deng et al. [37] have analysed eye movement of drivers to predict their fixation and understand their attention allocation on scenes, and Pan et al. [182] have explored the use of eye-tracking technologies to predict drivers' lane-changing intention. Passive eye monitoring is often for diagnostic applications [46, 220] where people's visual behaviours are recorded for offline processing [46].

Multi-modal gaze interaction refers to the use of gaze alongside other input modalities, such as voice, touch, or hand gestures. Multi-modal gaze interaction can overcome the limitations of other modalities when they are not accurate or when users have difficulty interacting with them. Gaze can also enhance other modalities; for example, gaze has been combined with touch-based PINs for a more secure authentication solution [121], and gaze is used to improve voice interaction by fixating on an object [167]. Both explicit and implicit gaze interactions can be employed in multi-modal interactions, offering a flexible and intuitive user experience.

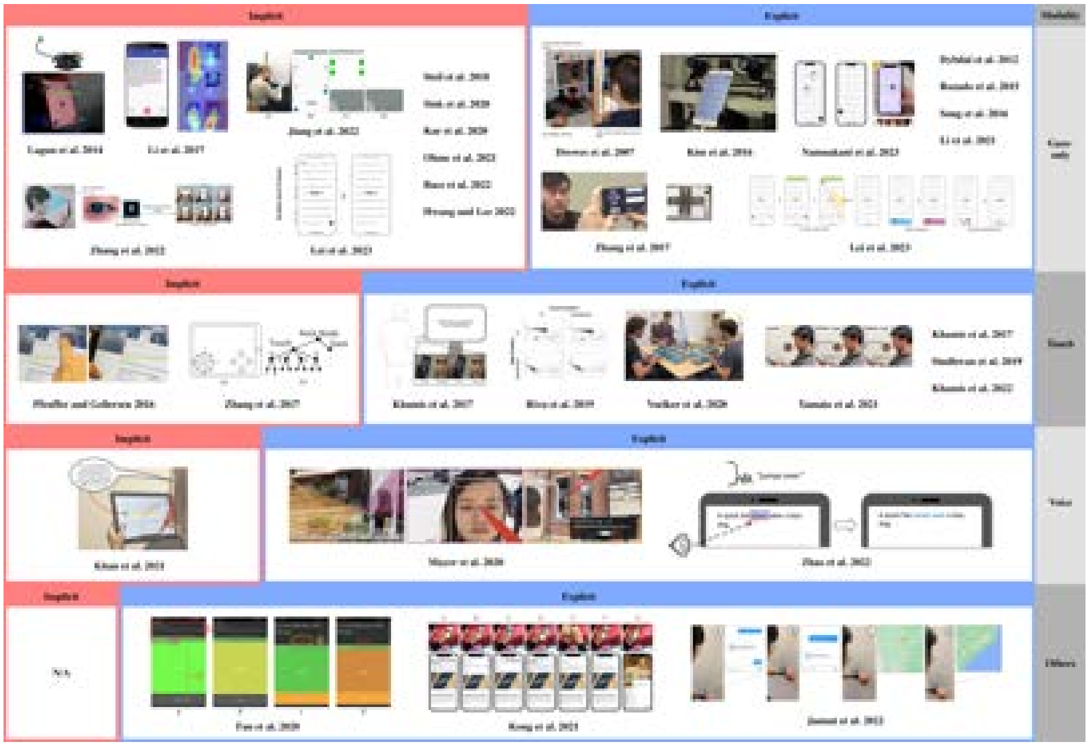


Fig. 9. Examples of gaze applications on handheld mobile devices

In the context of handheld mobile devices, the applications mainly fall into the first two categories, and the trend is moving from gaze-only interaction to multi-model interaction. As listed in Table 5, earlier applications have used explicit gaze gestures to perform particular commands. With the improvement of gaze detection, gaze has served as implicit interaction input; for example, deriving users' interest or attention by monitoring their gaze in the background [112, 177, 180]. Wrist-worn gaze control [80] provides a gaze-based smart home control setting, which uses off-screen gaze gestures on a wrist-worn unit for IoT control. Gaze+Hold [199] provides a gaze-based interface, where a user can use gaze gesture combined with blink and fixation to complete most of the mouse functionality. For example, the selection of an object is performed by first fixating on an object and then closing and opening one eye. The downside of this type of application is that the gaze input might be slow and less accurate than the single-source input, and users may potentially experience fatigue [199].

Also, the applications can be classified as gaze-only and multi-modal interaction [118]. The former uses gaze solely as the interaction input, and this is widely adopted by gaze applications in handheld mobile devices. The latter uses gaze to complement other interaction modalities, such as touching or tilting. Pfeuffer and Gellersen [193] have explored the gaze and touch interaction to extend the area that the thumb cannot reach. For example, a user can look at an element on the screen and tap anywhere, and the system will activate the element. GTmoPass [120] is an authentication method that uses handheld mobile devices to enter multi-modal passwords; that is, combining gaze gestures (e.g., left to right) with touch input.

Table 5. Gaze applications on handheld mobile devices

Project	Year	Sensor	Feature	Modality	Leverage	Application
[40]	2007	Eye Response ERICA	gaze gesture	-	E	interface control
[48]	2012	external IR cam	dwell & gesture	-	E	target selection
[145]	2014	Tobii X60	dwell	-	I	attention analysis
[206]	2015	modified prototype	gaze gesture	-	E	interface control
[193]	2016	Tobii EyeX	dwell	touch	I	interface control
[224]	2016	phone camera	gaze basic events	-	E	user authentication
[125]	2016	Facelab 5	gaze basic events	-	I	Website usability test
[150]	2017	phone camera	dwell	-	I	intention inference
[120]	2017	phone camera	gaze gesture	touch	E	user authentication
[261]	2017	Tobii eyeX	gaze basic events	touch	I	gaze adaptive UI
[267]	2017	phone camera	gaze gesture	-	E	gaze input
[121]	2017	external RGB cam	gaze gesture	touch	E	user authentication
[227]	2018	phone camera	gaze basic events	-	I	attention inference
[219]	2019	Tobii 4C	dwell	touch	E	text editing aids
[202]	2019	Tobii 4C	dwell	touch	E	text interface control
[167]	2020	phone camera	dwell	voice	E	map navigation
[220]	2020	phone camera	eye image	-	I	ocular exam
[239]	2020	phone camera	dwell	touch	E	cross-device control
[58]	2020	phone camera	dwell	eyelid	E	interface control
[112]	2020	phone camera	gaze basic events	-	I	attention analysis
[254]	2021	phone camera	gaze gesture	touch	E	gaze-assist input
[133]	2021	phone camera	dwell	hand motion	E	interface control
[180]	2021	Tobii X2	gaze basic events	-	I	attention analysis
[153]	2021	phone camera	dwell	-	E	target selection
[123]	2021	Tobii 4C	dwell	voice	I	implicit note-taking
[122]	2022	external RGB cam	gaze gesture	touch	E	user authentication
[274]	2022	phone camera	dwell	voice	E	text correction
[10]	2022	external RGB cam	gaze & face	-	I	user privacy
[266]	2022	phone camera	eye image	-	I	holding posture detection
[275]	2022	external RGB cam	gaze basic events	*	E	gaze command definition
[103]	2022	Tobii X2	gaze basic events	-	I	attention analysis
[107]	2022	SMI Glasses	gaze basic events	-	I	learning process of typing
[105]	2022	watch camera	face position	hand motion	E	spatial user interfaces
[175]	2023	phone camera	dwell, pursuit, gesture	-	E	gaze UI usability test
[147]	2023	phone camera	dwell, pursuit, gesture	-	I&E	gaze UI usability test

Note: *Feature* means gaze feature for interaction; *Modality* means the other modality with gaze; *Leverage* means the way of leveraging gaze; *E* means Explicit; *I* means Implicit; gaze basic events mean: fixation, saccade, smooth pursuit etc; *: including eyelids, mouth, and head.

7.3 Wider Application Domains

A large number of applications have employed gaze as an interactive modality to complement other modalities in different application domains, including accessibility and productivity, collaboration across devices, device and robot control, and security and privacy.

7.3.1 Accessibility and Productivity. Gaze can be an intuitive interaction modality to be complemented with other modalities to enhance accessibility and productivity. This area of applications can be further grouped into gaming [168], web browsing [142], typing [219], and computer interaction [51, 148, 202, 213]. The main idea is to detect users' attention and intention from their gaze, and use their gaze dwell time and gesture to trigger different actions. ReType [219] is a gaze-assisted positioning technique, which makes use of users' gaze to position the text where they are interested. This interaction method reduces the use of the mouse and allows users to perform editing operations while keeping their hands on the keyboard. GazeHelp [148] assists graphic design activities using real-time gaze information. It works as a plugin to Adobe PhotoShop to allow users to select tools with gaze, create windows at the gaze point, and block the current artboard when the gaze is away from the display. It integrates gaze points and a mechanical switch for object selection, positioning, and manipulation. Gaze has been applied to empower disadvantaged people; for example, creating visual design [34] and coding [190].

7.3.2 Collaboration. Prior works also point out that gaze interaction can improve multi-user collaboration efficiency and user experience [85, 144, 192, 214]. Gaze-sharing user interface [214] detects the user's attention by sensing the user's eye movements and shares them with other collaborators to enhance team collaboration. GazeChat [85] is a remote communication system that renders gaze-aware 3D profile photos. It uses an off-the-shelf webcam to track users' gaze in video calls such as Zoom or Teams and then renders gaze to animate the participants' profile images.

In addition to team collaboration, gaze can also support multi-device interaction. It is increasingly common for a user to have multiple devices or multiple screens. In a collaborative environment with one or more tablets, GazeConduits [239] uses the phone's front camera to sense and detect the user's gaze to identify which tablet the user is looking at and making content selections and actions based on the user's gaze gestures. GazeMirror [239] allows users to mirror the content between devices by coordinating gaze and four-finger multi-touch gesture. Another gaze+touch project [192] allows multiple users to zoom in and out on a single shared map without interference or occlusion issues.

7.3.3 Control and Interaction. Gaze is playing an important role in controlling devices [110] and robots [152]. Krishna et al. [140] propose a control interface that uses a tablet to process and estimate gaze. The project uses dwell time as a trigger for interaction instruction and manipulates the robot arm to perform a pick-and-drop operation. Jungwirth et al. [110] have designed gaze-triggered actions that use object contours as visual guidance; for example, a user can trace the contour of a lamp to turn on the light.

7.3.4 Security and Privacy. Gaze supports security and privacy in applications. GazeConduits [239] provides user awareness in a poker game; for example, showing playing cards when the gaze looks and hiding them when the gaze leaves. It also maintains the user's space around the table by detecting who enters and leaves the space at what time and accordingly adjusting the content display. GazeRoomLock [72] combines gaze and head pose for user authentication in VR applications. EyeVeri [224] applies signal processing and pattern matching techniques to explore conscious and unconscious gaze patterns for access authentication. In addition, gaze-based interaction can support authentication and privacy protection on personal devices [116] such as password entry [120] and protection against shoulder surfer [197].

