InvisibleEye: Mobile Eye Tracking Using Multiple Low-Resolution Cameras and Learning-Based Gaze Estimation

MARC TONSEN, Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
JULIAN STEIL, Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
YUSUKE SUGANO*, Osaka University, Japan
ANDREAS BULLING, Max Planck Institute for Informatics, Saarland Informatics Campus, Germany

Analysis of everyday human gaze behaviour has significant potential for ubiquitous computing, as evidenced by a large body of work in gaze-based human-computer interaction, attentive user interfaces, and eye-based user modelling. However, current mobile eye trackers are still obtrusive, which not only makes them uncomfortable to wear and socially unacceptable in daily life, but also prevents them from being widely adopted in the social and behavioural sciences. To address these challenges we present *InvisibleEye*, a novel approach for mobile eye tracking that uses millimetre-size RGB cameras that can be fully embedded into normal glasses frames. To compensate for the cameras' low image resolution of only a few pixels, our approach uses multiple cameras to capture different views of the eye, as well as learning-based gaze estimation to directly regress from eye images to gaze directions. We prototypically implement our system and characterise its performance on three large-scale, increasingly realistic, and thus challenging datasets: 1) eye images synthesised using a recent computer graphics eye region model, 2) real eye images recorded of 17 participants under controlled lighting, and 3) eye images recorded of four participants over the course of four recording sessions in a mobile setting. We show that *InvisibleEye* achieves a top person-specific gaze estimation accuracy of 1.79° using four cameras with a resolution of only 5×5 pixels. Our evaluations not only demonstrate the feasibility of this novel approach but, more importantly, underline its significant potential for finally realising the vision of invisible mobile eye tracking and pervasive attentive user interfaces.

CCS Concepts: • **Human-centered computing** → **Interaction devices**; • **Computing methodologies** → *Computer vision*; *Machine learning approaches*;

Additional Key Words and Phrases: Mobile Eye Tracking; Appearance-Based Gaze Estimation;

ACM Reference format:
Marc Tonsen, Julian Steil, Yusuke Sugano, and Andreas Bulling. 2017. InvisibleEye: Mobile Eye Tracking Using Multiple Low-Resolution Cameras and Learning-Based Gaze Estimation. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 106 (September 2017), 21 pages.
DOI: 10.1145/3130971

[*] Work mainly conducted while at the Max Planck Institute for Informatics

Authors' addresses: M. Tonsen, J. Steil and A. Bulling, Max Planck Institute for Informatics, Saarland Informatics Campus, Campus E1 4, 66123 Saarbrücken, Germany. email: {tonsen,jsteil,bulling}@mpi-inf.mpg.de; Y. Sugano, Graduate School of Information Science and Technology, Osaka University, 1-5 Yamadaoka, Suita-shi, 565-0871 Osaka, Japan. email: sugano@ist.osaka-u.ac.jp.

