

# MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation

Xucong Zhang, Yusuke Sugano\*, Mario Fritz, Andreas Bulling

**Abstract**—Learning-based methods are believed to work well for unconstrained gaze estimation, i.e. gaze estimation from a monocular RGB camera without assumptions regarding user, environment, or camera. However, current gaze datasets were collected under laboratory conditions and methods were not evaluated across multiple datasets. Our work makes three contributions towards addressing these limitations. First, we present the MPIIGaze dataset, which contains 213,659 full face images and corresponding ground-truth gaze positions collected from 15 users during everyday laptop use over several months. An experience sampling approach ensured continuous gaze and head poses and realistic variation in eye appearance and illumination. To facilitate cross-dataset evaluations, 37,667 images were manually annotated with eye corners, mouth corners, and pupil centres. Second, we present an extensive evaluation of state-of-the-art gaze estimation methods on three current datasets, including MPIIGaze. We study key challenges including target gaze range, illumination conditions, and facial appearance variation. We show that image resolution and the use of both eyes affect gaze estimation performance, while head pose and pupil centre information are less informative. Finally, we propose GazeNet, the first deep appearance-based gaze estimation method. GazeNet improves on the state of the art by 22% (from a mean error of 13.9 degrees to 10.8 degrees) for the most challenging cross-dataset evaluation.

**Index Terms**—Unconstrained Gaze Estimation, Cross-Dataset Evaluation, Convolutional Neural Network, Deep Learning

## 1 INTRODUCTION

GAZE estimation is well established as a research topic in computer vision because of its relevance for several applications, such as gaze-based human-computer interaction [1] or visual attention analysis [2], [3]. Most recent learning-based methods leverage large amounts of both real and synthetic training data [4], [5], [6], [7] for person-independent gaze estimation. They have thus brought us one step closer to the grand vision of *unconstrained gaze estimation*: 3D gaze estimation in everyday environments and without any assumptions regarding users' facial appearance, geometric properties of the environment and camera, or image formation properties of the camera itself. Unconstrained gaze estimation using monocular RGB cameras is particularly promising given the proliferation of such cameras in portable devices [8] and public displays [9].

While learning-based methods have demonstrated their potential for person-independent gaze estimation, methods have not been evaluated across different datasets to properly study their generalisation capabilities. In addition, current datasets have been collected under controlled laboratory conditions that are characterised by limited variability in appearance and illumination and the assumption of accurate head pose estimates. These limitations not only bear the risk of significant dataset bias – an important problem also identified in other areas in computer vision, such as object recognition [10] or salient object detection [11]. They also impede further progress towards unconstrained gaze estimation, given that it currently remains unclear how state-of-the-art methods perform on real-world images and across multiple datasets.

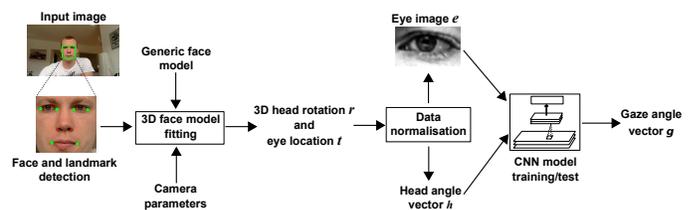


Fig. 1: Overview of GazeNet— appearance-based gaze estimation using a deep convolutional neural network (CNN).

This work aims to shed light on these questions and make the next step towards unconstrained gaze estimation. To facilitate cross-dataset evaluations, we first introduce the MPIIGaze dataset, which contains 213,659 images that we collected from 15 laptop users over several months in their daily life (see Fig. 2). To ensure frequent sampling during this time period, we opted for an experience sampling approach in which participants were regularly triggered to look at random on-screen positions on their laptop. This way, MPIIGaze not only offers an unprecedented realism in eye appearance and illumination variation but also in personal appearance – properties not available in any existing dataset. Methods for unconstrained gaze estimation have to handle significantly different 3D geometries between user, environment, and camera. To study the importance of such geometry information, we ground-truth annotated 37,667 images with six facial landmarks (eye and mouth corners) and pupil centres. These annotations make the dataset also interesting for closely related computer vision tasks, such as pupil detection. The full dataset including annotations is available at <https://www.mpi-inf.mpg.de/MPIIGaze>.

Second, we conducted an extensive evaluation of several state-of-the-art methods on three current datasets: MPIIGaze, EYE-DIAP [12], and UT Multiview [6]. We include a recent learning-

- X. Zhang, M. Fritz, and A. Bulling are with the Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken, Germany. E-mail: {xczhang, mfritz, bulling}@mpi-inf.mpg.de
- Y. Sugano is with the Graduate School of Information Science and Technology, Osaka University, Japan. E-mail: sugano@ist.osaka-u.ac.jp
- \*Work conducted while at the Max Planck Institute for Informatics

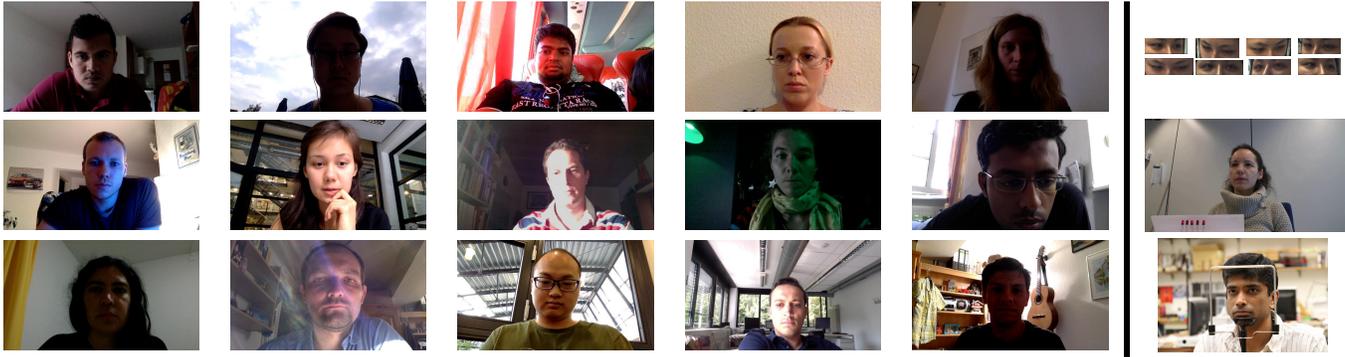


Fig. 2: Sample images from the MPIIGaze dataset showing the considerable variability in terms of place and time of recording, eye appearance, and illumination (particularly directional light and shadows). For comparison, the last column shows sample images from other current publicly available datasets (cf. Table 1): UT Multiview [6] (top), EYEDIAP [12] (middle), and Columbia [13] (bottom).

by-synthesis approach that trains the model with synthetic data and fine-tunes it on real data [14]. We first demonstrate the significant performance gap between previous within- and cross-dataset evaluation conditions. We then analyse various challenges associated with the unconstrained gaze estimation task, including gaze range, illumination conditions, and personal differences. Our experiments show these three factors are responsible for 25%, 35% and 40% performance gap respectively, when extending or restricting the coverage of training data. These analyses reveal that, although largely neglected in previous research, illumination conditions represent an important source of error, comparable to differences in personal appearance.

Finally, we propose GazeNet, the first deep appearance-based gaze estimation method based on a 16-layer VGG deep convolutional neural network. GazeNet outperforms the state of the art by 22% on MPIIGaze and 8% on EYEDIAP for the most difficult cross-dataset evaluation. Our evaluations represent the first account of the state of the art in cross-dataset gaze estimation and, as such, provide valuable insights for future research on this important but so far under-investigated computer vision task.

An earlier version of this work was published in [15]. Parts of the text and figures are reused from that paper. The specific changes implemented in this work are: 1) Extended annotation of 37,667 images with six facial landmarks (four eye corners and two mouth corners) and pupil centres, 2) updated network architecture to a 16-layer VGGNet, 3) new cross-dataset evaluation when training on synthetic data, 4) new evaluations of key challenges of the domain-independent gaze estimation task, specifically differences in gaze range, illumination conditions, and personal appearance, and 5) new evaluations on the influence of image resolution, the use of both eyes, and the use of head pose and pupil centre information on gaze estimation performance.

## 2 RELATED WORK

### 2.1 Gaze Estimation Methods

Gaze estimation methods can generally be distinguished as model-based or appearance-based [22]. Model-based methods use a geometric eye model and can be further divided into corneal-reflection and shape-based methods. Corneal-reflection methods rely on eye features detected using reflections of an external infrared light source on the outermost layer of the eye, the cornea. Early works on corneal reflection-based methods were limited to stationary settings [23], [24], [25], [26] but were later extended

to handle arbitrary head poses using multiple light sources or cameras [27], [28]. Shape-based methods [29], [30], [31], [32] infer gaze directions from the detected eye shape, such as the pupil centres or iris edges. Although model-based methods have recently been applied to more practical application scenarios [8], [33], [34], [35], [36], their gaze estimation accuracy is still lower, since they depend on accurate eye feature detections for which high-resolution images and homogeneous illumination are required. These requirements have largely prevented these methods from being widely used in real-world settings or on commodity devices.

In contrast, appearance-based gaze estimation methods do not rely on feature point detection but directly regress from eye images to 3D gaze directions. While early methods assumed a fixed head pose [37], [38], [39], [40], [41], [42], more recent methods allow for free 3D head movement in front of the camera [43], [44], [45], [46], [47]. Because they do not rely on any explicit shape extraction stage, appearance-based methods can handle low-resolution images and long distances. However, these methods require more person-specific training data than model-based approaches to cover the significant variability in eye appearance caused by free head motion and were therefore mainly evaluated for a specific domain or person. An open research challenge in gaze estimation is to learn gaze estimators that do not make any assumptions regarding the user, environment, or camera.

### 2.2 Person-Independent Gaze Estimation

The need to collect person-specific training data represents a fundamental limitation for both model-based and appearance-based methods. To reduce the burden on the user, several previous works used interaction events, such as mouse clicks or key presses, as a proxy for users' on-screen gaze position [48], [49]. Alternatively, visual saliency maps [50], [51] or pre-recorded human gaze patterns of the presented visual stimuli [52] were used as probabilistic training data to learn the gaze estimation function. However, the need to acquire user input fundamentally limits the applicability of these approaches to interactive settings.