7.3.5 Healthcare and Other Areas. As an important bio-signal, eye movement patterns have been widely used in the diagnosis of mental health such as depression[3], Parkinson's[81], autism [236], and dyslexia [201]. For example, autism can be diagnosed by analysing joint attention from interactions between eye movement and objects in a room [236]. Gaze is also utilised for assessing the usability of tools or systems [21, 191]. Bace et al. [21] have employed an unsupervised approach to detect gaze contact and attempted to compute and quantify attentional metrics. This allows for analysing how users allocate attention during interactions.

7.4 Summary

We have reviewed a broader scope of applications that make use of gaze as an interaction modality. Most of these applications use commercial eye-tracking devices, including Tobii [51, 148] and Eye tribe [34]. We have seen an increasing number of applications for collaboration, control and interaction, security and authentication developed on mobile devices. In terms of accessibility and productivity, there is a tendency to explore complex actions from gaze, which can be challenging for handheld devices. For example, typing on the phone's keyboard often requires the gaze to switch back and forth between the target area and the keyboard. This makes it challenging to locate the gaze position accurately and thus can trigger the Midas touch problem. The key to tackling the Midas touch problem is to analyse the eye movement metrics through the cognitive process and separate users' true intention from the unintentional activities [51, 61, 183, 199]. The other applications that combine fixation and eye-opening/closing actions to select and manipulate objects can be further explored. A potential limitation is that the mobile devices have smaller icons that are presented closely together, which can make the estimation of gaze and inference of gaze trajectories more challenging.

8 FUTURE RESEARCH CHALLENGES AND OPPORTUNITIES

This section focuses on addressing key questions related to the challenges of handheld mobile devices in advancing unconstrained gaze estimation, robust gaze data processing methods and facilitating interaction involving gaze. Our discussion will be centred around three main questions:

- **RQ1**: How can we achieve robust gaze estimation in unconstrained environments?
- **RQ2**: How can we develop gaze analysis and processing methods that can tolerate the inherent instabilities of dynamic gaze estimation?
- **RQ3:** How can we utilise estimated gaze for a broader range of applications?

8.1 Roadmap for Unconstrained Gaze Estimation

The mobile setting is bringing new challenges to gaze estimation. First of all, the distance and angle between the screen and the eyes might not be ideal or stable. The consequence is that the camera might only capture a partial face or an occluded face; for example, the face could be covered by sunglasses or a scarf (see Figure 7). Secondly, users tend to change their holding postures, motion states, and whereabouts, which may compromise the gaze estimation and/or calibration model. Thirdly, the variety of environmental conditions, such as lighting may compromise the quality of images for face and eye detection. In the following, we will present potential directions to tackle these challenges towards achieving unconstrained gaze estimation.

8.1.1 Augmenting with Sensors and Multi-Cameras. One of the major differences between mobile devices and desktop/AR/VR devices is that modern mobile devices are augmented with a rich set of sensors. These sensors provide more information that can reveal the context of the user. For example, we can infer the holding posture from the gyroscope, physical activities from the accelerometer, and the lighting condition from light sensors. Also, we have witnessed an improvement in the camera and sensor technologies on mobile devices. For example, some mobile devices have between three and six cameras with higher resolutions. Combining data from these built-in sensors with the camera may improve the accuracy of gaze estimation [25].

8.1.2 Continuous Calibration. Calibration is often necessary to tune the gaze estimation model and to allow for more accurate estimation adapting to the current user, screen, and environment. *Continuous calibration* is an interesting direction to explore; for example, quickly adapting the model continuously and obtaining the calibration points without distracting users from their tasks at hand.

A potential direction is to allow for implicit calibration, which leverages users' interaction and other sensor data in mobile devices, including accelerometer, gyroscope, magnetometer, and compass. Gaze behaviour, such as smooth pursuit, has been explored for calibration. For example, correlating eye or hand movements with the trajectory of moving objects has been utilised for in-use calibration [23, 176]. The trajectory of a straight saccade has been used to calibrate the distortion of eye tracker [97].

Current handheld mobile devices integrate a number of inertial sensors, and their data can be employed for dynamic adjustment. Pino et al. [194] propose to detect changes in the position and holding postures of the phone from gyroscopes and accelerometers. The detected change will inform whether to use current eye images for gaze estimation. Gaze-based input requires a high level of stability in the input process, and device movements and other possible disturbances can decrease the quality of input. A promising direction is to use inertial sensor data to select the input for gaze estimation algorithms or select and tune the gaze estimation algorithms for the current input.

Research can investigate real-time feedback and calibration mechanisms to enhance the resilience of gaze estimation in handheld mobile devices. Real-time feedback on users' gaze behaviour can be used to calibrate the model on-the-fly, leading to high resilience even in the face of changing conditions.

8.1.3 Diversity in Datasets. As presented in Section 5, current datasets predominantly contain static poses and are limited in terms of capturing gaze estimation in dynamic conditions. There is a need to collect a large amount of data with versatile conditions in more naturalistic settings. However, such data can be expensive to collect; therefore, the current research direction is to explore unsupervised [79], self-supervised [30, 251] and weakly supervised techniques [134].

One increasing concern is *bias* in face recognition models; for example, the state-of-the-art models are trained on the datasets that over-represent socio-demographic groups with certain skin tones and facial structures, which makes them less accurate for marginalised groups [245]. For example, Buolamwini and Gebru have evaluated three commercial gender classification systems and found that the error rate for classifying lighter-skinned males is 0.8% while the error rate for classifying darker-skinned females can reach up to 34.7% [19]. The bias in the dataset has led to low-quality facial feature extraction on under-represented groups, which can result in low accuracy in gaze estimation. Therefore, future data collection should cover a wider ethnic group of subjects.

8.1.4 Model Deployment. Deploying high-performance gaze estimation models on handheld mobile devices presents challenges in model compression and model deployment.

Model Compression. The existing gaze estimation models, especially the ones based on deep learning, can be computationally expensive and take up much memory; for example, a deep learning model can have billions of parameters. Even though today's smartphones have much more computational resources, continuously estimating gaze in real-time can still be challenging, and it can compromise their battery life and reduce user experience. Model compression is a popular direction to reduce the size of deep learning models while maintaining model performance. There are several approaches, such as parameter quantisation, parameter pruning, low-rank factorisation and knowledge distillation. Guo et al. [78] apply knowledge distillation and pruning to reduce the size of the CNN model while maintaining its performance.

Model Deployment. Many of the existing gaze estimation models are deployed on the cloud or edge so that mobile devices only perform pre-processing steps and then pass the extracted features to the deep learning models. However, this adds to the communication cost, results in high latency, and increases privacy risk. Future work will look into deploying the models on users' own devices. However, this raises a practical issue – model framework compatibility.

Most of the deep learning models are implemented in frameworks such as PyTorch or TensorFlow, and these frameworks do not have the same support on different handheld devices. The ONNX (Open Neural Network Exchange) [11] can be used as a medium to transform models across different deep learning frameworks and thus can help deploy models to various platforms of handheld mobile devices. It only supports operators that are commonly supported by all the deep learning frameworks, but not for specialised operators for specific frameworks. This limits the deployment of advanced, customised, gaze estimation models on handheld devices. An engineering perspective of research is how to support gaze estimation on a wide range of different devices.

8.1.5 Open Standards.

It is important to highlight the current lack of international or national industrial standards specifically for eye-tracking technologies and systems. Existing standards such as *ISO 9241-971:2020 Ergonomics of human system interaction* and *ISO 13407:1999: Human-centred design processes for interactive systems* focus on usability assessment and human-centred design principles. Similarly, *WCAG 2.1: Web Content Accessibility Guidelines* is for conducting usability studies with people with disabilities, which are relevant to the use of oculomotor systems.

There is a need for the industry to develop appropriate standards that regulate the overall performance and usability of eye-tracking systems. These standards should address various aspects, including device accuracy, calibration procedures, data processing, and user privacy. Establishing such standards could ensure consistent quality and interoperability across different eye-tracking systems, facilitating their adoption and fostering innovation in the field.

8.2 Roadmap for Gaze Analytics

As described in Section 6, there are no established algorithms and toolkits for processing raw gaze points or vectors into high-level gaze patterns or events. Appearance-based gaze estimation on handheld mobile devices can be much less accurate than commercial eye-tracking devices in terms of large error, low frequency, and high instability. The existing methods of removing outliers, smoothing data, or merging intra-threshold movements might lead to significant information loss, compromising the quality of gaze analytics. Therefore, a future research direction is to develop gaze event detection and processing algorithms to tolerate imperfect gaze estimation. This may require a further understanding of user behaviour patterns and cognitive processes in various applications. Overcoming these challenges will not only facilitate more robust and accurate gaze interaction but also pave the way for a wide range of novel applications in human-computer interaction.

8.3 Roadmap for a Wider adoption of Gaze Applications

8.3.1 Privacy Implication.

Gaze interaction can have high privacy implications. First of all, gaze can reveal users' intentions, interests, emotions, and personality traits [16, 93, 189, 243]. Secondly, gaze estimation from handheld mobile devices is based on users' facial images. Thirdly, the camera may capture the background, such as bystanders' activities. Research effort should be devoted to protecting the foreground and background users' privacy; for example, how to securely store face/eye images, prevent leaking gaze information to third-party applications on the same devices, and explore the possibility of protecting the privacy of bystanders [116].

8.3.2 Gaze Interaction.

One obstacle to the wider adoption of gaze interaction on handheld mobile devices is the imprecision of estimation [15] and sensitivity to the environments and interaction positions. Imprecision can increase users' fatigue and lower user experience. Furthermore, the screen size of handheld mobile devices might limit the choices of gaze pattern combinations; for example, saccade detection on a small screen can be challenging and smooth pursuit or dwell and gesture combination might be a better option.

Promising applications for gaze interaction on handheld mobile devices include healthcare, usability testing, cognitive process understanding, interface and device control, and security and privacy protection. For example, gaze-engaged interface or device control in accessibility contexts can offer an alternative input method for users with motor impairments, and gaze patterns can be used as bio-metric identifiers to enhance security and privacy protection.

Apart from using the front camera to capture the gaze behaviour of the target user, leveraging other cameras, such as rear or external cameras, to capture and project multiple gaze directions and identify objects of interest or perform interaction [117] has rarely been explored. This approach holds promise for opening up a new research area, as it has the potential to link gaze with physical objects or locations in the real world through socialised gaze cues, enabling a more immersive user experience. Gaze-engaged multi-modal interaction is another promising area, as exploring such interactions can complement gaze and other modalities, leading to novel human-computer interaction.