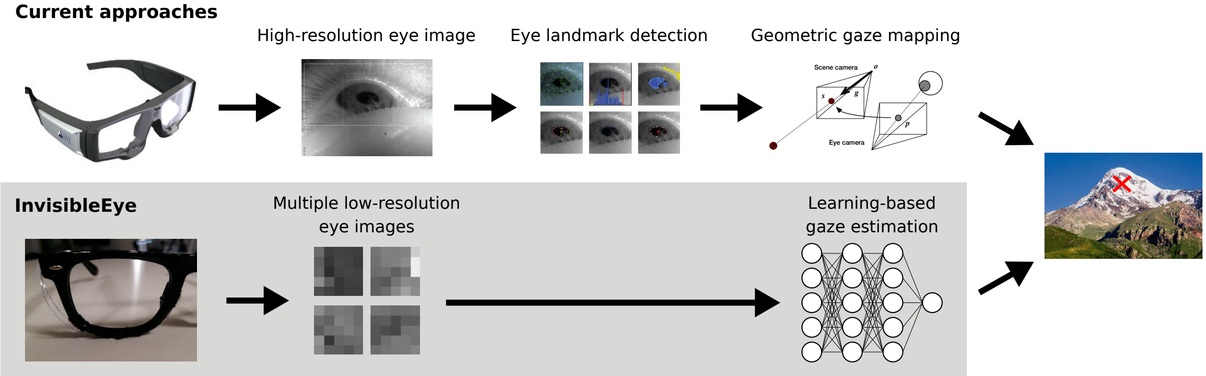


Fig. 1. (top) Classic approaches require high-resolution imaging sensors, resulting in rather bulky and obtrusive headsets, as well as hand-optimised algorithms for eye landmark detection and geometric gaze mapping. (bottom) *InvisibleEye* is a novel approach for mobile eye tracking that uses millimetre-size RGB cameras that can be fully embedded into normal glasses frames. To compensate for the cameras' low image resolution of only a few pixels, our approach uses multiple cameras in parallel and learning-based gaze estimation to regress to gaze position in the scene camera coordinate system (red cross).

1 INTRODUCTION

Human gaze has a long history as a means for hands-free interaction with ubiquitous computing systems and has, more recently, also been shown to be a rich source of information about the user [13, 18, 36]. Prior work has demonstrated that gaze can be used for fast, accurate, and natural interaction with both ambient [31, 53, 63, 65, 73] and body-worn displays, including smartwatches [3, 21]. Eye movements are closely linked to everyday human behaviour and cognition and can therefore be used for computational user modelling, such as for eye-based recognition of daily activities [14, 15], visual memory recall [10], visual search targets [49, 50, 70], and intents [7], or personality traits [25] – including analyses over long periods of time for life-logging applications [17, 52]. Interest in gaze has been fuelled by recent technical advances and significant reductions in the cost of mobile eye trackers that can be worn in daily life and thus provide access to users' everyday gaze behaviour [9].

However, despite its appeal, mobile eye tracking still suffers from several fundamental usability problems. First, current mobile trackers are still rather uncomfortable to wear, especially during long-term recordings. The main reason for this is high-quality imaging sensors that are large and thus often occlude the user's field of view. In addition, the sensors themselves as well as the additional electronics and wiring required to operate them makes current headsets heavy and cause discomfort or even pain. Second, current mobile eye trackers limit users' mobility given that they require a wired connection to a recording computer both as a power supply and for real-time image processing (often in the form of a laptop worn in a backpack). Eye trackers that do not require a wired connection instead store data on the device itself but, on the downside, are not well-suited for real-time applications. In addition, tetherless headsets require a battery, which adds to their weight and further limits their recording time. Finally, the obtrusive design of current eye trackers leads to low social acceptance and unnatural behaviour of both the wearer and people they interact with [41, 45], thus fundamentally limiting the practical usefulness of mobile eye tracking as a tool in the social and behavioural sciences.

To address these issues, we argue that it is ultimately necessary to fully integrate eye tracking into regular glasses, i.e. to effectively make eye tracking visually and physically unnoticeable to both the wearer and others. We believe that a key requirement for such unnoticeable (*invisible*) integration is to reduce the size of an eye tracker's core component: the imaging sensors. Smaller sensors would not only significantly reduce the device's weight but could also be positioned in the visual periphery to avoid occlusions within the users' field of view. In addition, the low resolution common to these sensors generates significantly less data that could more easily be processed on the device itself, stored, or transmitted wirelessly, thus removing the need for a separate recording device altogether. Finally, the reduced computation required to process low-resolution images decreases the load on the processor, which in turn could help to extend the recording time, which is limited to a few hours for current mobile eye trackers.

As a first step towards realising the above vision, we present *InvisibleEye*, a novel mobile eye tracker that uses millimetre-size imaging sensors with a resolution of only a few pixels that can be fully embedded into a normal glasses frame (see Figure 1). Traditional image processing and computer vision methods for eye landmark detection (most importantly pupil and pupil centre) and gaze estimation in mobile eye trackers require high-quality eye region images and are thus not suitable for such low-resolution sensors. Inspired by recent advances in remote gaze estimation in computer vision [71, 72], we instead propose a learning-based approach that does not require robust detection of eye landmarks but directly regresses from low-resolution eye images to 3D gaze directions. To compensate for the low resolution of each individual imaging sensor, and thus to improve overall gaze estimation accuracy, *InvisibleEye* uses multiple sensors positioned around the eye in parallel. In this work we learn a person-specific model for each user using training data recorded beforehand. Calibration-free (person-independent) gaze estimation is an open research challenge and an important direction for future work. We evaluate *InvisibleEye* on three large-scale, increasingly realistic datasets: 1) 200,000 eye images synthesised using a recent computer graphics method [69], which allows us to explore the influence of the number of cameras, camera positioning, and image resolution on gaze estimation performance in a principled way, 2) 280,000 real eye images recorded with a first prototype implementation in a laboratory setting with controlled lighting during a calibration-like procedure, and 3) 240,000 real eye images recorded using a second prototype over the course of four recording sessions in a mobile setting in which four participants gazed at a physical targets from various angles. The second dataset will be made publicly available upon acceptance[1]. We demonstrate that our approach can achieve a person-specific gaze estimation accuracy of 1.79° in the mobile setting using four cameras with an image resolution of only 5×5 pixels.