Other methods aimed to learn gaze estimators that generalise to arbitrary persons without requiring additional input. A large body of works focused on cross-person evaluations in which the model is trained and tested on data from different groups of participants. For example, Schneider et al. performed a cross-person evaluation on the Columbia dataset [13] with 21 gaze points for one frontal head pose of 56 participants [5]. Funes et al. followed

	Participants	Head poses	Gaze targets	Illumination conditions	Face annotations	Amount of data	Collection duration	3D anno.
Villaneuva et al. [16]	103	1	12	1	1,236	1,236	1 day	No
TabletGaze [17]	51	continuous	35	1	none	1,428 min	1 day	No
GazeCapture [18]	1,474	continuous	continuous	daily life	none	2,445,504	1 day	No
Columbia [13]	56	5	21	1	none	5,880	1 day	Yes
McMurrough et al. [19]	20	1	16	1	none	97 min	1 day	Yes
Weidenbacher et al. [20]	20	19	2-9	1	2,220	2,220	1 day	Yes
OMEG [21]	50	3 + continuous	10	1	unknown	333 min	1 day	Yes
EYEDIAP [12]	16	continuous	continuous	2	none	237 min	2 days	Yes
UT Multiview [6]	50	8 + synthesised	160	1	64,000	64,000	1 day	Yes
<b>MPIIGaze (ours)</b>	<b>15</b>	<b>continuous</b>	<b>continuous</b>	<b>daily life</b>	<b>37,667</b>	<b>213,659</b>	<b>9 days ~ 3 months</b>	<b>Yes</b>

Table 1: Overview of publicly available appearance-based gaze estimation datasets showing the number of participants, head poses and on-screen gaze targets (discrete or continuous), illumination conditions, images with annotated face and facial landmarks, amount of data (number of images or duration of video), collection duration per participant, as well as the availability of 3D annotations of gaze directions and head poses. Datasets suitable for cross-dataset evaluation (i.e. that have 3D annotations) are listed below the double line.

a similar approach, but only evaluated on five participants [4]. To reduce data collection and annotation efforts, Sugano et al. presented a clustered random forest method that was trained on a large number of synthetic eye images [6]. The images were synthesised from a smaller number of real images captured using a multi-camera setup and controlled lighting in a laboratory. Later works evaluated person-independent gaze estimation methods on the same dataset [53], [54]. Krafka et al. recently presented a method for person-independent gaze estimation that achieved 1.71 cm on an iPhone and 2.53 cm on an iPad screen [18]. However, the method assumed a fixed camera-screen relationship and therefore cannot be used for cross-dataset gaze estimation.

### 2.3 Unconstrained Gaze Estimation

Despite significant advances in person-independent gaze estimation, all previous works assumed training and test data to come from the same dataset. We were first to study the practically more relevant but also significantly more challenging task of unconstrained gaze estimation via cross-dataset evaluation [15]. We introduced a method based on a multimodal deep convolutional neural network that outperformed all state-of-the-art methods by a large margin. More recently, we proposed another method that, in contrast to a long-standing line of work in computer vision, only takes the full face image as input, resulting again in significant performance improvements for both 2D and 3D gaze estimation [55]. In later works, Wood et al. demonstrated that large-scale methods for unconstrained gaze estimation could benefit from parallel advances in computer graphics techniques for eye region modelling. These models were used to synthesise large amounts of highly realistic eye region images, thereby significantly reducing both data collection and annotation efforts [14]. Their latest model is fully morphable [56] and can synthesise large numbers of images in a few hours on commodity hardware [7].

### 2.4 Datasets

Several gaze estimation datasets have been published in recent years (see Table 1 for an overview). Early datasets were severely limited with respect to variability in head poses, on-screen gaze

targets, illumination conditions, number of images, face and facial landmark annotations, collection duration per participant, and annotations of 3D gaze directions and head poses [13], [16], [19], [20]. More recent datasets are larger and cover the head pose and gaze ranges continuously. The OMEG dataset includes 200 image sequences from 50 people with fixed and free head movement but discrete visual targets [21]. TabletGaze includes 16 videos recorded from 51 people looking at different points on a tablet screen [17]. The EYEDIAP dataset contains 94 video sequences of 16 participants who looked at three different targets (discrete and continuous markers displayed on a monitor, and floating physical targets) under both static and free head motion conditions [12]. The UT Multiview dataset also contains dense gaze samples of 50 participants and 3D reconstructions of eye regions that can be used to synthesise images for arbitrary head poses and gaze targets [6]. However, all of these datasets were still recorded under controlled laboratory settings and therefore only include a few illumination conditions. While the recent GazeCapture dataset [18] includes a large number of participants, the limited number of images and similar illumination conditions per participant make it less interesting for unconstrained gaze estimation. Even more importantly, the lack of 3D annotations limits its use to within-dataset evaluations. Several large-scale datasets were published for visual saliency prediction, such as the crowd-sourced iSUN dataset [57], but their focus is on bottom-up saliency prediction, and input face or eye images are not available.

## 3 THE MPIIGAZE DATASET

To be able to evaluate methods for unconstrained gaze estimation, a dataset with varying illumination conditions, head poses, gaze directions, and personal appearance was needed. To fill this gap, we collected the MPIIGaze dataset that contains a large number of images from different participants, covering several months of their daily life (see Fig. 2 for sample images from our dataset). The long-term recording resulted in a dataset that is one order of magnitude larger and significantly more variable than existing datasets (cf. Table 1). All images in the dataset come with 3D annotations of gaze target and detected eye/head positions, which

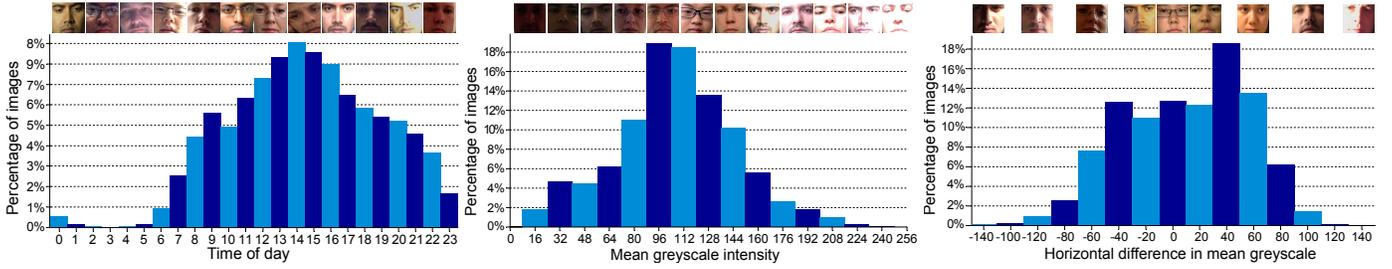


Fig. 3: Key characteristics of our dataset. Percentage of images collected at different times of day (left), having different mean grey-scale intensities within the face region (middle), and having horizontally different mean grey-scale intensities between the left to right half of the face region (right). Representative sample images are shown at the top.

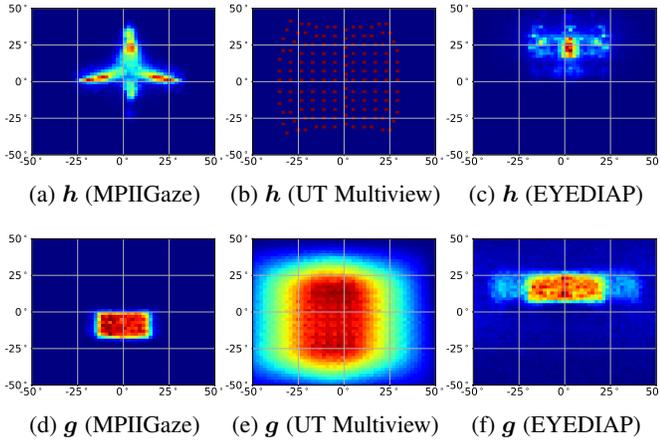


Fig. 4: Distributions of head angle ( $h$ ) and gaze angle ( $g$ ) in degrees for MPIIGaze, UT Multiview, and the screen target sequences in EYEDIAP (cf. Table 1).

is required for cross-dataset training and evaluation. Our dataset also provides manual facial landmark annotations on a subset of images, which enables a principled evaluation of gaze estimation performance and makes the dataset useful for other face-related tasks, such as eye or pupil detection.

### 3.1 Collection Procedure

We designed our data collection procedure with two main objectives in mind: 1) to record images of participants outside of controlled laboratory conditions, i.e. during their daily routine, and 2) to record participants over several months to cover a wider range of recording locations and times, illuminations, and eye appearances. We opted for recording images on laptops not only because they are suited for long-term daily recordings but also because they are an important platform for eye tracking applications [1]. Laptops are personal devices, therefore typically remaining with a single user, and they are used throughout the day and over long periods of time. Although head pose and gaze range are a bit limited compared to the fully unconstrained case due to the screen size, they have a strong advantage in that the data recording can be carried out in a mobile setup. They also come with high-resolution front-facing cameras and their large screen size allows us to cover a wide range of gaze directions. We further opted to use an experience sampling approach to ensure images were collected regularly throughout the data collection period [58].

We implemented custom software running as a background service on participants’ laptops, and opted to use the well-established moving dot stimulus [59], to collect ground-truth annotations. Every 10 minutes the software automatically asked participants to look at a random sequence of 20 on-screen positions (a recording session), visualised as a grey circle shrinking in size and with a white dot in the middle. Participants were asked to fixate on these dots and confirm each by pressing the spacebar exactly once when the circle was about to disappear. If they missed this small time window of about 500 ms, the software asked them to record the same on-screen location again right after the failure. While we cannot completely eliminate the possibility of bad ground truth, this approach ensured that participants had to concentrate and look carefully at each point during the recording.

Otherwise, participants were not constrained in any way, in particular as to how and where they should use their laptops. Because our dataset covers different laptop models with varying screen size and resolution, on-screen gaze positions were converted to 3D positions in the camera coordinate system. We obtained the intrinsic parameters from each camera beforehand using the camera calibration procedure from OpenCV [60]. The 3D position of the screen plane in the camera coordinate system was estimated using a mirror-based calibration method in which the calibration pattern was shown on the screen and reflected to the camera using a mirror [61]. Both calibrations are required for evaluating gaze estimation methods across different devices. 3D positions of the six facial landmarks were recorded from all participants using an external stereo camera prior to the data collection, which could be used to build the 3D face model.

### 3.2 Dataset Characteristics

We collected a total of 213,659 images from 15 participants (six female, five with glasses) aged between 21 and 35 years. 10 participants had brown, 4 green, and one grey iris colour. Participants collected the data over different time periods ranging from 9 days to 3 months. The number of images collected for each participant varied from 1,498 to 34,745. Note that we only included images in which a face could be detected (see Section 4.1). Fig. 3 (left) shows a histogram of times of the recording sessions. Although there is a natural bias towards working hours, the figure shows the high variation in recording times. Consequently, our dataset also covers significant variation in illumination. To visualise the different illumination conditions, Fig. 3 (bottom) shows a histogram of mean grey-scale intensities inside the face region. Fig. 3 (right) further shows a histogram of the mean intensity differences from

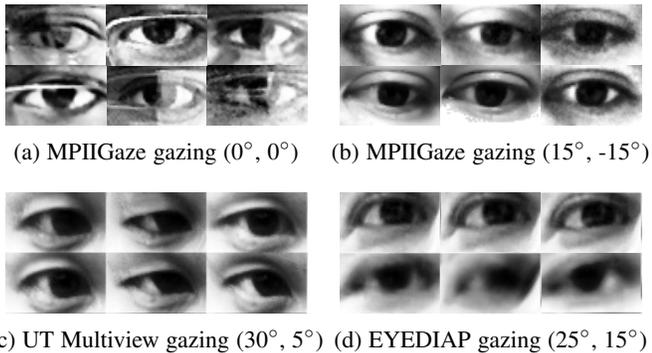


Fig. 5: Sample images from a single person for roughly the same gaze directions from MPIIGaze with (a) and without (b) glasses, UT Multiview (c), and EYEDIAP (d).

the right side to the left side of the face region, indicative of strong directional light for a substantial number of images.