9 CONCLUSION

Mobile human-computer interaction is one of the most popular areas for innovation. This paper has presented a review of gaze estimation and interaction on handheld mobile devices. There are many promising developments in this area: more powerful mobile devices with high-resolution cameras, accurate gaze estimation algorithms based on deep learning, and an increasing number of novel applications that leverage the use of gaze to enable hands-free interaction or complement other interaction modalities. This paper summarises these latest developments and points to the future research challenges and opportunities, especially in gaze estimation accuracy in terms of robustness and continuous calibration, the computational cost in terms of model size and power consumption, and gaze interaction in terms of richer eye movement exploration. It is desirable that the gaze estimation algorithm is adaptable to different environmental conditions and has a low computational overhead. Continuous and implicit calibration will be key factor in supporting unconstrained gaze estimation.

Gaze estimation algorithms are becoming increasingly mature and, although still far from unconstrained estimation, can be applied on handheld mobile devices by installing apps that deploy models. We review the existing types of gaze interactions, summarise various types of eye movements, cognitive theories and combinations of gestures used in gaze interactions, and give a generalised workflow of gaze interactions in the hope that more people will see the role of gaze in interactions and use it to develop more and novel interactions and applications.

REFERENCES

[1] Ahmed A. Abdelrahman, Thorsten Hempel, Aly Khalifa, and Ayoub Al-Hamadi. 2022. L2CS-Net: Fine-Grained Gaze Estimation in Unconstrained Environments. arXiv:2203.03339 [cs.CV]

[2] Robert G Alexander, Stephen L Macknik, and Susana Martinez-Conde. 2018. Microsaccade characteristics in neurological and ophthalmic disease. *Frontiers in neurology* 9 (2018), 144.

[3] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Gordon Parker, and Michael Breakspear. 2013. Eye movement analysis for depression detection. In *2013 IEEE International Conference on Image Processing*. IEEE, 4220–4224.

[4] Dima Amso, Sara Haas, and Julie Markant. 2014. An eye tracking investigation of developmental change in bottom-up attention orienting to faces in cluttered natural scenes. *PloS one* 9, 1 (2014), e85701.

[5] Rishi Athavale, Lakshmi Sritan Motati, and Rohan Kalahasty. 2022. One Eye is All You Need: Lightweight Ensembles for Gaze Estimation with Single Encoders. arXiv:2211.11936 [cs.CV]

[6] Edward Awh, Edward K Vogel, and S-H Oh. 2006. Interactions between attention and working memory. *Neuroscience* 139, 1 (2006), 201–208.

[7] Pradeep Raj Krishnappa Babu and Uttama Lahiri. 2019. Understanding the role of Proximity and Eye gaze in human–computer interaction for individuals with autism. *Journal of Ambient Intelligence and Humanized Computing* 5 (2019), 1–15.

[8] Mihai Bace, Vincent Becker, Chenyang Wang, and Andreas Bulling. 2020. Combining Gaze Estimation and Optical Flow for Pursuits Interaction. In *ETRA '20*. ACM, Article 2, 10 pages.

[9] Mihai Bâce, Teemu Leppänen, David Gil de Gomez, and Argenis Ramirez Gomez. 2016. UbiGaze: Ubiquitous Augmented Reality Messaging Using Gaze Gestures. In *SIGGRAPH ASIA 2016 Mobile Graphics and Interactive Applications*. ACM, Article 11, 5 pages.

[10] Mihai Bâce, Alia Saad, Mohamed Khamis, Stefan Schneegass, and Andreas Bulling. 2022. PrivacyScout: Assessing Vulnerability to Shoulder Surfing on Mobile Devices. *Proceedings on Privacy Enhancing Technologies* 1 (2022), 21.

[11] Bai, Junjie and Lu, Fang and Zhang, Ke and others. 2019. *Onnx: Open neural network exchange.* Github. https://github.com/onnx/onnx

[12] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 59–66.

[13] Jun Bao, Buyu Liu, and Jun Yu. 2022. An individual-difference-aware model for cross-person gaze estimation. *IEEE Transactions on Image Processing* 31 (2022), 3322–3333.

[14] Yiwei Bao, Yihua Cheng, Yunfei Liu, and Feng Lu. 2021. Adaptive feature fusion network for gaze tracking in mobile tablets. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 9936–9943.

[15] Michael Barz, Florian Daiber, Daniel Sonntag, and Andreas Bulling. 2018. Error-Aware Gaze-Based Interfaces for Robust Mobile Gaze Interaction. In *ETRA '18*. ACM, Article 24, 10 pages.

[16] Michael Barz, Sven Stauden, and Daniel Sonntag. 2020. Visual Search Target Inference in Natural Interaction Settings with Machine Learning. In *ETRA '20*. ACM, Article 1, 8 pages.

[17] Glenn Beach, Charles J Cohen, Jeff Braun, and Gary Moody. 1998. Eye tracker system for use with head mounted displays. In *SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 98CH36218)*, Vol. 5. IEEE, 4348–4352.

[18] Andreas Bulling, Jamie A. Ward, and Hans Gellersen. 2012. Multimodal Recognition of Reading Activity in Transit Using Body-Worn Sensors. *ACM Trans. Appl. Percept.* 9, 1, Article 2 (mar 2012), 21 pages.

[19] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[20] Antimo Buonocore, Julianne Skinner, and Ziad M Hafed. 2019. Eye position error influence over "open-loop" smooth pursuit initiation. *Journal of Neuroscience* 39, 14 (2019), 2709–2721.

[21] Mihai Bâce, Sander Staal, and Andreas Bulling. 2019. Accurate and Robust Eye Contact Detection During Everyday Mobile Device Interactions. arXiv:1907.11115 [cs.HC]

[22] Lijun Cai, Lei Huang, and Changping Liu. 2015. Person-specific Face Spoofing Detection for Replay Attack Based on Gaze Estimation. In *Biometric Recognition*. Springer, 201–211.

[23] Marcus Carter, Eduardo Velloso, John Downs, Abigail Sellen, Kenton O'Hara, and Frank Vetere. 2016. PathSync: Multi-User Gestural Interaction with Touchless Rhythmic Path Mimicry. In *CHI '16*. ACM, 3415–3427.

[24] Feridun M. Celebi, Elizabeth S. Kim, Quan Wang, Carla A. Wall, and Frederick Shic. 2014. A Smooth Pursuit Calibration Technique. In *ETRA '14*. ACM, 377–378.

[25] Yuhu Chang, Changyang He, Yingying Zhao, Tun Lu, and Ning Gu. 2021. A High-Frame-Rate Eye-Tracking Framework for Mobile Devices. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1445–1449.

[26] Zhaokang Chen and Bertram E. Shi. 2019. Appearance-Based Gaze Estimation Using Dilated-Convolutions. In *Computer Vision – ACCV 2018*, C.V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler (Eds.). Springer, 309–324.

[27] Shiwei CHENG, Qiufeng PING, Jialing WANG, and Yijian CHEN. 2022. EasyGaze: Hybrid eye tracking approach for handheld mobile devices. *Virtual Reality & Intelligent Hardware* 4, 2 (2022), 173–188.

[28] Yihua Cheng, Yiwei Bao, and Feng Lu. 2022. PureGaze: Purifying Gaze Feature for Generalizable Gaze Estimation. *AAAI* 36, 1 (Jun. 2022), 436–443.

[29] Yihua Cheng and Feng Lu. 2021. Gaze Estimation using Transformer. arXiv:2105.14424 [cs.CV]

[30] Yihua Cheng, Feng Lu, and Xucong Zhang. 2018. Appearance-based gaze estimation via evaluation-guided asymmetric regression. In *the European Conference on Computer Vision (ECCV)*. IEEE, 105–121.

[31] Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. 2021. Appearance-based Gaze Estimation With Deep Learning: A Review and Benchmark. arXiv:2104.12668 [cs.CV]

[32] Andrew SA Chetwood, Ka-Wai Kwok, Loi-Wah Sun, George P Mylonas, James Clark, Ara Darzi, and Guang-Zhong Yang. 2012. Collaborative eye tracking: a potential training tool in laparoscopic surgery. *Surgical endoscopy* 26, 7 (2012), 2003–2009.

[33] Kevin Cortacero, Tobias Fischer, and Yiannis Demiris. 2019. RT-BENE: a dataset and baselines for real-time blink estimation in natural environments. In *the IEEE/CVF International Conference on Computer Vision Workshops*. IEEE, 0–0.

[34] Chris Creed, Maite Frutos-Pascual, and Ian Williams. 2020. Multimodal Gaze Interaction for Creative Design. In *CHI '20*. ACM, 1–13.

[35] Wen Cui, Jinshi Cui, and Hongbin Zha. 2017. Specialized gaze estimation for children by convolutional neural network and domain adaptation. In *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 3305–3309.

[36] Edwin S Dalmaijer, Sebastiaan Mathôt, and Stefan Van der Stigchel. 2014. PyGaze: An open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. *Behavior research methods* 46, 4 (2014), 913–921.

[37] Tao Deng, Hongmei Yan, Long Qin, Thuyen Ngo, and B. S. Manjunath. 2020. How Do Drivers Allocate Their Potential Attention? Driving Fixation Prediction via Convolutional Neural Networks. *IEEE Transactions on Intelligent Transportation Systems* 21, 5 (2020), 2146–2154.

[38] Jacob W Dink and Brock Ferguson. 2015. eyetrackingR: An R library for eye-tracking data analysis.

[39] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 [cs.CV]

[40] Heiko Drewes, Alexander De Luca, and Albrecht Schmidt. 2007. Eye-Gaze Interaction for Mobile Phones. In *International Conference on Mobile Technology, Applications, and Systems and the 1st International Symposium on Computer Human Interaction in Mobile Technology*. ACM, 364–371.

[41] Heiko Drewes, Mohamed Khamis, and Florian Alt. 2019. DialPlates: Enabling Pursuits-Based User Interfaces with Large Target Numbers. In *MUM '19*. ACM, Article 10, 10 pages.

[42] Heiko Drewes, Ken Pfeuffer, and Florian Alt. 2019. Time- and Space-Efficient Eye Tracker Calibration. In *ETRA '19*. ACM, Article 7, 8 pages.

[43] Heiko Drewes and Albrecht Schmidt. 2007. Interacting with the Computer Using Gaze Gestures. In *Human-Computer Interaction – INTERACT 2007*, Cécilia Baranauskas, Philippe Palanque, Julio Abascal, and Simone Diniz Junqueira Barbosa (Eds.). Springer Berlin Heidelberg, 475–488.

[44] Lingyu Du and Guohao Lan. 2022. FreeGaze: Resource-efficient Gaze Estimation via Frequency Domain Contrastive Learning. arXiv:2209.06692 [cs.CV]

[45] Neeru Dubey, Shreya Ghosh, and Abhinav Dhall. 2019. Unsupervised learning of eye gaze representation from the web. In *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–7.