The specific contributions of this work are three-fold: First, we propose a novel approach for mobile eye tracking that leverages multiple tiny, low-resolution cameras that can be fully and thus invisibly integrated into a normal glasses frame. Second, we introduce a first-of-its-kind dataset of 280,000 close-up eye images that have been captured from multiple views and that are annotated with corresponding ground-truth gaze directions in both a stationary controlled and mobile setting. Third, we present extensive evaluations of two prototypical implementations of our approach on these datasets plus synthetic data and characterise their performance across key design parameters including image resolution, number of cameras, and camera angle and positioning.

2 RELATED WORK

Our work is related to previous works on 1) mobile eye tracking, 2) gaze estimation using multiple cameras, and 3) datasets for the development and evaluation of gaze estimation algorithms.

[1] The dataset is available at *http://www.mpi-inf.mpg.de/invisibleeye*

2.1 Mobile Eye Tracking

Many approaches for mobile eye tracking have been explored in the past, including some at low cost [28, 29, 42, 48]. The traditional computational pipeline for mobile gaze estimation involves 1) eye landmark detection, in particular detecting the pupil center, and ellipse fitting either using special-purpose image processing techniques [22, 24, 27, 32–34, 57] or machine learning [23], and 2) gaze mapping, traditionally using a geometric eye model [43, 44, 58, 62] or, more recently, by directly mapping 2D pupil positions to 3D gaze directions [38]. Instead of using two cameras, Nakazawa and Nitschke relied on only an eye camera and proposed a geometric approach to estimate gaze using corneal imaging [40]. All of these video-based methods rely on high-quality eye images and cameras, and therefore all suffer from the disadvantages discussed in the introduction.

Although a large body of works investigated learning-based gaze estimation, they mostly focused on remote settings, i.e. settings in which the camera is placed in front of the user, for example under a display [30, 35, 55, 71]. More closely related to ours is the work by Mayberry et al., who used a subset of pixels from an eye image to estimate gaze direction with an accuracy of up to 3° [39]. However, they still assumed high-resolution eye images as input, and did not fully explore the potential of the learning-based approach, in particular in terms of input image resolution. Although Abdulin et al. investigated the impact of image resolution of an eye camera and found that the iris-diameter resolution should be at least 50 pixels for model-based approaches [2], the minimum image resolution for learning-based approaches has not been fully investigated in prior work. In contrast, our work is first to utilise multiple low-resolution eye cameras that can be fully embedded into an ordinary glasses frame in combination with a learning-based gaze estimation method.

In an attempt to further integrate mobile eye tracking, a smaller number of works investigated alternative measurement techniques, such as electrooculography (EOG). EOG involves attaching electrodes on the skin around the eyes to measure the electric potential differences caused by eye movements. While EOG is computationally light-weight compared to video-based approaches, and thus promises full and low-power integration [11, 12, 37], due to drift and a low signal-to-noise ratio EOG is only suited for measuring relative movement of the eye. Borsato et al. instead used the sensor of a computer mouse to track the episcleral surface of the eye (the white part of the eye) using optic flow [8]. Using this approach they reported an accuracy of 2.1° of error at a 1 kHz sampling rate. However, the tracking was lost during every blink and the system had to be recalibrated each time, rendering it impractical for actual use. A few other works explored the use of phototransistors for mobile eye tracking that can, potentially, be fully integrated into a glasses frame. For example, Ishiguro et al. used infrared illumination in combination with four infrared sensitive phototransistors attached to a glasses frame to record relative movement of the eyes [26]. Their use of phototransistors allowed for a fairly compact, occlusion-free, and low-power design but the proposed system was only evaluated in a usability study without a quantitative analysis. With the goal of obtaining actual gaze estimates, Topal et al. used up to six infrared sensitive phototransistors per eye and trained a support vector machine to regress the gaze point from the signals achieving an average angular error of about 0.93° [61]. However, their evaluation was also limited to a constrained laboratory setting.