The 2D histograms in Fig. 4 visualise the distributions of head and gaze angles  $h, g$  in the normalised space, colour-coded from blue (minimum) to red (maximum), for MPIIGaze in comparison with two other recent datasets, EYEDIAP (all screen target sequences) [12] and UT Multiview [6] (see Section 4.2 for a description of the normalisation procedure). The UT Multiview dataset (see Fig. 4b and 4e) was only recorded under a single controlled lighting condition, but provides good coverage of the gaze and pose spaces. For the EYEDIAP dataset, Fig. 4c and 4f show distributions of 2D screen targets that are comparable to our setting, yet gaze angle distributions do not overlap, due to different camera and gaze target plane setups (see Fig. 4a and 4d). For our MPIIGaze dataset, gaze directions tend to be below the horizontal axis in the camera coordinate system because the laptop-integrated cameras were positioned above the screen, and the recording setup biased the head pose to a near-frontal pose. The gaze angles in our dataset are in the range of  $[-1.5, 20]$  degrees in the vertical and  $[-18, +18]$  degrees in the horizontal direction.

Finally, Fig. 5 shows sample eye images from each dataset after normalisation. Each group of images was randomly selected from a single person for roughly the same gaze directions. Compared to the UT Multiview and EYEDIAP datasets (see Fig. 5c and 5d), MPIIGaze contains larger appearance variations even inside the eye region (see Fig. 5b), particularly for participants wearing glasses (see Fig. 5a).

### 3.3 Facial Landmark Annotation

We manually annotated a subset of images with facial landmarks to be able to evaluate the impact of face alignment errors on gaze estimation performance. To this end, we annotated the evaluation subset used in [15] that consists of a randomly-selected 1,500 left eye and 1,500 right eye images of all 15 participants. Because eye images could be selected from the same face, this subset contains a total of 37,667 face images.

The annotation was conducted in a semi-automatic manner. We first applied a state-of-the-art facial landmark detection method [62], yielding six facial landmarks per face image: the four eye and two mouth corners. We then showed these landmarks to two experienced human annotators and asked them to flag those images that contained incorrect landmark locations or wrong face detections (see Fig. 6b). 5,630 out of 37,667 images were flagged

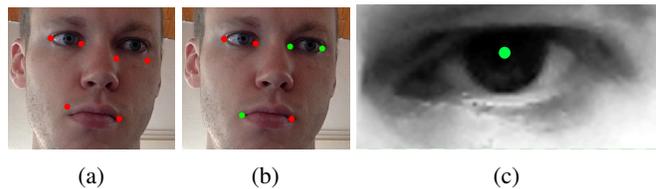


Fig. 6: We manually annotated 37,667 images with seven facial landmarks: the corners of the left and right eye, the mouth corners, and the pupil centres. We used a semi-automatic annotation approach: (a) Landmarks were first detected automatically (in red) and, (b) if needed, corrected manually post-hoc (in green). We also manually annotated the pupil centre without any detection (c). Note that this is only for completeness and we do not use the pupil centre as input for our method later.

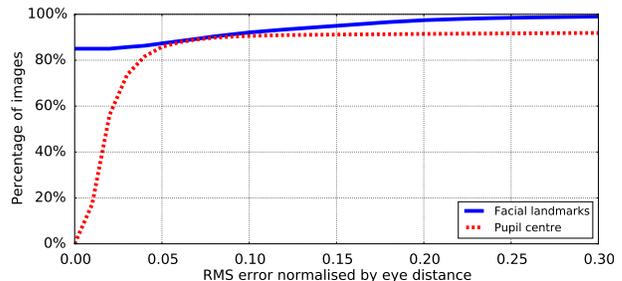


Fig. 7: Percentage of images for different error levels in the detection of facial landmarks (blue solid line) and pupil centres (red dashed line). The x-axis shows the root-mean-square (RMS) distance between the detected and annotated landmarks, normalised by the distance between both eyes.

for manual annotation in this process. Subsequently, landmark locations for all of these images were manually corrected by the same annotators. Since automatic pupil centre localisation remains challenging [63], we cropped the eye images using the manually-annotated facial landmarks and asked the annotators to annotate the pupil centres (see Fig. 6c).

Fig. 7 shows the detection error for facial landmarks and pupil centres when compared to the manual annotation. We calculated the error as the average root-mean-square (RMS) distances between the detected and annotated landmarks per face image. As can be seen from the figure, 85% of the images had no error in the detected facial landmarks. 0.98% of the images had normalised RMS error less than 0.3. This error roughly corresponds to the size of one eye and indicates that in these cases the face detection method failed to correctly detect the target face. For the pupil centre (red line), the error for each eye image is the RMS between the detected and annotated pupil centre normalised by the distance between both eyes. A normalised RMS error of 0.01 roughly corresponds to the size of the pupil, and 80% of the images had lower pupil detection performance.

## 4 METHOD

Prior work performed person-independent gaze estimation using 2D regression in the screen coordinate system [17], [18]. Because this requires a fixed position of the camera relative to the screen, these methods are limited to the specific device configuration, i.e. do not directly generalise to other devices. The recent success of deep learning combined with the availability of large-scale

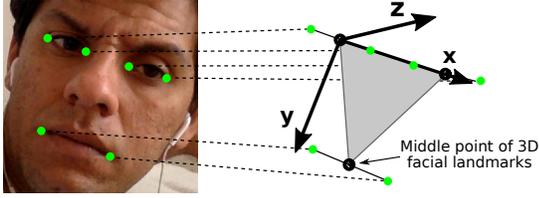


Fig. 8: Definition of the head coordinate system defined based on the triangle connecting three midpoints of the eyes and mouth. The x-axis goes through the midpoints of both while the y-axis is perpendicular to the x-axis inside the triangle plane. The z-axis is perpendicular to this triangle plane.

datasets, such as MPIIGaze, opens up promising new directions towards unconstrained gaze estimation that was not previously possible. In particular, large-scale methods promise to learn gaze estimators that can handle the significant variability in domain properties as well as user appearance. Fig. 1 shows an overview of our GazeNet method based on a multimodal convolutional neural network (CNN). We first use state-of-the-art face detection [64] and facial landmark detection [62] methods to locate landmarks in the input image obtained from the calibrated monocular RGB camera. We then fit a generic 3D facial shape model to estimate 3D poses of the detected faces and apply the space normalisation technique proposed in [6] to crop and warp the head pose and eye images to the normalised training space. A CNN is finally used to learn a mapping from the head poses and eye images to 3D gaze directions in the camera coordinate system.

#### 4.1 Face Alignment and 3D Head Pose Estimation

Our method first detects the user’s face in the image with a HOG-based method [64]. We assume a single face in the images and take the largest bounding box if the detector returned multiple face proposals. We discard all images in which the detector fails to find any face, which happened in about 5% of all cases. Afterwards, we use a continuous conditional neural fields (CCNF) model framework to detect facial landmarks [62].

While previous works assumed accurate head poses, we use a generic mean facial shape model  $\mathbf{F}$  for the 3D pose estimation to evaluate the whole gaze estimation pipeline in a practical setting. The generic mean facial shape  $\mathbf{F}$  is built as the averaged shape across all the participants, which could also be derived from any other 3D face models. We use the same definition of the face model and head coordinate system as [6]. The face model  $\mathbf{F}$  consists of 3D positions of six facial landmarks (eye and mouth corners, cf. Fig.1). As shown in Fig.8, the right-handed head coordinate system is defined according to the triangle connecting three midpoints of the eyes and mouth. The x-axis is defined as the line connecting midpoints of the two eyes in the direction from the right eye to the left eye, and the y-axis is defined to be perpendicular to the x-axis inside the triangle plane in the direction from the eye to the mouth. The z-axis is hence perpendicular to the triangle, and pointing backwards from the face. Obtaining the 3D rotation matrix  $\mathbf{R}_r$  and translation vector  $\mathbf{t}_r$  of the face model from the detected 2D facial landmarks  $\mathbf{p}$  is a classical *Perspective-n-Point*, problem which is estimating the 3D pose of an object given its 3D model and the corresponding 2D projections in the image. We fit  $\mathbf{F}$  to detected facial landmarks by estimating the initial solution using the EPnP algorithm [65] and further refine the pose by minimising the Levenberg-Marquardt distance.

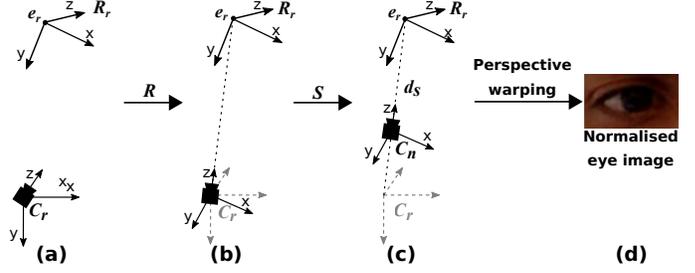


Fig. 9: Procedure for eye image normalisation. (a) Starting from the head pose coordinate system centred at one of the eye centres  $e_r$  (top) and the camera coordinate system (bottom); (b) the camera coordinate system is rotated with  $\mathbf{R}$ ; (c) the head pose coordinate system is scaled with matrix  $\mathbf{S}$ ; (d) the normalised eye image is cropped from the input image by the image transformation matrix corresponding to these rotations and scaling.

#### 4.2 Eye Image Normalisation

Given that our key interest is in cross-dataset evaluation, we normalise the image and head pose space as introduced in [6]. Fundamentally speaking, object pose has six degrees of freedom, and in the simplest case the gaze estimator has to handle eye appearance changes in this 6D space. However, if we assume that the eye region is planar, arbitrary scaling and rotation of the camera can be compensated for by its corresponding perspective image warping. Therefore, the appearance variation that needs to be handled inside the appearance-based estimation function has only two degrees of freedom. The task of pose-independent appearance-based gaze estimation is to learn the mapping between gaze directions and eye appearances, which cannot be compensated for by virtually rotating and scaling the camera.