[46] Andrew T Duchowski. 2018. Gaze-based interaction: A 30 year retrospective. *Computers & Graphics* 73 (2018), 59–69.

[47] Andrew T Duchowski and Andrew T Duchowski. 2017. *Eye tracking methodology: Theory and practice*. Springer.

[48] Morten Lund Dybdal, Javier San Agustin, and John Paulin Hansen. 2012. Gaze Input for Mobile Devices by Dwell and Gestures. In *ETRA '12*. ACM, 225–228.

[49] Mohamad A Eid, Nikolas Giakoumidis, and Abdulmotaleb El Saddik. 2016. A novel eye-gaze-controlled wheelchair system for navigating unknown environments: case study with a person with ALS. *IEEE Access* 4 (2016), 558–573.

[50] Mahmoud Elbattah, Jean-Luc Guérin, Romuald Carette, Federica Cilia, and Gilles Dequen. 2020. NLP-Based Approach to Detect Autism Spectrum Disorder in Saccadic Eye Movement. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 1581–1587.

[51] Carlos Elmadjian and Carlos H Morimoto. 2021. GazeBar: Exploiting the Midas Touch in Gaze Interaction. In *CHI EA '21*. ACM, Article 248, 7 pages.

[52] Kara J Emery, Marina Zannoli, James Warren, Lei Xiao, and Sachin S Talathi. 2021. OpenNEEDS: A Dataset of Gaze, Head, Hand, and Scene Signals During Exploration in Open-Ended VR Environments. In *ETRA '21*. 1–7.

[53] Ralf Engbert and Reinhold Kliegl. 2003. Microsaccades uncover the orientation of covert attention. *Vision research* 43, 9 (2003), 1035–1045.

[54] Augusto Esteves, Eduardo Velloso, Andreas Bulling, and Hans Gellersen. 2015. Orbits: Gaze Interaction for Smart Watches Using Smooth Pursuit Eye Movements. In *UIST '15*. ACM, 457–466.

[55] EyeTech. 2023. *EyeOn Air – Eye Tracking Communication Aid*. eyetechds. Retrieved 2023-03-10 from https://eyetechds.com/eyeon-air/

[56] Myrthe Faber, Robert Bixler, and Sidney K D'Mello. 2018. An automated behavioral measure of mind wandering during computerized reading. *Behavior Research Methods* 50, 1 (2018), 134–150.

[57] Myrthe Faber, Kristina Krasich, Robert E Bixler, James R Brockmole, and Sidney K D'Mello. 2020. The eye–mind wandering link: Identifying gaze indices of mind wandering across tasks. *Journal of experimental psychology: human perception and performance* 46, 10 (2020), 1201.

[58] Mingming Fan, Zhen Li, and Franklin Mingzhe Li. 2020. Eyelid gestures on mobile devices for people with motor impairments. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. Association for Computing Machinery, 1–8.

[59] Anna Maria Feit, Lukas Vordemann, Seonwook Park, Caterina Bérubé, and Otmar Hilliges. 2020. Detecting Relevance during Decision-Making from Eye Movements for UI Adaptation. In *ACM Symposium on Eye Tracking Research & Applications*. Association for Computing Machinery, 1–11.

[60] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. 2018. Joint 3d face reconstruction and dense alignment with position map regression network. In *the European Conference on Computer Vision (ECCV)*. Springer, 534–551.

[61] Paul Festor, Ali Shafti, Alex Harston, Michey Li, Pavel Orlov, and A. Aldo Faisal. 2022. MIDAS: Deep learning human action intention prediction from natural eye movement patterns. arXiv:2201.09135 [cs.CV]

[62] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. 2018. RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments. In *the European Conference on Computer Vision (ECCV)*. Springer, 334–352.

[63] Wolfgang Fuhl, Shahram Eivazi, Benedikt Hosp, Anna Eivazi, Wolfgang Rosenstiel, and Enkelejda Kasneci. 2018. BORE: Boosted-Oriented Edge Optimization for Robust, Real Time Remote Pupil Center Detection. In *ETRA '16*. ACM, Article 48, 5 pages.

[64] Wolfgang Fuhl, David Geisler, Thiago Santini, Tobias Appel, Wolfgang Rosenstiel, and Enkelejda Kasneci. 2018. CBF: Circular Binary Features for Robust and Real-Time Pupil Center Detection. In *ETRA '18*. ACM, Article 8, 6 pages.

[65] Wolfgang Fuhl, Gjergji Kasneci, and Enkelejda Kasneci. 2021. TEyeD: Over 20 Million Real-World Eye Images with Pupil, Eyelid, and Iris 2D and 3D Segmentations, 2D and 3D Landmarks, 3D Eyeball, Gaze Vector, and Eye Movement Types. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 367–375.

[66] Wolfgang Fuhl, Thomas Kübler, Katrin Sippel, Wolfgang Rosenstiel, and Enkelejda Kasneci. 2015. ExCuSe: Robust Pupil Detection in Real-World Scenarios. In *Computer Analysis of Images and Patterns*, George Azzopardi and Nicolai Petkov (Eds.). Springer, 39–51.

[67] Wolfgang Fuhl, Thiago Santini, Gjergji Kasneci, Wolfgang Rosenstiel, and Enkelejda Kasneci. 2017. PupilNet v2.0: Convolutional Neural Networks for CPU based real time Robust Pupil Detection. arXiv:1711.00112 [cs.CV]

[68] Wolfgang Fuhl, Thiago C. Santini, Thomas Kübler, and Enkelejda Kasneci. 2016. ElSe: Ellipse Selection for Robust Pupil Detection in Real-World Environments. In *ETRA '16*. ACM, 123–130.

[69] Kikuro Fukushima, Junko Fukushima, Tateo Warabi, and Graham R Barnes. 2013. Cognitive processes involved in smooth pursuit eye movements: behavioral evidence, neural substrate and clinical correlation. *Frontiers in systems neuroscience* 7 (2013), 4.

[70] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. 2014. EYEDIAP: A Database for the Development and Evaluation of Gaze Estimation Algorithms from RGB and RGB-D Cameras. In *ETRA '14*. ACM, 255–258.

[71] Jason Geller, Matthew B Winn, Tristian Mahr, and Daniel Mirman. 2020. GazeR: A package for processing gaze position and pupil size data. *Behavior research methods* 52, 5 (2020), 2232–2255.

[72] Ceenu George, Daniel Buschek, Andrea Ngao, and Mohamed Khamis. 2020. GazeRoomLock: Using Gaze and Head-Pose to Improve the Usability and Observation Resistance of 3D Passwords in Virtual Reality. In *Augmented Reality, Virtual Reality, and Computer Graphics*, Lucio Tommaso De Paolis and Patrick Bourdot (Eds.). Springer, 61–81.

[73] Upamanyu Ghose, Arvind A Srinivasan, W Paul Boyce, Hong Xu, and Eng Siong Chng. 2020. PyTrack: An end-to-end analysis toolkit for eye tracking. *Behavior research methods* 52, 6 (2020), 2588–2603.

[74] Shreya Ghosh, Abhinav Dhall, Munawar Hayat, Jarrod Knibbe, and Qiang Ji. 2022. Automatic Gaze Analysis: A Survey of Deep Learning based Approaches. arXiv:2108.05479 [cs.CV]

[75] Darren R Gitelman. 2002. ILAB: a program for postexperimental eye movement analysis. *Behavior Research Methods, Instruments, & Computers* 34, 4 (2002), 605–612.

[76] Grand View Research. 2022. *Eye Tracking Market Size & Share Report, 2022 - 2030*. grandviewresearch. https://www.grandviewresearch.com/industry-analysis/eye-tracking-market

[77] Nishan Gunawardena, Jeewani Anupama Ginige, and Bahman Javadi. 2022. Eye-Tracking Technologies in Mobile Devices Using Edge Computing: A Systematic Review. *ACM Comput. Surv.* 55, 8, Article 158 (dec 2022), 33 pages.

[78] Tianchu Guo, Yongchao Liu, Hui Zhang, Xiabing Liu, Youngjun Kwak, Byung In Yoo, Jae-Joon Han, and Changkyu Choi. 2019. A Generalized and Robust Method Towards Practical Gaze Estimation on Smart Phone. arXiv:1910.07331 [cs.CV]

[79] Zidong Guo, Zejian Yuan, Chong Zhang, Wanchao Chi, Yonggen Ling, and Shenghao Zhang. 2020. Domain adaptation gaze estimation by embedding with prediction consistency. In *the Asian Conference on Computer Vision*. Springer, 292–307.

[80] John Paulin Hansen, Haakon Lund, Florian Biermann, Emillie Møllenbach, Sebastian Sztuk, and Javier San Agustin. 2016. Wrist-Worn Pervasive Gaze Interaction. In *ETRA '16*. ACM, 57–64.

[81] Katarzyna Harezlak and Pawel Kasprowski. 2018. Application of eye tracking in medicine: A survey, research issues and challenges. *Computerized Medical Imaging and Graphics* 65 (2018), 176–190.

[82] Junfeng He, Khoi Pham, Nachiappan Valliappan, Pingmei Xu, Chase Roberts, Dmitry Lagun, and Vidhya Navalpakkam. 2019. On-Device Few-Shot Personalization for Real-Time Gaze Estimation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 1149–1158.

[83] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR '16*. IEEE, 770–778.

[84] Qiuhai He, Xiaopeng Hong, Xiujuan Chai, Jukka Holappa, Guoying Zhao, Xilin Chen, and Matti Pietikäinen. 2015. OMEG: Oulu Multi-Pose Eye Gaze Dataset. In *Image Analysis*, Rasmus R. Paulsen and Kim S. Pedersen (Eds.). Springer, 418–427.

[85] Zhenyi He, Keru Wang, Brandon Yushan Feng, Ruofei Du, and Ken Perlin. 2021. GazeChat: Enhancing Virtual Conferences with Gaze-Aware 3D Photos. In *UIST '21*. ACM, 769–782.

[86] Henna Heikkilä and Kari-Jouko Räihä. 2012. Simple Gaze Gestures and the Closure of the Eyes as an Interaction Technique. In *ETRA '12*. ACM, 147–154.

[87] Oliver Hein and Wolfgang Zangemeister. 2017. Topology for gaze analyses-Raw data segmentation. Retrieved April 19, 2023 from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7141061/

[88] Helena Hemmingsson and Maria Borgestig. 2020. Usability of eye-gaze controlled computers in Sweden: A total population survey. *International journal of environmental research and public health* 17, 5 (2020), 1639.

[89] Roy S Hessels, Richard Andersson, Ignace TC Hooge, Marcus Nyström, and Chantal Kemner. 2015. Consequences of eye color, positioning, and head movement for eye-tracking data quality in infant research. *Infancy* 20, 6 (2015), 601–633.