The detailed procedure for the eye image normalisation is shown in Fig.9. Given the head rotation matrix  $\mathbf{R}_r$  and the eye position in the camera coordinate system  $e_r = \mathbf{t}_r + e_h$  where  $e_h$  is the position of the midpoint of the two eye corners defined in the head coordinate system (Fig. 9 (a)), we need to compute the conversion matrix  $\mathbf{M} = \mathbf{S}\mathbf{R}$  for normalisation. As illustrated in Fig. 9 (b),  $\mathbf{R}$  is the inverse of the rotation matrix that rotates the camera so that the camera looks at  $e_r$  (i.e., the eye position is located along the z-axis of the rotated camera), the x-axis of the head coordinate system is perpendicular to the y-axis of the camera coordinate system. The scaling matrix  $\mathbf{S} = \text{diag}(1, 1, d_n/\|e_r\|)$  (Fig. 9 (c)) is then defined so that the eye position  $e_r$  is located at a distance  $d_n$  from the origin of the scaled camera coordinate system.

$\mathbf{M}$  describes a 3D scaling and rotation that brings the eye centre to a fixed position in the (normalised) camera coordinate system, and is used for interconversion of 3D positions between the original and the normalised camera coordinate system. If we denote the original camera projection matrix obtained from camera calibration as  $\mathbf{C}_r$  and the normalised camera projection matrix as  $\mathbf{C}_n$ , the same conversion can be applied to the original image pixels via perspective warping using the image transformation matrix  $\mathbf{W} = \mathbf{C}_n\mathbf{M}\mathbf{C}_r^{-1}$  (Fig. 9 (d)).  $\mathbf{C}_n = [f_x, 0, c_x; 0, f_y, c_y; 0, 0, 1]$ , where  $f$  and  $c$  indicate the focal length and principal point of the normalised camera, which are arbitrary parameters of the normalised space. The whole normalisation process is applied to both right and left eyes in the same manner, with  $e_r$  defined according to the corresponding eye position.

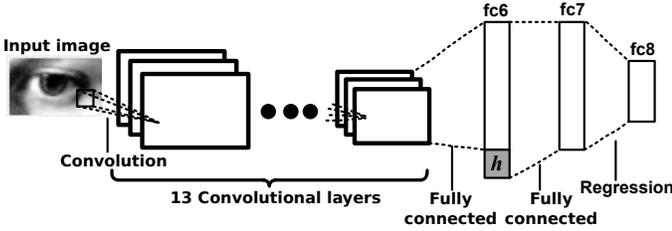


Fig. 10: Architecture of the proposed GazeNet. The head angle  $h$  is injected into the first fully connected layer. The 13 convolutional layers are inherited from a 16-layer VGG network [66].

This yields a set of an eye image  $I$ , a head rotation matrix  $R_n = MR_r$ , and a gaze angle vector  $g_n = Mg_r$  in the normalised space.  $g_r$  is the 3D gaze vector originating from  $e_r$  in the original camera coordinate system. The normalised head rotation matrix  $R_n$  is then converted to a three-dimensional rotation angle vector  $h_n$ . Since rotation around the z-axis is always zero after normalisation,  $h_n$  can be represented as a two-dimensional rotation vector (horizontal and vertical orientations)  $h$ .  $g_n$  is also represented as a two-dimensional rotation vector  $g$  assuming a unit length. We define  $d_n$  to be 600 mm and focal length  $f_x$  and  $f_y$  of the normalised camera projection matrix  $C_n$  to be 960, so that it is compatible with the UT Multiview dataset [6]. The resolution of the normalised eye image is set to  $I$  in  $60 \times 36$  pixels, and thus  $c_x$  and  $c_y$  are set to 30 and 18, respectively. Eye images  $I$  are converted to grey scale and histogram-equalised after normalisation to make the normalised eye images compatible between different datasets, facilitating cross-dataset evaluations.

### 4.3 GazeNet Architecture

The task for the CNN is to learn a mapping from the input features (2D head angle  $h$  and eye image  $e$ ) to gaze angles  $g$  in the normalised space. In the unconstrained setting, the distance to the target gaze plane can vary. The above formulation thus has the advantage that training data does not have to consider the angle of convergence between both eyes. As pointed out in [6], the difference between the left and right eyes is irrelevant in the person-independent evaluation scenario: By flipping eye images horizontally and mirroring  $h$  and  $g$  around the horizontal direction, both eyes can be handled using a single regression function.

Our method is based on the 16-layer VGGNet architecture [66] that includes 13 convolutional layers, two fully connected layers, and one classification layer with five max pooling layers in between. Following prior work on face [62], [67] and gaze [41], [68] analysis, we use a grey-scale single channel image as input with a resolution of  $60 \times 36$  pixels. We changed the stride of the first and second pooling layer from two to one to reflect the smaller input resolution. The output of the network is a 2D gaze angle vector  $\hat{g}$  consisting of two gaze angles, yaw  $\hat{g}_\phi$  and pitch  $\hat{g}_\theta$ . We extended the vanilla VGGNet architecture into a multimodal model to also take advantage of head pose information [69]. To this end we injected head pose information  $h$  into the first fully connected layer (fc6) (see Fig. 10). As a loss function we used the sum of the individual  $L_2$  losses measuring the distance between the predicted  $\hat{g}$  and true gaze angle vector  $g$ .

## 5 EXPERIMENTS

We first evaluated GazeNet for cross-dataset and cross-person evaluation. We then explored key challenges in unconstrained

gaze estimation including differences in gaze ranges, illumination conditions, and personal appearance. Finally, we studied other closely related topics, such as the influence of image resolution, the use of both eyes, and the use of head pose and pupil centre information on gaze estimation performance. GazeNet was implemented using the Caffe library [70]. We used the weights of the 16-layer VGGNet [66] pre-trained on ImageNet for all our evaluations, and fine-tuned the whole network in 15,000 iterations with a batch size of 256 on the training set. We used the Adam solver [71] with the two momentum values set to  $\beta_1 = 0.9$  and  $\beta_2 = 0.95$ . An initial learning rate of 0.00001 was used and multiplied by 0.1 after every 5,000 iterations.

### Baseline Methods

We further evaluated the following baseline methods:

- **MnistNet:** The four-layer (two convolutional and two fully connected layers) MnistNet architecture [72] has been used as the first CNN-based method for appearance-based gaze estimation [15]. We used the implementation provided by [70] and trained weights from scratch. The learning rate was set to be 0.1 and the loss was also changed to the Euclidean distance between estimated and ground-truth gaze directions.
- **Random Forests (RF):** Random forests were recently demonstrated to outperform existing methods for person-independent appearance-based gaze estimation [6]. We used the implementation provided by the authors, and the same parameters as in [6], and we resized input eye images to  $15 \times 9$  according to the implementation in [6], which has been optimised.
- **$k$ -Nearest Neighbours (kNN):** As shown in [6], a simple kNN regression estimator can perform well in scenarios that offer a large amount of dense training images. We used the same kNN implementation and also incorporated a training images clustering in head angle space.
- **Adaptive Linear Regression (ALR):** Because it was originally designed for a person-specific and sparse set of training images [41], ALR does not scale well to large datasets. We therefore used the same approximation as in [4], i.e. we selected five training persons for each test person with lowest interpolation weights. We further selected random subsets of images from the neighbours of the test image in head pose space. We used the same image resolution as for RF.
- **Support Vector Regression (SVR):** Schneider et al. used SVR with a polynomial kernel under a fixed head pose [5]. We used a linear SVR [73] for scalability given the large amount of training data. We also used a concatenated vector of HOG and LBP features ( $6 \times 4$  blocks,  $2 \times 2$  cells for HOG) as suggested in [5]. However, we did not use manifold alignment since it does not support pose-independent training.
- **Shape-based approach (EyeTab):** In addition to the appearance-based methods, we evaluated one state-of-the-art shape-based method that estimates gaze by fitting a limbus model (a fixed-diameter disc) to detected iris edges [8]. We used the implementation provided by the authors.

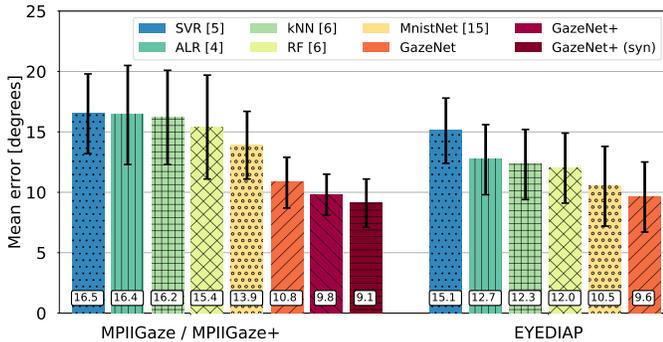


Fig. 11: Gaze estimation error for cross-dataset evaluation with training on 64,000 eye images in UT Multiview and testing on 45,000 eye images of MPIIGaze or MPIIGaze+ (left) and EYEDIAP (right). Bars show mean error across all participants; error bars indicate standard deviations.

## Datasets

As in [15], in all experiments that follow, we used a random subset of the full dataset consisting of 1,500 left eye images and 1,500 right eye images from each participant. Because one participant only offered 1,448 face images, we randomly oversampled data of that participant to 3,000 eye images. From now on we refer to this subset as *MPIIGaze*, while we call the same subset with manual facial landmark annotations *MPIIGaze+*. To evaluate the generalisation capabilities of the proposed method, in addition to MPIIGaze, we used all screen target sequences with both VGA and HD videos of the EYEDIAP dataset for testing [12]. We did not use the floating target sequences in the EYEDIAP dataset since they contain many extreme gaze directions that are not covered by UT Multiview. We further used the SynthesEyes dataset [14] that contains 11,382 eye samples from 10 virtual participants.

## Evaluation Procedure

For cross-dataset evaluation, each method was trained on UT Multiview or SynthesEyes, and tested on MPIIGaze, MPIIGaze+ or EYEDIAP. We used the UT Multiview dataset as the training set for each method because it covers the largest area in head and gaze angle spaces compared to EYEDIAP and our MPIIGaze datasets (see Fig.4). Note that SynthesEyes has the same head and gaze angle ranges as UT Multiview dataset. For cross-person evaluation, we performed a leave-one-person-out cross-validation for all participants on MPIIGaze+.

## 5.1 Performance Evaluation

We first report the performance evaluation for the cross-dataset setting, for which all the methods were trained and tested on two different datasets respectively, followed by the cross-person evaluation setting, for which all methods were evaluated with leave-one-person-out cross-validation.