[90] Roy S Hessels, Chantal Kemner, Carlijn van den Boomen, and Ignace TC Hooge. 2016. The area-of-interest problem in eyetracking research: A noise-robust solution for face and sparse stimuli. *Behavior research methods* 48, 4 (2016), 1694–1712.

[91] Teresa Hirzle, Maurice Cordts, Enrico Rukzio, and Andreas Bulling. 2020. A Survey of Digital Eye Strain in Gaze-Based Interactive Systems. In *ETRA '20*. ACM, Article 9, 12 pages.

[92] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[93] Sabrina Hoppe, Tobias Loetscher, Stephanie A. Morey, and Andreas Bulling. 2018. Eye Movements During Everyday Behavior Predict Personality Traits. *Frontiers in Human Neuroscience* 12 (2018), 105. https://www.frontiersin.org/articles/10.3389/fnhum.2018.00105

[94] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *CVPR '17*. IEEE, 4700–4708.

[95] Gary B. Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. 2008. Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*. Erik Learned-Miller and Andras Ferencz and Frédéric Jurie, Inria, 1–14. https://hal.inria.fr/inria-00321923

[96] Haoxian Huang, Luqian Ren, Zhuo Yang, Yinwei Zhan, Qieshi Zhang, and Jujian Lv. 2022. GAZEATTENTIONNET: Gaze Estimation with Attentions. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2435–2439.

[97] Michael Xuelin Huang and Andreas Bulling. 2019. SacCalib: Reducing Calibration Distortion for Stationary Eye Trackers Using Saccadic Eye Movements. In *ETRA '19*. ACM, Article 71, 10 pages.

[98] Michael Xuelin Huang, Tiffany C.K. Kwok, Grace Ngai, Stephen C.F. Chan, and Hong Va Leong. 2016. Building a Personalized, Auto-Calibrating Eye Tracker from User Interactions. In *CHI '16*. ACM, 5169–5179.

[99] Michael Xuelin Huang, Jiajia Li, Grace Ngai, and Hong Va Leong. 2016. Stressclick: Sensing stress from gaze-click patterns. In *the 24th ACM international conference on Multimedia*. 1395–1404.

[100] Michael Xuelin Huang, Jiajia Li, Grace Ngai, and Hong Va Leong. 2017. ScreenGlint: Practical, In-Situ Gaze Estimation on Smartphones. In *CHI '17*. ACM, 2546–2557.

[101] Qiong Huang, Ashok Veeraraghavan, and Ashutosh Sabharwal. 2017. TabletGaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. *Machine Vision and Applications* 28, 5 (2017), 445–461.

[102] Sinh Huynh, Rajesh Krishna Balan, and JeongGil Ko. 2021. iMon: Appearance-based gaze tracking system on mobile devices. *IMWUT* 5, 4 (2021), 1–26.

[103] Yoon Min Hwang and Kun Chang Lee. 2022. An eye-tracking paradigm to explore the effect of online consumers' emotion on their visual behaviour between desktop screen and mobile screen. *Behaviour & Information Technology* 41, 3 (2022), 535–546.

[104] Robert J. K. Jacob. 1991. The Use of Eye Movements in Human-Computer Interaction Techniques: What You Look at is What You Get. *ACM Trans. Inf. Syst.* 9, 2 (apr 1991), 152–169.

[105] Marium-E Jannat, Thuan T Vo, and Khalad Hasan. 2022. Face-Centered Spatial User Interfaces on Smartwatches. In *CHI EA '22*. ACM, Article 393, 7 pages.

[106] Sumit Jha and Carlos Busso. 2022. Estimation of driver's gaze region from head position and orientation using probabilistic confidence regions. *IEEE Transactions on Intelligent Vehicles* 8, 1 (2022), 59–72.

[107] Xinhui Jiang, Jussi PP Jokinen, Antti Oulasvirta, and Xiangshi Ren. 2022. Learning to type with mobile keyboards: Findings with a randomized keyboard. *Computers in Human Behavior* 126 (2022), 106992.

[108] Xinhui Jiang, Yang Li, Jussi P.P. Jokinen, Viet Ba Hirvola, Antti Oulasvirta, and Xiangshi Ren. 2020. How We Type: Eye and Finger Movement Strategies in Mobile Typing. In *CHI '20*. ACM, 1–14.

[109] Swati Jindal and Roberto Manduchi. 2022. Contrastive Representation Learning for Gaze Estimation. arXiv:2210.13404 [cs.CV]

[110] Florian Jungwirth, Michael Haslgrübler, and Alois Ferscha. 2018. Contour-Guided Gaze Gestures: Using Object Contours as Visual Guidance for Triggering Interactions. In *ETRA '18*. ACM, Article 28, 10 pages.

[111] Anuradha Kar and Peter Corcoran. 2017. A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms. *IEEE Access* 5 (2017), 16495–16519.

[112] Pragma Kar, Samiran Chattopadhyay, and Sandip Chakraborty. 2020. Gestatten: Estimation of User's Attention in Mobile MOOCs From Eye Gaze and Gaze Gesture Tracking. *Proc. ACM Hum.-Comput. Interact.* 4, EICS (2020), 1–32.

[113] Keith S. Karn. 2000. "Saccade Pickers" vs. "Fixation Pickers": The Effect of Eye Tracking Instrumentation on Research. In *ETRA '00*. ACM, 87–88.

[114] Pawel Kasprowski and Katarzyna Harezlak. 2016. Implicit Calibration Using Predicted Gaze Targets. In *ETRA '16*. ACM, 245–248.

[115] Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-Based Interaction. In *UbiComp '14 Adjunct*. ACM, 1151–1160.

[116] Christina Katsini, Yasmeen Abdrabou, George E. Raptis, Mohamed Khamis, and Florian Alt. 2020. The Role of Eye Gaze in Security and Privacy Applications: Survey and Future HCI Research Directions. In *CHI '20*. ACM, 1–21.

[117] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. 2019. Gaze360: Physically unconstrained gaze estimation in the wild. In *the IEEE/CVF International Conference on Computer Vision*. IEEE, 6912–6921.

[118] Mohamed Khamis, Florian Alt, and Andreas Bulling. 2018. The Past, Present, and Future of Gaze-Enabled Handheld Mobile Devices: Survey and Lessons Learned. In *MobileHCI '18*. ACM, Article 38, 17 pages.

[119] Mohamed Khamis, Anita Baier, Niels Henze, Florian Alt, and Andreas Bulling. 2018. Understanding Face and Eye Visibility in Front-Facing Cameras of Smartphones Used in the Wild. In *CHI '18*. ACM, 1–12.

[120] Mohamed Khamis, Regina Hasholzner, Andreas Bulling, and Florian Alt. 2017. GTmoPass: Two-Factor Authentication on Public Displays Using Gaze-Touch Passwords and Personal Mobile Devices. In *PerDis '17*. ACM, Article 8, 9 pages.

[121] Mohamed Khamis, Mariam Hassib, Emanuel von Zezschwitz, Andreas Bulling, and Florian Alt. 2017. GazeTouchPIN: Protecting Sensitive Data on Mobile Devices Using Secure Multimodal Authentication. In *ICMI '17*. ACM, 446–450.

[122] Mohamed Khamis, Karola Marky, Andreas Bulling, and Florian Alt. 2022. User-centred multimodal authentication: securing handheld mobile devices using gaze and touch input. *Behaviour & Information Technology* 41, 10 (2022), 2061–2083.

[123] Anam Ahmad Khan, Joshua Newn, Ryan M Kelly, Namrata Srivastava, James Bailey, and Eduardo Velloso. 2021. GAVIN: Gaze-assisted voice-based implicit note-taking. *ACM Transactions on Computer-Human Interaction (TOCHI)* 28, 4 (2021), 1–32.

[124] Joohwan Kim, Michael Stengel, Alexander Majercik, Shalini De Mello, David Dunn, Samuli Laine, Morgan McGuire, and David Luebke. 2019. NVGaze: An Anatomically-Informed Dataset for Low-Latency, Near-Eye Gaze Estimation. In *CHI '19*. ACM, 1–12.

[125] Jaewon Kim, Paul Thomas, Ramesh Sankaranarayana, Tom Gedeon, and Hwan-Jin Yoon. 2016. Pagination versus Scrolling in Mobile Web Search. In *ACM International on Conference on Information and Knowledge Management*. ACM, 751–760.

[126] Jaewon Kim, Paul Thomas, Ramesh Sankaranarayana, Tom Gedeon, and Hwan-Jin Yoon. 2016. Understanding eye movements on mobile devices for better presentation of search results. *Journal of the Association for Information Science and Technology* 67, 11 (2016), 2607–2619.

[127] Jung-Hwa Kim, Seung-June Choi, and Jin-Woo Jeong. 2019. Watch & Do: A smart iot interaction system with object detection and gaze estimation. *IEEE Transactions on Consumer Electronics* 65, 2 (2019), 195–204.

[128] Davis E. King. 2009. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* 10 (2009), 1755–1758.

[129] Reinhold Kliegl and Richard K Olson. 1981. Reduction and calibration of eye monitor data. *Behavior Research Methods & Instrumentation* 13, 2 (1981), 107–111.

[130] Christof Koch and Shimon Ullman. 1987. *Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry*. Springer Netherlands, 115–141.

[131] Oleg V Komogortsev, Denise V Gobert, Sampath Jayarathna, Sandeep M Gowda, et al. 2010. Standardization of automated analyses of oculomotor fixation and saccadic behaviors. *IEEE Transactions on biomedical engineering* 57, 11 (2010), 2635–2645.

[132] Oleg V Komogortsev and Alex Karpov. 2013. Automated classification and scoring of smooth pursuit eye movements in the presence of fixations and saccades. *Behavior research methods* 45, 1 (2013), 203–215.

[133] Andy Kong, Karan Ahuja, Mayank Goel, and Chris Harrison. 2021. EyeMU Interactions: Gaze + IMU Gestures on Mobile Devices. In *ICMI '21*. ACM, 577–585.

[134] Rakshit Kothari, Shalini De Mello, Umar Iqbal, Wonmin Byeon, Seonwook Park, and Jan Kautz. 2021. Weakly-supervised physically unconstrained gaze estimation. In *CVPR '21*. IEEE, 9980–9989.

[135] Rakshit Kothari, Zhizhuo Yang, Christopher Kanan, Reynold Bailey, Jeff B Pelz, and Gabriel J Diaz. 2020. Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities. *Scientific reports* 10, 1 (2020), 1–18.

[136] Marek Kowalski, Jacek Naruniec, and Tomasz Trzcinski. 2017. Deep alignment network: A convolutional neural network for robust face alignment. In *CVPR '17 workshops*. IEEE, 88–97.

[137] Eileen Kowler, Jason F Rubinstein, Elio M Santos, and Jie Wang. 2019. Predictive smooth pursuit eye movements. *Annual review of vision science* 5 (2019), 223–246.

[138] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. 2016. Eye tracking for everyone. In *CVPR '16*. IEEE, 2176–2184.

[139] Vassilios Krassanakis, Vassiliki Filippakopoulou, and Byron Nakos. 2014. EyeMMV toolbox: An eye movement post-analysis tool based on a two-step spatial dispersion threshold for fixation identification. *Journal of Eye Movement Research* 7, 1 (Feb. 2014), 1–10.

[140] Vinay Krishna Sharma, Kamalpreet Saluja, Vimal Mollyn, and Pradipta Biswas. 2020. Eye Gaze Controlled Robotic Arm for Persons with Severe Speech and Motor Impairment. In *ETRA '20*. ACM, Article 12, 9 pages.

[141] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* 60, 6 (may 2017), 84–90.

[142] Chandan Kumar, Raphael Menges, Daniel Müller, and Steffen Staab. 2017. Chromium Based Framework to Include Gaze Interaction in Web Browser. In *WWW '17 Companion*. International World Wide Web Conferences Steering Committee, 219–223.

[143] Manu Kumar, Terry Winograd, and Andreas Paepcke. 2007. Gaze-Enhanced Scrolling Techniques. In *CHI EA '07*. ACM, 2531–2536.

[144] Grete Helena Kütt, Teerapaun Tanprasert, Jay Rodolitz, Bernardo Moyza, Samuel So, Georgia Kenderova, and Alexandra Papoutsaki. 2020. Effects of shared gaze on audio-versus text-based remote collaborations. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2 (2020),

1–25.

[145] Dmitry Lagun, Chih-Hung Hsieh, Dale Webster, and Vidhya Navalpakkam. 2014. Towards Better Measurement of Attention and Satisfaction in Mobile Search. In *SIGIR '14*. ACM, 113–122.

[146] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[147] Yaxiong Lei, Yuheng Wang, Tyler Caslin, Alexander Wisowaty, Xu Zhu, Mohamed Khamis, and Juan Ye. 2023. DynamicRead: Exploring Robust Gaze Interaction Methods for Reading on Handheld Mobile Devices under Dynamic Conditions. *Proc. ACM Hum.-Comput. Interact.* 7, ETRA23 (5 2023), 17.

[148] Ryan Lewien. 2021. GazeHelp: Exploring Practical Gaze-Assisted Interactions for Graphic Design Tools. In *ETRA '21 (ETRA '21 Adjunct)*. ACM, Article 1, 4 pages.

[149] Runtong Li, Huimin Ma, Rongquan Wang, and Jiawei Ding. 2021. Device-Adaptive 2D Gaze Estimation: A Multi-Point Differential Framework. In *Image and Graphics*, Yuxin Peng, Shi-Min Hu, Moncef Gabbouj, Kun Zhou, Michael Elad, and Kun Xu (Eds.). Springer, 485–497.

[150] Yixuan Li, Pingmei Xu, Dmitry Lagun, and Vidhya Navalpakkam. 2017. Towards Measuring and Inferring User Interest from Gaze. In *WWW '17 Companion*. International World Wide Web Conferences Steering Committee, 525–533.

[151] Yuqing Li, Yinwei Zhan, and Zhuo Yang. 2020. Evaluation of appearance-based eye tracking calibration data selection. In *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*. IEEE, 222–224.

[152] Zhenxing Li, Deepak Akkil, and Roope Raisamo. 2020. Gaze-based kinaesthetic interaction for virtual reality. *Interacting with Computers* 32, 1 (2020), 17–32.

[153] Zhi Li, Maozheng Zhao, Yifan Wang, Sina Rashidian, Furqan Baig, Rui Liu, Wanyu Liu, Michel Beaudouin-Lafon, Brooke Ellison, Fusheng Wang, IV Ramakrishnan, and Xiaojun Bi. 2021. BayesGaze: A Bayesian Approach to Eye-Gaze Based Target Selection. In *Graphics Interface 2021*. Canadian Information Processing Society, 231 – 240.

[154] Dongze Lian, Lina Hu, Weixin Luo, Yanyu Xu, Lixin Duan, Jingyi Yu, and Shenghua Gao. 2018. Multiview multitask gaze estimation with deep convolutional neural networks. *IEEE transactions on neural networks and learning systems* 30, 10 (2018), 3010–3023.

[155] Dongze Lian, Ziheng Zhang, Weixin Luo, Lina Hu, Minye Wu, Zechao Li, Jingyi Yu, and Shenghua Gao. 2019. RGBD Based Gaze Estimation via Multi-Task CNN. *AAAI* 33, 01 (Jul. 2019), 2488–2495.

[156] Daniel J. Liebling and Susan T. Dumais. 2014. Gaze and Mouse Coordination in Everyday Work. In *UbiComp '14: Adjunct*. ACM, 1141–1150.

[157] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *International Conference on Computer Vision (ICCV)*. IEEE, 3730–3738.

[158] Chaochao Lu and Xiaoou Tang. 2014. Learning the Face Prior for Bayesian Face Recognition. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer, 119–134.

[159] Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato. 2014. Adaptive linear regression for appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence* 36, 10 (2014), 2033–2046.

[160] Päivi Majaranta, Ulla-Kaija Ahola, and Oleg Špakov. 2009. Fast Gaze Typing with an Adjustable Dwell Time. In *CHI '09*. ACM, 357–360.

[161] Päivi Majaranta and Andreas Bulling. 2014. *Eye Tracking and Eye-Based Human–Computer Interaction.* Springer London, 39–65.

[162] Päivi Majaranta, Jari Laitinen, Jari Kangas, and Poika Isokoski. 2019. Inducing Gaze Gestures by Static Illustrations. In *ETRA '19*. ACM, Article 75, 5 pages.

[163] Barry R Manor and Evian Gordon. 2003. Defining the temporal threshold for ocular fixation in free-viewing visuocognitive tasks. *Journal of neuroscience methods* 128, 1-2 (2003), 85–93.

[164] Susana Martinez-Conde, Stephen L Macknik, and David H Hubel. 2004. The role of fixational eye movements in visual perception. *Nature reviews neuroscience* 5, 3 (2004), 229–240.

[165] Susana Martinez-Conde, Jorge Otero-Millan, and Stephen L Macknik. 2013. The impact of microsaccades on vision: towards a unified theory of saccadic function. *Nature Reviews Neuroscience* 14, 2 (2013), 83–96.

[166] Pranay Mathur, Nikhil Khedekar, and Kostas Alexis. 2021. Resource-aware Online Parameter Adaptation for Computationally -constrained Visual-Inertial Navigation Systems. In *2021 20th International Conference on Advanced Robotics (ICAR)*. IEEE, 842–848.

[167] Sven Mayer, Gierad Laput, and Chris Harrison. 2020. Enhancing Mobile Voice Assistants with WorldGaze. In *CHI '20*. ACM, 1–10.

[168] Raphael Menges, Chandan Kumar, Ulrich Wechselberger, Christoph Schaefer, Tina Walber, and Steffen Staab. 2017. Schau genau! A gaze-controlled 3D game for entertainment and education. *Journal of Eye Movement Research* 10, 6 (2017), 220.

[169] Caitlin Mills, Julie Gregg, Robert Bixler, and Sidney K D'Mello. 2021. Eye-mind reader: An intelligent reading interface that promotes long-term comprehension by detecting and responding to mind wandering. *Human–Computer Interaction* 36, 4 (2021), 306–332.

[170] Callum Mole, Jami Pekkanen, William EA Sheppard, Gustav Markkula, and Richard M Wilkie. 2021. Drivers use active gaze to monitor waypoints during automated driving. *Scientific Reports* 11, 1 (2021), 1–18.

[171] Kenneth Alberto Funes Mora and Jean-Marc Odobez. 2013. Person independent 3d gaze estimation from remote rgb-d cameras. In *2013 IEEE International Conference on Image Processing*. IEEE, 2787–2791.

[172] Martez E. Mott, Shane Williams, Jacob O. Wobbrock, and Meredith Ringel Morris. 2017. Improving Dwell-Based Gaze Typing with Dynamic, Cascading Dwell Times. In *CHI '17*. ACM, 2558–2570.

[173] Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. 2018. Detecting Low Rapport During Natural Interactions in Small Groups from Non-Verbal Behaviour. In *IUI '18*. ACM, 153–164.

[174] Vikrant Nagpure and Kenji Okuma. 2023. Searching Efficient Neural Architecture With Multi-Resolution Fusion Transformer for Appearance-Based Gaze Estimation. In *the IEEE/CVF Winter Conference on Applications of Computer Vision*. IEEE, 890–899.

[175] Omar Namnakani, Yasmeen Abdrabou, Jonathan Grizou, Augusto Esteves, and Mohamed Khamis. 2023. Comparing Dwell Time, Pursuits and Gaze Gestures for Gaze Interaction on Handheld Mobile Devices. In *CHI '23*. ACM, Article 258, 17 pages.

[176] Matei Negulescu, Jaime Ruiz, and Edward Lank. 2012. A Recognition Safety Net: Bi-Level Threshold Recognition for Mobile Motion Gestures. In *MobileHCI '12*. ACM, 147–150.

[177] Anelise Newman, Barry McNamara, Camilo Fosco, Yun Bin Zhang, Pat Sukhum, Matthew Tancik, Nam Wook Kim, and Zoya Bylinskii. 2020. TurkEyes: A Web-Based Toolbox for Crowdsourcing Attention Data. In *CHI '20*. ACM, 1–13.

[178] Jun O Oh, Hyung Jin Chang, and Sang-Il Choi. 2022. Self-attention with convolution and deconvolution for efficient eye gaze estimation from a full face image. In *CVPR '22*. IEEE, 4992–5000.

[179] Reo Ogusu and Takao Yamanaka. 2019. LPM: learnable pooling module for efficient full-face gaze estimation. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 1–5.

[180] Jakob Ohme, Ewa Maslowska, and Cornelia Mothes. 2022. Mobile News Learning — Investigating Political Knowledge Gains in a Social Media Newsfeed with Mobile Eye Tracking. *Political Communication* 39, 3 (2022), 339–357.

[181] Cristina Palmero, Javier Selva, Mohammad Ali Bagheri, and Sergio Escalera. 2018. Recurrent CNN for 3D Gaze Estimation using Appearance and Shape Cues. arXiv:1805.03064 [cs.CV]

[182] Yunxian Pan, Qinyu Zhang, Yifan Zhang, Xianliang Ge, Xiaoqing Gao, Shiyan Yang, and Jie Xu. 2022. Lane-change intention prediction using eye-tracking technology: A systematic review. *Applied Ergonomics* 103 (2022), 103775.