### 5.1.1 Cross-Dataset Evaluation

Fig. 11 shows the mean angular errors of the different methods when trained on UT Multiview dataset and tested on both MPIIGaze, or MPIIGaze+, and EYEDIAP datasets. Bars correspond to mean error across all participants in each dataset, and error bars indicate standard deviations across persons. As can be seen from the figure, our GazeNet shows the lowest error on both datasets

(10.8 degrees on MPIIGaze, 9.6 degrees on EYEDIAP). This represents a significant performance gain of 22% (3.1 degrees) on MPIIGaze and 8% on EYEDIAP (0.9 degrees),  $p < 0.01$  using a paired Wilcoxon signed rank test [74], over the state-of-the-art method [15]. Performance on MPIIGaze and MPIIGaze+ is generally worse than on the EYEDIAP dataset, which indicates the fundamental difficulty of the in-the-wild setting covered by our dataset. We also evaluated performance on the different sequences of EYEDIAP (not shown in the figure). Our method achieved 10.0 degrees on the HD sequences and 9.2 degrees on the VGA sequences. This difference is most likely caused by differences in camera angles and image quality. The shape-based EyeTab method performs poorly on MPIIGaze (47.1 degrees mean error and 7% misdetection rate), which shows the advantage of appearance-based approaches in this challenging cross-dataset setting.

The input image size for some baselines, like RF, kNN and ALR, has been optimized to be  $15 \times 9$  pixels, which was lower than the  $60 \times 36$  pixels used in our method. To make the comparison complete, we also evaluated our GazeNet with  $15 \times 9$  pixels input images and achieved 11.4 degrees gaze estimation error on MPIIGaze, thereby still outperforming the other baseline methods.

Compared to GazeNet, GazeNet+ uses the manually annotated facial landmark locations MPIIGaze+ instead of the detected ones. In this case the mean error is reduced from 10.8 degrees to 9.8 degrees, which indicates that the face detection and landmark alignment accuracy is still a dominant error factor in practice. Furthermore, GazeNet+ (syn) implements the strategy proposed in [14]. That is, we first trained the model with synthetic data and then fine-tuned it on the UT Multiview dataset. This approach further reduced the gaze estimation error to 9.1 degrees. For comparison, the naive predictor that always outputs the average gaze direction of all training eye images in UT Multiview (not shown in the figure) achieves an estimation error of 34.2 degrees on MPIIGaze and 42.4 degrees on EYEDIAP.

While GazeNet achieved significant performance improvements for this challenging generalisation task, the results underline the difficulty of unconstrained gaze estimation. They also reveal a critical limitation of previous laboratory-based datasets such as UT Multiview with respect to variation in eye appearance, compared to MPIIGaze, which was collected in the real world. The learning-by-synthesis approach presented in [14] is promising given that it allows the synthesis of variable eye appearance and illumination conditions. This confirms the importance of the training data and indicates that future efforts should focus on addressing the gaze estimation task both in terms of training data as well as methodology to bridge the gap to the within-dataset scenario.

### 5.1.2 Cross-Person Evaluation

Although results of the previous cross-dataset evaluation showed the advantage of our GazeNet, they still fall short of the cross-person performance reported in [6]. To discuss the challenges of person-independent gaze estimation within MPIIGaze, we performed a cross-person evaluation using a leave-one-person-out approach. Fig. 12 shows the mean angular errors of this cross-person evaluation. Since the model-based EyeTab method has been shown to perform poorly in our setting, we opted to instead show a learning-based result using the detected pupil (iris centre) positions. More specifically, we used the pupil positions detected using [8] in the normalised eye image space as a feature for kNN regression, and performed the leave-one-person-out evaluation.

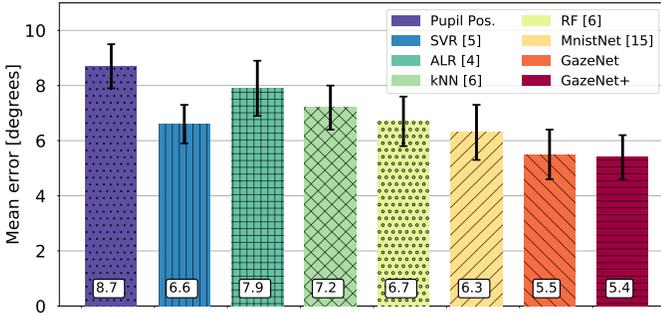


Fig. 12: Gaze estimation error on MPIIGaze and MPIIGaze+ for cross-person evaluation using a leave-one-person-out approach. Bars show the mean error across participants; error bars indicate standard deviations; numbers on the bottom are the mean estimation error in degrees. GazeNet+ refers to the result for MPIIGaze+.

As can be seen from the figure, all methods performed better than in the cross-dataset evaluation, which indicates the importance of domain-specific training data for appearance-based gaze estimation methods. Although the performance gain is smaller in this setting, our GazeNet still significantly (13%) outperformed the second-best MnistNet with 5.5 degrees mean error ( $p < 0.01$ , paired Wilcoxon signed rank test). While the pupil position-based approach worked better than the original EyeTab method, its performance was still worse than the different appearance-based methods. In this case there is dataset-specific prior knowledge about gaze distribution, and the mean prediction error that always outputs the average gaze direction of all training images becomes 13.9 degrees. Because the noise in facial landmark detections is included in the training set, there was no noticeable improvement when testing our GazeNet on MPIIGaze+ (shown as GazeNet+ in Fig. 12). It contradicts the observation with the previous cross-dataset evaluation that testing on MPIIGaze+ can bring one degree of improvement compared to MPIIGaze with detected facial landmarks (from 10.8 to 9.8 degrees).

## 5.2 Key Challenges

The previous results showed a performance gap between cross-dataset and cross-person evaluation settings. To better understand this gap, we additionally studied several key challenges. In all analyses that follow, we used GazeNet+ in combination with MPIIGaze+ to minimise error in face detection and alignment.

### 5.2.1 Differences in Gaze Ranges

As discussed in [15] and [14], one of the most important challenges for unconstrained gaze estimation is differences in gaze ranges between the training and testing domains. Although handling the different gaze angles has been researched by combining geometric and appearance-based methods [75], it is still challenging for appearance-based gaze estimation methods. The first bar in Fig. 13 (*UT*) corresponds to the cross-dataset evaluation using the *UT* Multiview dataset for training and MPIIGaze+ for testing. In this case, as illustrated in Fig. 4, the training data covers wider gaze ranges than the testing data. The second bar (*UT Sub*) corresponds to the performance of the model trained on a subset of the *UT* Multiview dataset that consists of 3,000 eye images per participant selected so as to have the same head pose and gaze angle distributions as MPIIGaze+. If the training dataset is tailored

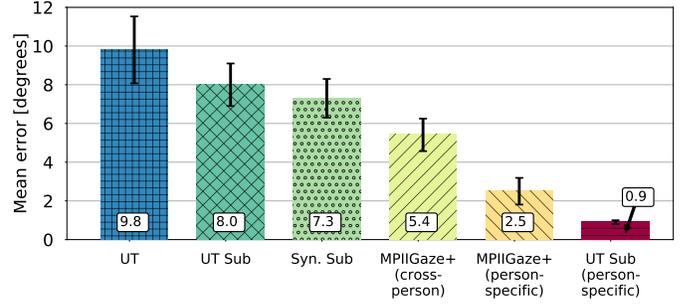


Fig. 13: Gaze estimation error on MPIIGaze+ using GazeNet+ for different training strategies and evaluation settings. Bars show the mean error across participants; error bars indicate standard deviations; numbers on the bottom are the mean estimation error in degrees. From left to right: 1) training on *UT* Multiview, 2) training on *UT* Multiview subset, 3) training on synthetic images targeted to the gaze and head pose ranges, 4) training on MPIIGaze+ with cross-person evaluation, 5) training on MPIIGaze+ with person-specific evaluation, and 6) training on *UT* Multiview subset with person-specific evaluation.

to the target domain and the specific gaze range, we achieve about 18% improvement in performance (from 9.8 to 8.0 degrees).

The top of Fig. 14 shows the gaze estimation errors in horizontal gaze direction with training on *UT* Multiview, *UT* Multiview subset, and MPIIGaze+, and testing on MPIIGaze+. The dots correspond to the average error for that particular gaze direction, while the line is the result of a quadratic polynomial curve fitting. The lines correspond to the *UT*, *UT Sub* and MPIIGaze+ (*cross-person*) bars in Fig. 13. As can be seen from the figure, for the model trained on *UT* Multiview subset, gaze estimation error increased for images that were close to the edge of the gaze range. In contrast, the model trained on the whole *UT* Multiview showed more robust performance across the full gaze direction range. The most likely reason for this difference is given by Fig. 14, which shows the percentage of images for the horizontal gaze directions for the training samples of MPIIGaze+ and *UT* Multiview. As can be seen from the figure, while *UT Sub* and MPIIGaze+ have the same gaze direction distribution, *UT* Multiview and MPIIGaze+ differ substantially. This finding demonstrates the fundamental shortcoming of previous works that only focused on cross-person evaluations and thereby implicitly or explicitly assumed a single, and thus restricted, gaze range. As such, this finding highlights the importance not only of cross-dataset evaluations, but also of developing methods that are robust to (potentially very) different gaze ranges found in different settings.

### 5.2.2 Differences in Illumination Conditions

Illumination conditions are another important factor in unconstrained gaze estimation and have been the main motivation for using fully synthetic training data that can cover a wider range of different illuminations [14]. The third bar in Fig. 13 (*Syn Sub*) corresponds to the same fine-tuned model as GazeNet+ (*syn*) in Fig. 11, but with the gaze range restricted to the same head pose and gaze angle distributions as MPIIGaze+. The fourth bar in Fig. 13 (MPIIGaze+ (*cross-person*)) shows the results of within-dataset cross-person evaluation on MPIIGaze+. For the second to the fourth bar in Fig. 13, the training data has nearly the same head angle and gaze direction range. The only difference is in

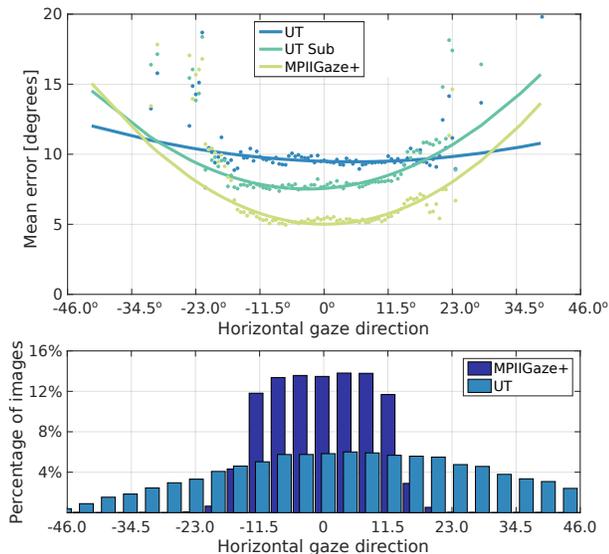


Fig. 14: Top: Gaze estimation error on MPIIGaze+ for the model trained with UT Multiview, UT Multiview subset, and MPIIGaze+ for different horizontal gaze directions. Bottom: Percentage of images for the horizontal gaze directions of MPIIGaze+ and UT.

the variation in illumination conditions in the training data. While the use of synthetic training data results in improved performance (from 8.0 degrees to 7.3 degrees), there is still a large gap between cross-dataset and cross-person settings.

This tendency is further illustrated in Fig. 15, in which we evaluated gaze estimation error with respect to lighting directions with our GazeNet. Similar to Fig. 3, we plotted the mean gaze estimation error according to the mean intensity difference between the left and right face region. The different colours represent the models trained with UT Multiview subset, synthetic subset and MPIIGaze+. They also correspond to *UT Sub*, *Syn. Sub* and *MPIIGaze+* (*cross-person*) in Fig. 13. The dots are averaged error for horizontal difference in the mean intensity in the face region, and lines are with quadratic polynomial curve fitting. Similar to Fig. 14, the bottom of Fig. 15 shows the percentage of images for mean greyscale intensity difference between the left and right half of the face region. We cannot show the distribution for *UT Sub* and *Syn. Sub* since their face images are not available. Compared to the model trained solely on the UT subset, the model with synthetic data shows better performance across different lighting conditions. While there still remains an overall performance gap from the domain-specific performance, the effect of synthetic data is more visible in the area with extreme lighting directions.

### 5.2.3 Differences in Personal Appearance

To further study the unconstrained gaze estimation task, we then evaluated person-specific gaze estimation performance, i.e. where training and testing data come from the same person. The results of this evaluation on MPIIGaze+ are shown as the second last bar (*MPIIGaze+* (*person-specific*)) in Fig. 13. Since there are 3,000 eye images for each participant in MPIIGaze+, we picked the first 2,500 eye images for training and the rest for testing. Similarly, the last bar (*UT Sub* (*p.s.*)) in Fig. 13 shows the person-specific evaluation within the UT subset, also with 2,500 eye images for training and 500 eye images for testing. The performance gap between *MPIIGaze+* (*cross-person*) and *MPIIGaze+*

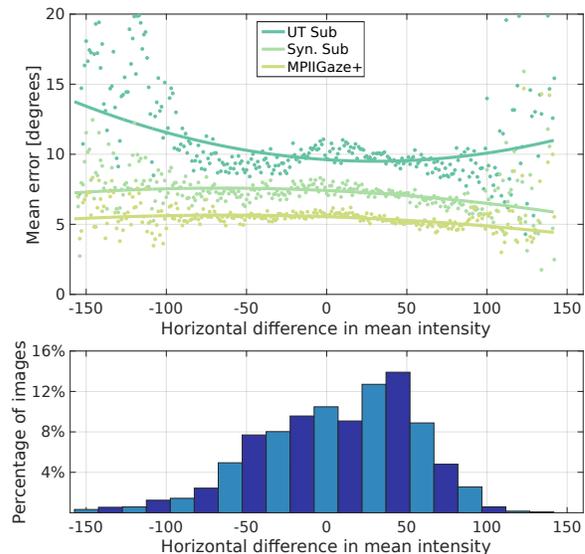


Fig. 15: Top: Gaze estimation error on MPIIGaze+ across mean greyscale intensity differences between the left and right half of the face region for models trained on UT subset, SynthesEyes subset, and MPIIGaze+. Bottom: Corresponding percentage of images for all mean greyscale intensity differences.

(*person-specific*) illustrates the fundamental difficulty of person-independent gaze estimation. The difference between *MPIIGaze+* (*person-specific*) and *UT Sub* (*p.s.*) also shows, however, that in-the-wild settings are challenging even for the person-specific case.

Fig. 16 shows the estimation error of each participant in both cross-dataset (trained on the UT Multiview) and person-specific (leave-one-person-out training on MPIIGaze+) settings with our GazeNet. Bars correspond to mean error for each participant and the error bars indicate standard deviations. Example faces from each participant are shown at the bottom. As the figure shows, for the cross-dataset evaluation the worst performance was achieved for participants wearing glasses (P5, P8, and P10). This is because the UT Multiview dataset does not include training images covering this case, although glasses can cause noise in the eye appearance as shown in Fig. 5a. For the person-specific evaluation, glasses are not the biggest error source, given that corresponding images are available in the training set. It can also be seen that the performance differences between participants are smaller in the person-specific evaluation. This indicates a clear need for developing new methods that can robustly handle differences in personal appearance for unconstrained gaze estimation.

## 5.3 Further Analyses

Following the previous evaluations of unconstrained gaze estimation performance and key challenges, we now provide further analyses on closely related topics, specifically the influence of image resolution, the use of both eyes, and the use of head pose and pupil centre information on gaze estimation performance.

### 5.3.1 Image Resolution

We first explored the influence of image resolution on gaze estimation performance, since it is conceivable that this represents a challenge for unconstrained gaze estimation. To this end, we evaluated the performance for the cross-dataset evaluation setting

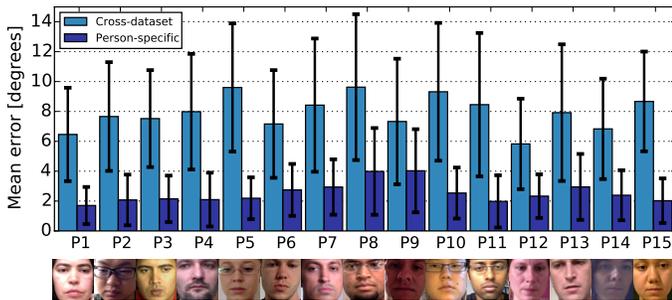


Fig. 16: Gaze estimation error for each participant for two evaluation schemes: *cross-dataset*, where the model was trained on UT Multiview and tested on MPIIGaze+, and *person-specific*, where the model was trained and tested on the same person from MPIIGaze+. Sample images are shown at the bottom.

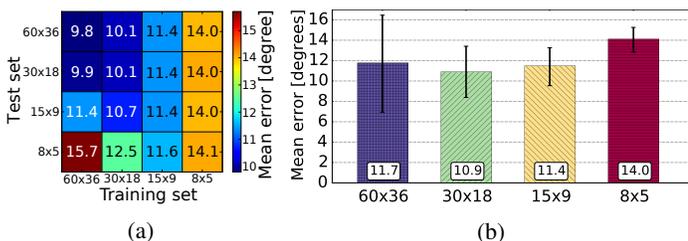


Fig. 17: Gaze estimation error of the models trained on UT Multiview and tested on MPIIGaze+ for different image resolutions. Test images were resized to the resolution of the training images. (a) Combinations of different training and test set resolutions with cell numbers indicating the average error in degrees. (b) The mean estimation error for the models trained with certain image resolutions across all images. Bars show the mean error across participants in degrees; error bars indicate standard deviations.

(trained on UT Multiview and tested on MPIIGaze+) for different training and testing resolutions with our GazeNet. Starting from the default input resolution  $60 \times 36$  in our model, we reduced the size to  $30 \times 18$ ,  $15 \times 9$  and  $8 \times 5$ . We always resized the test images according to the training resolution with bicubic interpolation. During training, we modified the stride of the first convolutional and max pooling layers of our GazeNet accordingly so that the input became the same starting from the second convolutional layer, regardless of the original image input resolution. Fig. 17a summarises the results of this evaluation with resolutions of training images along the x-axis, and resolutions of testing images on the y-axis. In general, if the test images have higher resolution than the training images, higher resolution results in better performance. Performance becomes significantly worse if the test images are smaller than the training images.

Fig. 17b shows the mean error of these models trained on one image resolution and tested across all testing resolutions, with the error bar denoting the standard deviation across all images. For the reason discussed above, the overall performance of the highest-resolution model is worse than that of the second  $30 \times 18$  model. This shows that higher resolution does not always mean better performance for unconstrained gaze estimation.

### 5.3.2 Use of Both Eyes

Previous methods typically used a single eye image as input. However, it is reasonable to assume that for some cases, such

as strong directional lighting, performance can be improved by using information from both eyes. To study this in more detail, we selected all images from MPIIGaze+ with two annotated eyes. We then evaluated different means of merging information from both eyes. The gaze estimation error when averaging across both eye images using the model trained on the UT Multiview dataset is 9.8 degrees with a standard deviation of 2.1 degrees. The best-case performance, i.e. always selecting the eye showing lower gaze estimation error, is 8.4 degrees with a standard deviation of 1.9 degrees. The gap between these two bars illustrates the limitations of the single eye-based estimation approach.

One approach to integrate estimation results from both eyes is to geometrically merge 3D gaze vectors after the appearance-based estimation pipeline. Given two 3D gaze vectors from both eyes, we thus further computed the mean gaze vector originating from the centre of both eyes. Ground-truth gaze vectors were also defined from the same origin, and the mean error across all faces using this approach was 7.2 degrees (standard deviation 1.4 degrees). It can be seen that even such a simple late fusion approach improves the estimation performance, indicating the potential of more sophisticated methods for fusing information from both eyes.

### 5.3.3 Use of Head Pose Information

To handle arbitrary head poses in the gaze estimation task, 3D head pose information has been used for the data normalisation as described in Sec. 4.2. After normalisation, 2D head angle vectors  $h$  were injected into the network as an additional geometry feature. The left side of Fig. 18 shows a comparison between different architectures of the multi-modal CNN on the UT Multiview dataset. We followed the same three-fold cross-validation setting as in [6]. The best performance reported in [6] is 6.5 degrees mean estimation error achieved by the head pose-clustered Random Forest. However, when the same clustering architecture is applied to the MnistNet (*Clustered MnistNet*), the performance became worse than for the model without clustering. In addition, our GazeNet (*Clustered GazeNet*) did not show any noticeable difference with the clustering structure. This indicates the higher learning flexibility of the CNN, which contributed to the large performance gain in the estimation task. The role of the additional head pose feature is also different in the two CNN architectures. While the MnistNet architecture achieved better performance with the help of the head pose feature, the effect of the head pose feature became marginal in the case of the GazeNet. Even though deeper networks like GazeNet can in general achieve better performance, achieving better performance with shallower networks is still important in some practical use cases where there is limited computational power, such as on mobile devices.

The right side of Fig. 18 shows a comparison of models with and without the head pose feature in the cross-dataset setting (trained on UT and tested on MPIIGaze+). The effect of the additional head pose feature is marginal in this case, but this is likely because the head pose variation in the MPIIGaze dataset is already limited to near-frontal cases. We performed an additional experiment to compare the gaze estimation performance when using the head pose estimated from the personal and the generic 3D face model. We achieved 9.8 degrees and 9.7 degrees for the cross-dataset evaluation, respectively, suggesting that the generic face model is sufficiently accurate for the gaze estimation task.