[183] Mohsen Parisay, Charalambos Poullis, and Marta Kersten-Oertel. 2021. EyeTAP: Introducing a multimodal gaze-based technique using voice inputs with a comparative analysis of selection techniques. *International Journal of Human-Computer Studies* 154 (oct 2021), 102676.

[184] Joonbeom Park, Seonghoon Park, and Hojung Cha. 2021. GAZEL: Runtime Gaze Tracking for Smartphones. In *2021 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 1–10.

[185] Seonwook Park, Emre Aksan, Xucong Zhang, and Otmar Hilliges. 2020. Towards End-to-End Video-Based Eye-Tracking. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer, 747–763.

[186] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. 2019. Few-shot adaptive gaze estimation. In *the IEEE/CVF International Conference on Computer Vision*. IEEE, 9368–9377.

[187] Seonwook Park, Adrian Spurr, and Otmar Hilliges. 2018. Deep pictorial gaze estimation. In *the European Conference on Computer Vision (ECCV)*. IEEE, 721–738.

[188] O Parkhi, A Vedaldi, and A Zisserman. 2015. Deep face recognition. In *the British Machine Vision Conference 2015 (BMVC 2015)*. British Machine Vision Association, 1–12.

[189] Timo Partala, Maria Jokiniemi, and Veikko Surakka. 2000. Pupillary Responses to Emotionally Provocative Stimuli. In *ETRA '00*. ACM, 123–129.

[190] Bharat Paudyal, Chris Creed, Maite Frutos-Pascual, and Ian Williams. 2020. Voiceye: A Multimodal Inclusive Development Environment. In *DIS '20*. ACM, 21–33.

[191] Yesaya Tommy Paulus and Gerard Bastiaan Remijn. 2021. Usability of various dwell times for eye-gaze-based object selection with eye tracking. *Displays* 67 (2021), 101997.

[192] Ken Pfeuffer, Jason Alexander, and Hans Gellersen. 2021. Multi-User Gaze-Based Interaction Techniques on Collaborative Touchscreens. In *ETRA '21*. ACM, Article 26, 7 pages.

[193] Ken Pfeuffer and Hans Gellersen. 2016. Gaze and Touch Interaction on Tablets. In *UIST '16*. ACM, 301–311.

[194] Carmelo Pino and Isaak Kavasidis. 2012. Improving mobile device interaction by eye tracking analysis. In *2012 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 1199–1202.

[195] Tobbi Pro. 2015. *How to position the eye tracker and participant in a study*. Tobii AB. Retrieved April 19, 2023 from https://www.tobiipro.com/learn-and-support/learn/steps-in-an-eye-tracking-study/run/how-to-position-the-participant-and-the-eye-tracker/

[196] Francine Prokoski. 2000. History, current status, and future of infrared identification. In *Proceedings IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications (Cat. No. PR00640)*. IEEE, 5–14.

[197] Kirill Ragozin, Yun Suen Pai, Olivier Augereau, Koichi Kise, Jochen Kerdels, and Kai Kunze. 2019. Private Reader: Using Eye Tracking to Improve Reading Privacy in Public Spaces. In *MobileHCI '19*. ACM, Article 18, 6 pages.

[198] Vijay Rajanna and Tracy Hammond. 2018. A Gaze Gesture-Based Paradigm for Situational Impairments, Accessibility, and Rich Interactions. In *ETRA '18*. ACM, Article 102, 3 pages.

[199] Argenis Ramirez Ramirez Gomez, Christopher Clarke, Ludwig Sidenmark, and Hans Gellersen. 2021. Gaze+Hold: Eyes-Only Direct Manipulation with Continuous Gaze Modulated by Closure of One Eye. In *ETRA '21*. ACM, Article 10, 12 pages.

[200] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. 2015. Where are they looking?. In *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.), Vol. 28. Curran Associates, Inc., 1–9.

[201] Luz Rello and Miguel Ballesteros. 2015. Detecting Readers with Dyslexia Using Machine Learning with Eye Tracking Measures. In *International Web for All Conference*. ACM, Article 16, 8 pages.

[202] Sheikh Rivu, Yasmeen Abdrabou, Thomas Mayer, Ken Pfeuffer, and Florian Alt. 2019. GazeButton: Enhancing Buttons with Eye Gaze Interactions. In *ETRA '19*. ACM, Article 73, 7 pages.

[203] D Ar Robinson, JL Gordon, and SE Gordon. 1986. A model of the smooth pursuit eye movement system. *Biological cybernetics* 55, 1 (1986), 43–57.

[204] Martin Rolfs. 2009. Microsaccades: small steps on a long way. *Vision research* 49, 20 (2009), 2415–2441.

[205] David Rozado, Javier S. Agustin, Francisco B. Rodriguez, and Pablo Varona. 2012. Gliding and Saccadic Gaze Gesture Recognition in Real Time. *ACM Trans. Interact. Intell. Syst.* 1, 2, Article 10 (jan 2012), 27 pages.

[206] David Rozado, T Moreno, Javier San Agustin, FB Rodriguez, and Pablo Varona. 2015. Controlling a smartphone using gaze gestures as the input mechanism. *Human–Computer Interaction* 30, 1 (2015), 34–63.

[207] David Rozado, Francisco B. Rodriguez, and Pablo Varona. 2012. Low cost remote gaze gesture recognition in real time. *Applied Soft Computing* 12, 8 (2012), 2072–2084.

[208] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature* 323, 6088 (1986), 533–536.

[209] Aimee E Ryan, Brendan Keane, and Guy Wallis. 2019. Microsaccades and covert attention: Evidence from a continuous, divided attention task. *Journal of Eye Movement Research* 12, 6 (2019), 1–11.

[210] Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying Fixations and Saccades in Eye-Tracking Protocols. In *ETRA '00*. ACM, 71–78.

[211] Thiago Santini, Wolfgang Fuhl, Thomas Kübler, and Enkelejda Kasneci. 2016. Bayesian Identification of Fixations, Saccades, and Smooth Pursuits. In *ETRA '16*. ACM, 163–170.

[212] Hosnieh Sattar, Sabine Muller, Mario Fritz, and Andreas Bulling. 2015. Prediction of search targets from fixations in open-world settings. In *CVPR '15*. IEEE, 981–990.

[213] Simon Schenk, Marc Dreiser, Gerhard Rigoll, and Michael Dorr. 2017. GazeEverywhere: Enabling Gaze-Only User Interaction on an Unmodified Desktop PC in Everyday Scenarios. In *CHI '17*. ACM, 3034–3044.

[214] Christian Schlösser, Benedikt Schröder, Linda Cedli, and Andrea Kienle. 2018. Beyond Gaze Cursor: Exploring Information-Based Gaze Sharing in Chat. In *COGAIN '18*. ACM, Article 10, 5 pages.

[215] B Shackel. 1960. Pilot study in electro-oculography. *The British journal of ophthalmology* 44, 2 (1960), 89.

[216] Lei Shi, Cosmin Copot, and Steve Vanlanduit. 2021. Gaze Gesture Recognition by Graph Convolutional Networks. *Frontiers in Robotics and AI* 8 (2021), 709952.

[217] Emiko Shishido, Shiori Ogawa, Seiko Miyata, Maeri Yamamoto, Toshiya Inada, and Norio Ozaki. 2019. Application of eye trackers for understanding mental disorders: Cases for schizophrenia and autism spectrum disorder. *Neuropsychopharmacology reports* 39, 2 (2019), 72–77.

[218] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs.CV]

[219] Shyamli Sindhwani, Christof Lutteroth, and Gerald Weber. 2019. ReType: Quick Text Editing with Keyboard and Gaze. In *CHI '19*. ACM, 13 pages.

[220] Joshua Sink, Stephen Blatt, David Yoo, Michael Henry, S Daniel Yang, Roshni Vasaiwala, Larissa Ghadiali, William Adams, and Charles S Bouchard. 2020. A novel telemedicine technique for evaluation of ocular exam findings via smartphone images. *Journal of Telemedicine and Telecare* 28, 3 (2020), 197–202.

[221] Smart Eye. 2022. *Driver Monitoring System - Smart Eye*. Smart Eye Co., Ltd. https://smarteye.se/solutions/automotive/driver-monitoring-system/

[222] Brian A. Smith, Qi Yin, Steven K. Feiner, and Shree K. Nayar. 2013. Gaze Locking: Passive Eye Contact Detection for Human-Object Interaction. In *UIST '13*. ACM, 271–280.

[223] Hiroyuki Sogo. 2013. GazeParser: an open-source and multiplatform library for low-cost eye tracking and analysis. *Behavior research methods* 45, 3 (2013), 684–695.

[224] Chen Song, Aosen Wang, Kui Ren, and Wenyao Xu. 2016. Eyeveri: A secure and usable approach for smartphone user authentication. In *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*. IEEE, 1–9.

[225] Cecie Starr, Christine Evers, and Lisa Starr. 2014. *Biology: concepts and applications*. Cengage Learning.

[226] Mikhail Startsev, Ioannis Agtzidis, and Michael Dorr. 2019. 1D CNN with BLSTM for automated classification of fixations, saccades, and smooth pursuits. *Behavior Research Methods* 51, 2 (2019), 556–572.

[227] Julian Steil, Philipp Müller, Yusuke Sugano, and Andreas Bulling. 2018. Forecasting User Attention during Everyday Mobile Interactions Using Device-Integrated and Wearable Sensors. In *MobileHCI '18*. ACM, Article 1, 13 pages.

[228] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. 2014. Learning-by-synthesis for appearance-based 3d gaze estimation. In *CVPR '14*. IEEE, 1821–1828.

[229] Tobbi. 2020. Data quality reports for 3 Tobii eye trackers - Tobii. Retrieved April 19, 2023 from https://www.tobii.com/resource-center/data-quality#cta-section

[230] Henri Tomas, Marcus Reyes, Raimarc Dionido, Mark Ty, Jonric Mirando, Joel Casimiro, Rowel Atienza, and Richard Guinto. 2021. Goo: A dataset for gaze object prediction in retail environments. In *CVPR '21*. IEEE, 3125–3133.

[231] Marc Tonsen, Julian Steil, Yusuke Sugano, and Andreas Bulling. 2017. Invisibleeye: Mobile eye tracking using multiple low-resolution cameras and learning-based gaze estimation. *IMWUT* 1, 3 (2017), 1–21.

[232] Nachiappan Valliappan, Na Dai, Ethan Steinberg, Junfeng He, Kantwon Rogers, Venky Ramachandran, Pingmei Xu, Mina Shojaeizadeh, Li Guo, Kai Kohlhoff, et al. 2020. Accelerating eye movement research via accurate and affordable smartphone eye tracking. *Nature communications* 11, 1 (2020), 1–12.