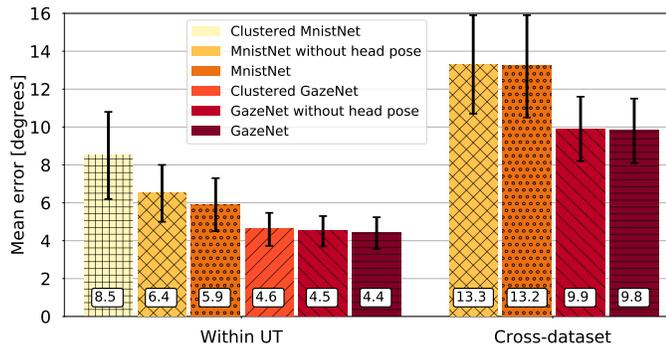


Fig. 18: Gaze estimation error when using the pose-clustered structure (*Clustered MnistNet* and *Clustered GazeNet*), without head angle vectors  $\mathbf{h}$  (*MnistNet without head pose* and *GazeNet without head pose*) for within-UT and cross-dataset (trained on UT, tested on MPIIGaze+) settings. Bars show the mean error across participants; error bars indicate standard deviations; numbers on the bottom are the mean estimation error in degrees.

### 5.3.4 Use of Pupil Centres

In GazeNet, we do not use pupil centre information as input. Although intuitively, eye shape features, such as pupil centres, can be a strong cue for gaze estimation, the model- or shape-based baseline performed relatively poorly for both the cross-dataset and cross-person evaluations. We therefore finally evaluated the performance of GazeNet when using the pupil centre as an additional feature for cross-person evaluation on MPIIGaze+. We detected the pupil centre location inside the normalised eye images using [8] and concatenated the pupil location to the geometry feature vector (head angle  $\mathbf{h}$ ). While there was an improvement between the models without and with the pupil centre feature, the improvement was relatively small (from 5.4 to 5.2 degrees). Performance improved more when using the manually annotated pupil centres, but still not significantly (5.0 degrees).

## 6 DISCUSSION

This work made an important step towards unconstrained gaze estimation, i.e. gaze estimation from a single monocular RGB camera without assumptions regarding users’ facial appearance, geometric properties of the environment or the camera and user therein. Unconstrained gaze estimation represents the practically most relevant but also most challenging gaze estimation task. Unconstrained gaze estimation is, for example, required for second-person gaze estimation from egocentric cameras or by a mobile robot. Through cross-dataset evaluation on our new MPIIGaze dataset, we demonstrated the fundamental difficulty of this task compared to the commonly used person-independent, yet still domain-specific, evaluation scheme. Specifically, gaze estimation performance dropped by up to 69% (from a gaze estimation error of 5.4 to 9.1 degrees) for the cross-dataset evaluation, as can be seen by comparing Figs. 11 and 12. The proposed GazeNet significantly outperformed the state of the art for both evaluation settings and in particular when pre-trained on synthetic data (see Fig. 11). The 3.1 degrees improvement that we achieved in the cross-dataset evaluation corresponds to around 2.9 cm on the laptop screen after backprojection. Performance on MPIIGaze was generally worse than on EYEDIAP, which highlights the difficulty but also the importance of developing and evaluating gaze estimators on images collected in real-world environments.

We further explored key challenges of this task, including differences in gaze ranges, illumination conditions, and personal appearance. Previous works either implicitly or explicitly side-stepped these challenges by restricting the gaze or head pose range [19], [76], studying a fixed illumination condition [6], [12], [21], or by only recording for short amounts of time and thereby limiting variations in personal appearance [13], [20]. Several recent works also did not study 3D gaze estimation but, instead, simplified the task to regression from eye images to 2D on-screen coordinates [17], [18]. While the 3D gaze estimation task generalises across hardware and geometric settings and thus facilities full comparison with other methods, the 2D task depends on the camera-screen relationship. Our evaluations demonstrated the fundamental shortcomings of such simplifications. They also showed that the development of 3D gaze estimation methods that properly handle all of these challenges, while important, remains largely unexplored. The ultimate goal of unconstrained gaze estimation is to obtain a generic estimator that can be distributed as a pre-trained library. While it is challenging to learn estimators that are robust and accurate across multiple domains, an intermediate solution might be to develop methods that adapt using domain-specific data automatically collected during deployment [2], [77].

The head angle vector plays different roles for the cross- and within-dataset evaluations. It is important to note that a 3D formulation is always required for unconstrained gaze estimation without restricting the focal length of the camera or the pose of the gaze target plane. 3D geometry, including the head pose, therefore has to be handled properly for unconstrained gaze estimation – a challenge still open at the moment. In this work we additionally explored the use of the head angle vector as a separate input to the CNN architecture as described in [15]. As shown in Fig. 18, while head pose information does result in a performance improvement for the shallower MnistNet architecture used in [15], it does not significantly improve the performance of GazeNet.

The state-of-the-art shape-based method [8] performed poorly in the cross-dataset evaluation, achieving only 47.1 degrees mean error. Similarly, adding the detected pupil centres as additional input to the CNN resulted in only a small performance improvement (see Section 5.3.4). While using eye shape and pupil centre features is typically considered to be a promising approach, both findings suggest that its usefulness may be limited for unconstrained gaze estimation, particularly on images collected in real-world settings – leaving aside the challenge of detecting these features robustly and accurately on such images in the first place.

## 7 CONCLUSION

In this work we made a case for unconstrained gaze estimation – a task that, despite its scientific and practical importance, has been simplified in several ways in prior work. To address some of these simplifications, we presented the new MPIIGaze dataset that we collected over several months in everyday life and that therefore covers significant variation in eye appearance and illumination. The dataset also offers manually annotated facial landmarks for a large subset of images and is therefore well-suited for cross-dataset evaluations. Through extensive evaluation of several state-of-the-art appearance- and model-based gaze estimation methods, we demonstrated both the critical need for and challenges of developing new methods for unconstrained gaze estimation. Finally, we proposed an appearance-based method based on a deep convolutional neural network that improves performance by 22%

for the most challenging cross-dataset evaluation on MPIIGaze. Taken together, our evaluations provide a detailed account of the state of the art in appearance-based gaze estimation and guide future research on this important computer vision task.

## ACKNOWLEDGMENTS

We would like to thank Laura Sesma for her help with the dataset handling and normalisation. This work was funded, in part, by the Cluster of Excellence on Multimodal Computing and Interaction (MMCI) at Saarland University, Germany, an Alexander von Humboldt Postdoctoral Fellowship, Germany, and a JST CREST Research Grant (JPMJCR14E1), Japan.

## REFERENCES

- [1] P. Majaranta and A. Bulling, "Eye tracking and eye-based human-computer interaction," in *Advances in Physiological Computing*, 2014, pp. 39–65.
- [2] Y. Sugano, X. Zhang, and A. Bulling, "Aggregaze: Collective estimation of audience attention on public displays," in *Proc. ACM Symp. User Interface Software and Technology*, 2016, pp. 821–831.
- [3] H. Sattar, S. Müller, M. Fritz, and A. Bulling, "Prediction of search targets from fixations in open-world settings," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 981–990.
- [4] K. A. Funes Mora and J.-M. Odobez, "Person independent 3d gaze estimation from remote rgb-d cameras," in *Proc. IEEE Int. Conf. on Image Processing*, 2013, pp. 2787–2791.
- [5] T. Schneider, B. Schauerer, and R. Stiefelhagen, "Manifold alignment for person independent appearance-based gaze estimation," in *Proc. IEEE Int. Conf. Pattern Recognition*, 2014, pp. 1167–1172.
- [6] Y. Sugano, Y. Matsushita, and Y. Sato, "Learning-by-synthesis for appearance-based 3d gaze estimation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 1821–1828.
- [7] E. Wood, T. Baltrušaitis, L.-P. Morency, P. Robinson, and A. Bulling, "Learning an appearance-based gaze estimator from one million synthesised images," in *Proc. ACM Symp. Eye Tracking Research Applications*, 2016, pp. 131–138.
- [8] E. Wood and A. Bulling, "Eyetable: Model-based gaze estimation on unmodified tablet computers," in *Proc. ACM Symp. on Eye Tracking Research and Applications*, 2014, pp. 207–210.
- [9] Y. Zhang, A. Bulling, and H. Gellersen, "Sideways: A gaze interface for spontaneous interaction with situated displays," in *Proc. ACM CHI Conf. on Human Factors in Computing Systems*, 2013, pp. 851–860.
- [10] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011, pp. 1521–1528.
- [11] Y. Li, X. Hou, C. Koch, J. Rehg, and A. Yuille, "The secrets of salient object segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 280–287.
- [12] K. A. Funes Mora, F. Monay, and J.-M. Odobez, "EYEDIAP: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras," in *Proc. ACM Symp. on Eye Tracking Research and Applications*, 2014, pp. 255–258.
- [13] B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar, "Gaze locking: passive eye contact detection for human-object interaction," in *Proc. ACM Symp. on User Interface Software and Technology*, 2013, pp. 271–280.
- [14] E. Wood, T. Baltrušaitis, X. Zhang, Y. Sugano, P. Robinson, and A. Bulling, "Rendering of eyes for eye-shape registration and gaze estimation," in *Proc. IEEE Int. Conf. Computer Vision*, 2015.
- [15] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 4511–4520.
- [16] A. Villanueva, V. Ponz, L. Sesma-Sanchez, M. Ariz, S. Porta, and R. Cabeza, "Hybrid method based on topography for robust detection of iris center and eye corners," *Trans. Multimedia Computing, Communications, and Applications*, vol. 9, no. 4, p. 25, 2013.
- [17] Q. Huang, A. Veeraraghavan, and A. Sabharwal, "Tabletgaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets," *Machine Vision and Applications*, vol. 28, no. 5, pp. 445–461, 2017.
- [18] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 2176–2184.
- [19] C. D. McMurrough, V. Metsis, J. Rich, and F. Makedon, "An eye tracking dataset for point of gaze detection," in *Proc. ACM Symp. on Eye Tracking Research and Applications*, 2012, pp. 305–308.
- [20] U. Weidenbacher, G. Layher, P.-M. Strauss, and H. Neumann, "A comprehensive head pose and gaze database," in *Proc. 3rd IET Int. Conf. Intelligent Environments*, 2007, pp. 455–458.
- [21] Q. He, X. Hong, X. Chai, J. Holappa, G. Zhao, X. Chen, and M. Pietikäinen, "Omeg: Oulu multi-pose eye gaze dataset," in *Image Analysis*, 2015, pp. 418–427.
- [22] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 478–500, 2010.
- [23] C. H. Morimoto, A. Amir, and M. Flickner, "Detecting eye position and gaze from a single camera and 2 light sources," in *Proc. IEEE Int. Conf. Pattern Recognition*, 2002, pp. 314–317.
- [24] S.-W. Shih and J. Liu, "A novel approach to 3-d gaze tracking using stereo cameras," *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 34, no. 1, pp. 234–245, 2004.
- [25] D. H. Yoo and M. J. Chung, "A novel non-intrusive eye gaze estimation using cross-ratio under large head motion," *Computer Vision and Image Understanding*, vol. 98, no. 1, pp. 25–51, 2005.
- [26] C. Hennessey, B. Noureddin, and P. Lawrence, "A single camera eye-gaze tracking system with free head motion," in *Proc. ACM Symp. on Eye Tracking Research and Applications*, 2006, pp. 87–94.
- [27] Z. Zhu and Q. Ji, "Eye gaze tracking under natural head movements," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005, pp. 918–923.
- [28] Z. Zhu, Q. Ji, and K. P. Bennett, "Nonlinear eye gaze mapping function estimation via support vector regression," in *Proc. IEEE Int. Conf. Pattern Recognition*, 2006, pp. 1132–1135.
- [29] T. Ishikawa, S. Baker, I. Matthews, and T. Kanade, "Passive driver gaze tracking with active appearance models," in *Proc. 11th World Congress on Intelligent Transportation Systems*, 2004.
- [30] J. Chen and Q. Ji, "3d gaze estimation with a single camera without ir illumination," in *Proc. IEEE Int. Conf. Pattern Recognition*, 2008, pp. 1–4.
- [31] H. Yamazoe, A. Utsumi, T. Yonezawa, and S. Abe, "Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions," in *Proc. ACM Symp. on Eye Tracking Research and Applications*, 2008, pp. 245–250.
- [32] R. Valenti, N. Sebe, and T. Gevers, "Combining head pose and eye location information for gaze estimation," *IEEE Trans. Image Processing*, vol. 21, no. 2, pp. 802–815, 2012.
- [33] L. Jianfeng and L. Shigang, "Eye-model-based gaze estimation by rgb-d camera," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops*, 2014, pp. 592–596.
- [34] K. A. Funes Mora and J.-M. Odobez, "Geometric generative gaze estimation (g3e) for remote rgb-d cameras," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 1773–1780.
- [35] L. Sun, M. Song, Z. Liu, and M.-T. Sun, "Real-time gaze estimation with online calibration," *IEEE MultiMedia*, vol. 21, no. 4, pp. 28–37, 2014.
- [36] S. Cristina and K. P. Camilleri, "Model-based head pose-free gaze estimation for assistive communication," *Computer Vision and Image Understanding*, vol. 149, pp. 157–170, 2016.
- [37] S. Baluja and D. Pomerleau, "Non-intrusive gaze tracking using artificial neural networks," DTIC Document, Tech. Rep., 1994.
- [38] K.-H. Tan, D. J. Kriegman, and N. Ahuja, "Appearance-based eye gaze estimation," in *Proc. IEEE Workshop Applications of Computer Vision*, 2002, pp. 191–195.
- [39] O. Williams, A. Blake, and R. Cipolla, "Sparse and semi-supervised visual mapping with the S<sup>3</sup>GP," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2006, pp. 230–237.
- [40] W. Sewell and O. Komogortsev, "Real-time eye gaze tracking with an unmodified commodity webcam employing a neural network," in *Ext. Abstr. ACM CHI Conf. on Human Factors in Computing Systems*, 2010, pp. 3739–3744.
- [41] F. Lu, Y. Sugano, T. Okabe, and Y. Sato, "Adaptive linear regression for appearance-based gaze estimation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 36, no. 10, pp. 2033–2046, 2014.
- [42] K. Liang, Y. Chahir, M. Molina, C. Tijus, and F. Jouen, "Appearance-based gaze tracking with spectral clustering and semi-supervised gaussian process regression," in *Proc. Conf. Eye Tracking South Africa*, 2013, pp. 17–23.
- [43] F. Lu, T. Okabe, Y. Sugano, and Y. Sato, "Learning gaze biases with head motion for head pose-free gaze estimation," *Image and Vision Computing*, vol. 32, no. 3, pp. 169–179, 2014.