[233] Daan R van Renswoude, Maartje EJ Raijmakers, Arnout Koornneef, Scott P Johnson, Sabine Hunnius, and Ingmar Visser. 2018. Gazepath: An eye-tracking analysis tool that accounts for individual differences and data quality. *Behavior research methods* 50, 2 (2018), 834–852.

[234] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[235] Eduardo Velloso, Marcus Carter, Joshua Newn, Augusto Esteves, Christopher Clarke, and Hans Gellersen. 2017. Motion Correlation: Selecting Objects by Matching Their Movement. *ACM Trans. Comput.-Hum. Interact.* 24, 3, Article 22 (apr 2017), 35 pages.

[236] Pranav Venuprasad, Tushal Dobhal, Anurag Paul, Tu N. M. Nguyen, Andrew Gilman, Pamela Cosman, and Leanne Chukoskie. 2019. Characterizing Joint Attention Behavior during Real World Interactions Using Automated Object and Gaze Detection. In *ETRA '19*. ACM, Article 21, 8 pages.

[237] Tore Vesterby, Jonas C. Voss, John Paulin Hansen, Arne John Glenstrup, Dan Witzner Hansen, and Mark Rudolph. 2005. Gaze-guided viewing of interactive movies. *Digital Creativity* 16, 4 (2005), 193–204.

[238] Mélodie Vidal, Andreas Bulling, and Hans Gellersen. 2013. Pursuits: Spontaneous Interaction with Displays Based on Smooth Pursuit Eye Movement and Moving Targets. In *UbiComp '13*. ACM, 439–448.

[239] Simon Voelker, Sebastian Hueber, Christian Holz, Christian Remy, and Nicolai Marquardt. 2020. GazeConduits: Calibration-Free Cross-Device Collaboration through Gaze and Touch. In *CHI '20*. ACM, 1–10.

[240] Sourabh Vora, Akshay Rangesh, and Mohan M Trivedi. 2017. On generalizing driver gaze zone estimation using convolutional neural networks. In *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, IEEE, 849–854.

[241] Kang Wang, Hui Su, and Qiang Ji. 2019. Neuro-inspired eye tracking with eye movement dynamics. In *CVPR '19*. IEEE, 9831–9840.

[242] Kang Wang, Rui Zhao, Hui Su, and Qiang Ji. 2019. Generalizing eye tracking with bayesian adversarial learning. In *CVPR '19*. IEEE, 11907–11916.

[243] Yao Wang, Mihai Bâ ce, and Andreas Bulling. 2023. Scanpath Prediction on Information Visualisations. *IEEE Transactions on Visualization and Computer Graphics* Early Access (2023), 1–15.

[244] Sam V Wass, Linda Forssman, and Jukka Leppänen. 2014. Robustness and precision: How data quality may influence key dependent variables in infant eye-tracker analyses. *Infancy* 19, 5 (2014), 427–460.

[245] Samuel Wehrli, Corinna Hertweck, Mohammadreza Amirian, Stefan Glüge, and Thilo Stadelmann. 2021. Bias, awareness, and ignorance in deep-learning-based face recognition. *AI and Ethics* 2 (2021), 509–522.

[246] Pierre Weill-Tessier, Jayson Turner, and Hans Gellersen. 2016. How Do You Look at What You Touch? A Study of Touch Interaction and Gaze Correlation on Tablets. In *ETRA '16*. ACM, 329–330.

[247] Andrew D. Wilson and Shane Williams. 2018. Autopager: Exploiting Change Blindness for Gaze-Assisted Reading. In *ETRA '18*. ACM, Article 46, 5 pages.

[248] Lior Wolf, Tal Hassner, and Itay Maoz. 2011. Face recognition in unconstrained videos with matched background similarity. In *CVPR '11*. IEEE, 529–534.

[249] Erroll Wood, Tadas Baltrusaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling. 2015. Rendering of eyes for eye-shape registration and gaze estimation. In *the IEEE International Conference on Computer Vision*. IEEE, 3756–3764.

[250] Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. 2016. Learning an Appearance-Based Gaze Estimator from One Million Synthesised Images. In *ETRA '16*. ACM, 131–138.

[251] Yong Wu, Gongyang Li, Zhi Liu, Mengke Huang, and Yang Wang. 2022. Gaze Estimation via Modulation-Based Adaptive Network With Auxiliary Self-Learning. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 8 (2022), 5510–5520.

[252] Zhengyang Wu, Srivignesh Rajendran, Tarrence Van As, Vijay Badrinarayanan, and Andrew Rabinovich. 2019. Eyenet: A multi-task deep network for off-axis eye gaze estimation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE.

[253] Yunyang Xiong, Hyunwoo J Kim, and Vikas Singh. 2019. Mixed effects neural networks (menets) with applications to gaze estimation. In *CVPR '19*. IEEE, 7743–7752.

[254] Yuki Yamato, Yutaro Suzuki, and Shin Takahashi. 2021. FGFlick: Augmenting Single-Finger Input Vocabulary for Smartphones with Simultaneous Finger and Gaze Flicks. In *Human-Computer Interaction – INTERACT 2021*, Carmelo Ardito, Rosa Lanzilotti, Alessio Malizia, Helen Petrie, Antonio Piccinno, Giuseppe Desolda, and Kori Inkpen (Eds.). Springer, 421–425.

[255] Dawei Yang, Xinlei Li, Xiaotian Dai, Rui Zhang, Lizhe Qi, Wenqiang Zhang, and Zhe Jiang. 2020. All in one network for driver attention monitoring. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2258–2262.

[256] Songzhou Yang, Yuan He, and Meng Jin. 2021. vGaze: Implicit Saliency-Aware Calibration for Continuous Gaze Tracking on Mobile Devices. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 1–10.

[257] Yu Yu and Jean-Marc Odobez. 2020. Unsupervised representation learning for gaze estimation. In *CVPR '20*. IEEE, 7314–7324.

[258] Jun-Seok Yun, Youngju Na, Hee Hyeon Kim, Hyung-Il Kim, and Seok Bong Yoo. 2022. HAZE-Net: High-Frequency Attentive Super-Resolved Gaze Estimation in Low-Resolution Face Images. In *the Asian Conference on Computer Vision*. IEEE, 3361–3378.

[259] Raimondas Zemblys, Diederick C Niehorster, and Kenneth Holmqvist. 2019. gazeNet: End-to-end eye-movement event detection with deep neural networks. *Behavior research methods* 51, 2 (2019), 840–864.

[260] Raimondas Zemblys, Diederick C Niehorster, Oleg Komogortsev, and Kenneth Holmqvist. 2018. Using machine learning to detect events in eye-tracking data. *Behavior research methods* 50, 1 (2018), 160–181.

[261] Cong Zhang, Qiyun He, Jiangchuan Liu, and Zhi Wang. 2017. Exploring viewer gazing patterns for touch-based mobile gamecasting. *IEEE Transactions on Multimedia* 19, 10 (2017), 2333–2344.

[262] Chi Zhang, Rui Yao, and Jinpeng Cai. 2018. Efficient eye typing with 9-direction gaze estimation. *Multimedia Tools and Applications* 77, 15 (2018).

[263] Guangtao Zhang, John Paulin Hansen, Katsumi Minakata, Alexandre Alapetite, and Zhongyu Wang. 2019. Eye-Gaze-Controlled Telepresence Robots for People with Motor Disabilities. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE.

[264] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.

[265] Xucong Zhang, Michael Xuelin Huang, Yusuke Sugano, and Andreas Bulling. 2018. *Training Person-Specific Gaze Estimators from User Interactions with Multiple Devices*. ACM, 1–12.

[266] Xiang Zhang, Kaori Ikematsu, Kunihiro Kato, and Yuta Sugiura. 2022. ReflecTouch: Detecting Grasp Posture of Smartphone Using Corneal Reflection Images. In *CHI '22*. ACM, Article 289, 8 pages.

[267] Xiaoyi Zhang, Harish Kulkarni, and Meredith Ringel Morris. 2017. Smartphone-Based Gaze Gesture Communication for People with Motor Disabilities. In *CHI '17*. ACM, 2878–2889.

[268] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. 2020. ETH-XGaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *European Conference on Computer Vision*. Springer, Springer, 365–381.

[269] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. 2019. Evaluation of Appearance-Based Methods and Implications for Gaze-Based Applications. In *CHI '19*. ACM, 13 pages.

[270] Xucong Zhang, Yusuke Sugano, Andreas Bulling, and Otmar Hilliges. 2020. Learning-based region selection for end-to-end gaze estimation. In *31st British Machine Vision Conference (BMVC 2020)*. British Machine Vision Association, BMVA, 86.

[271] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2015. Appearance-Based Gaze Estimation in the Wild. In *CVPR '15*. IEEE, 4511–4520.

[272] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2017. It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation. In *CVPR '17 Workshops*. IEEE, 2299–2308.

[273] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2019. MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 1 (2019), 162–175.

[274] Maozheng Zhao, Henry Huang, Zhi Li, Rui Liu, Wenzhe Cui, Kajal Toshniwal, Ananya Goel, Andrew Wang, Xia Zhao, Sina Rashidian, Furqan Baig, Khiem Phi, Shumin Zhai, IV Ramakrishnan, Fusheng Wang, and Xiaojun Bi. 2022. EyeSayCorrect: Eye Gaze and Voice Based Hands-Free Text Correction for Mobile Devices. In *IUI '22*. ACM, 470–482.

[275] Xuan Zhao, Mingming Fan, and Teng Han. 2022. "I Don't Want People to Look At Me Differently": Designing User-Defined Above-the-Neck Gestures for People with Upper Body Motor Impairments. In *CHI '22*. ACM, Article 1, 15 pages.

[276] Zeng Zhe, Felix Wilhelm Siebert, Antje Christine Venjakob, and Matthias Roetting. 2020. Calibration-free gaze interfaces based on linear smooth pursuit. *Journal of Eye Movement Research* 13, 1 (2020), 1–12.

[277] Xiaolong Zhou, Jianing Lin, Zhuo Zhang, Zhanpeng Shao, Shenyong Chen, and Honghai Liu. 2020. Improved itracker combined with bidirectional long short-term memory for 3D gaze estimation using appearance cues. *Neurocomputing* 390 (2020), 217–225.

[278] Wangjiang Zhu and Haoping Deng. 2017. Monocular free-head 3d gaze tracking with deep learning and geometry constraints. In *the IEEE International Conference on Computer Vision*. IEEE, 3143–3152.

[279] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z Li. 2017. Face alignment in full pose range: A 3d total solution. *IEEE transactions on pattern analysis and machine intelligence* 41, 1 (2017), 78–92.

[280] Ye Zhu, Yan Yan, and Oleg Komogortsev. 2020. Hierarchical HMM for eye movement classification. In *European Conference on Computer Vision*. Springer, Springer, 544–554.

Received May 2022; revised April 2023; accepted June 2023