- [44] F. Lu, Y. Sugano, T. Okabe, and Y. Sato, "Head pose-free appearance-based gaze sensing via eye image synthesis," in *Proc. IEEE Int. Conf. Pattern Recognition*, 2012, pp. 1008–1011.
- [45] K. A. Funes Mora and J.-M. Odobez, "Gaze estimation from multimodal kinect data," in *IEEE Conf. Computer Vision and Pattern Recognition Workshop*, 2012, pp. 25–30.
- [46] J. Choi, B. Ahn, J. Parl, and I. S. Kweon, "Appearance-based gaze estimation using kinect," in *Proc. IEEE Conf. Ubiquitous Robots and Ambient Intelligence*, 2013, pp. 260–261.
- [47] T. Gao, D. Harari, J. Tenenbaum, and S. Ullman, "When computer vision gazes at cognition," *arXiv preprint arXiv:1412.2672*, 2014.
- [48] Y. Sugano, Y. Matsushita, Y. Sato, and H. Koike, "An incremental learning method for unconstrained gaze estimation," in *Proc. European Conf. Computer Vision*, 2008, pp. 656–667.
- [49] M. X. Huang, T. C. Kwok, G. Ngai, H. V. Leong, and S. C. Chan, "Building a self-learning eye gaze model from user interaction data," in *Proc. Int. Conf. on Multimedia*, 2014, pp. 1017–1020.
- [50] J. Chen and Q. Ji, "Probabilistic gaze estimation without active personal calibration," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011, pp. 609–616.
- [51] Y. Sugano, Y. Matsushita, and Y. Sato, "Appearance-based gaze estimation using visual saliency," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, no. 2, pp. 329–341, 2013.
- [52] F. Alnajar, T. Gevers, R. Valenti, and S. Ghebreab, "Calibration-free gaze estimation using human gaze patterns," in *Proc. IEEE Int. Conf. Computer Vision*, 2013, pp. 137–144.
- [53] P. Yu, J. Zhou, and Y. Wu, "Learning reconstruction-based remote gaze estimation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 3447–3455.
- [54] L. A. Jeni and J. F. Cohn, "Person-independent 3d gaze estimation using face frontalization," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops*, 2016, pp. 87–95.
- [55] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Its written all over your face: Full-face appearance-based gaze estimation," in *IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.
- [56] E. Wood, T. Baltrušaitis, L.-P. Morency, P. Robinson, and A. Bulling, "A 3d morphable eye region model for gaze estimation," in *Proc. European Conf. Computer Vision*, 2016, pp. 3756–3764.
- [57] P. Xu, K. A. Ehinger, Y. Zhang, A. Finkelstein, S. R. Kulkarni, and J. Xiao, "Turkergaze: Crowdsourcing saliency with webcam based eye tracking," *arXiv preprint arXiv:1504.06755*, 2015.
- [58] R. Larson and M. Csikszentmihalyi, "The experience sampling method," *New Directions for Methodology of Social & Behavioral Science*, 1983.
- [59] M. Kassner, W. Patera, and A. Bulling, "Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction," in *Proc. Int. Conf. on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM, 2014, pp. 1151–1160.
- [60] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.
- [61] R. Rodrigues, J. a. Barreto, and U. Nunes, "Camera pose estimation using images of planar mirror reflections," in *Proc. European Conf. Computer Vision*, 2010, pp. 382–395.
- [62] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Continuous conditional neural fields for structured regression," in *Proc. European Conf. Computer Vision*, 2014, pp. 593–608.
- [63] M. Tonsen, X. Zhang, Y. Sugano, and A. Bulling, "Labelled pupils in the wild: a dataset for studying pupil detection in unconstrained environments," in *Proc. ACM Symp. Eye Tracking Research Applications*, 2016, pp. 139–142.
- [64] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.
- [65] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EPnP: An accurate o(n) solution to the PnP problem," *Int. Journal of Computer Vision*, vol. 81, no. 2, pp. 155–166, 2009.
- [66] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. Learning Representations*, 2015.
- [67] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, "Joint cascade face detection and alignment," in *Proc. European Conf. Computer Vision*. Springer, 2014, pp. 109–122.
- [68] Y. Sugano, Y. Matsushita, Y. Sato, and H. Koike, "Appearance-based gaze estimation with online calibration from mouse operations," *Trans. Human-Machine Systems*, vol. 45, no. 6, pp. 750–760, 2015.
- [69] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," in *Proc. Int. Conf. on Machine Learning*, 2011, pp. 689–696.
- [70] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. Int. Conf. Multimedia*, 2014, pp. 675–678.
- [71] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *The Int. Conf. on Learning Representations (ICLR)*, 2015.
- [72] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [73] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Lib-linear: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.
- [74] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, vol. 1, no. 6, pp. 80–83, 1945.
- [75] K. A. F. Mora and J.-M. Odobez, "Gaze estimation in the 3d space using rgb-d sensors-yowards head-pose and user invariance," *International Journal of Computer Vision*, vol. 118, no. 2, pp. 194–216, 2016.
- [76] V. Ponz, A. Villanueva, and R. Cabeza, "Dataset for the evaluation of eye detector for gaze estimation," in *Proc. ACM Conf. Ubiquitous Computing*, 2012, pp. 681–684.
- [77] X. Zhang, Y. Sugano, and A. Bulling, "Everyday eye contact detection using unsupervised gaze target discovery," in *Proc. ACM Symp. on User Interface Software and Technology*, 2017, pp. 193–203.



**Xucong Zhang** is a PhD student in the Perceptual User Interfaces Group at the Max Planck Institute for Informatics, Germany. Xucong Zhang received his BSc. from the China Agriculture University in 2007, and a MSc. from Beihang University in 2010. His research interests include computer vision, human-computer interaction, and learning-based gaze estimation.



**Yusuke Sugano** is Associate Professor at the Graduate School of Information Science and Technology, Osaka University, Japan. Yusuke Sugano received his BSc., MSc., and PhD degrees from the University of Tokyo, in 2005, 2007, and 2010, respectively. He was previously a Postdoc in the Perceptual User Interfaces Group at the Max Planck Institute for Informatics, and a project research associate at Institute of Industrial Science, the University of Tokyo. His research interests include computer vision and

human-computer interaction.



**Mario Fritz** is Senior Researcher at the Max Planck Institute for Informatics, Germany, where he heads the Scalable Learning and Perception Group. Mario Fritz received his MSc. in Computer Science from the University of Erlangen-Nuremberg, Germany, in 2004 and his PhD from the Technical University of Darmstadt, Germany, in 2008. From 2008 to 2011, he was a Postdoc at the International Computer Science Institute (ICS) and UC Berkeley, US funded by a Feodor Lynen Fellowship from the Alexander von Humboldt Foundation. His research interests include computer vision, machine learning, privacy, and natural language processing.



**Andreas Bulling** is head of the Perceptual User Interfaces Group at the Max Planck Institute for Informatics and the Cluster of Excellence on Multimodal Computing and Interaction at Saarland University, Germany. Andreas Bulling received his MSc. in Computer Science from the Karlsruhe Institute of Technology, Germany, in 2006 and his PhD in Information Technology and Electrical Engineering from ETH Zurich, Switzerland, in 2010. From 2010 to 2013, he was a Feodor Lynen and Marie Curie Research Fellow at the University of Cambridge, UK. His research interests include computer vision, ubiquitous computing, and human-computer interaction.