2.2 Multi-Camera Gaze Estimation

Several previous works investigated the use of multiple cameras for head pose estimation as a proxy to gaze, or gaze estimation directly. For example, Voit and Stiefelhagen equipped a room with multiple cameras to track horizontal head orientation of multiple users and, eventually, estimate who was looking at whom [66]. As a follow up work of [46], Ruddarraju et al. presented a method for detecting gaze in interaction [47]. Head pose was used to estimate a user's eye gaze and to measure if a user was looking at a previously defined region of interest. Utsumi et al. estimated users' head pose to choose the best out of multiple remote cameras positioned around the user to estimate gaze [64]. Arar et al. proposed a general framework for gaze estimation using multiple cameras placed around a computer screen by computing a weighted average of the estimations of each individual camera [4]. While all of these works explored multi-camera gaze estimation in remote settings, also using learning-based methods, our work is first to explore this approach for mobile eye tracking.

2.3 Gaze Estimation Datasets

In computer vision, but increasingly also in other fields, the availability of large-scale, annotated datasets to develop and evaluate learning-based methods has emerged as a critical requirement. Consequently, recent years have seen an increasing number of datasets being published, including for mobile gaze estimation. Swirski et al. presented a small dataset of 600 eye images recorded with a head-mounted camera, but the dataset only covered a single camera view and offered no variability in terms of participants or lighting conditions [57]. Tonsen et al. and Fuhl et al. provided large and challenging datasets with a lot of variability in personal appearance and illumination conditions but they, too, only included single-view recordings of one eye [22, 60]. While an ever-increasing number of datasets have been proposed, all of them target the tasks of pupil detection and ellipse fitting. To the best of our knowledge, none of the existing datasets offers ground truth gaze directions in addition to the eye images, thus limiting their use for developing and evaluating mobile gaze estimation pipelines. In contrast, we present the first-of-its-kind large-scale dataset of eye images that have been captured from multiple views and that are annotated with corresponding ground-truth gaze directions in both a stationary controlled and mobile everyday settings.

With the goal of reducing the time and effort required to record and annotate gaze estimation datasets, a relatively new line of work is exploring means to instead render highly realistic and perfectly annotated eye images using computer graphics techniques. Two representatives of this line of work are the methods by Swirski and Dodgson [59] as well as SynthesEyes and UnityEyes by Wood et al. [68, 69], the latter of which was more recently extended into a fully morphable 3D eye region model [67]. While both methods allow synthesis of annotated eye images for different camera positions, they differ in that [68] uses a more realistic eye region model and can simulate different lighting conditions. We therefore opted to use UnityEyes for part of our evaluation.

3 MULTI-VIEW LOW-RESOLUTION MOBILE EYE TRACKING

The goal of this work is to design a fully-integrated, *invisible* eye tracking device. As illustrated in Figure 1, our proposed system consists of eye cameras fully embedded into ordinary eyeglasses. While the scene camera is still expected to have higher resolution, the eye cameras are expected to be built with tiny low-resolution imaging sensors. Since the use of low-resolution and low-quality eye images leads to a fundamental difficulty in employing the conventional mobile eye tracking approaches through, e.g., eye landmark detection, we further propose to take a machine learning-based approach for gaze estimation. Here, the specific technical challenges are: 1) whether such tiny imaging sensors are available, and 2) what is the minimum image quality and resolution, as well as the minimum number of sensors, required for mobile learning-based gaze estimation. Considering previous works that have used individual photo transistors for gaze estimation [26, 61], in this work we explore eye image resolutions as low as 1×1 pixels.

In terms of sensor footprint, millimetre-size RGB cameras are available on the market mainly for medical imaging purposes such as endoscopy. Figure 2 shows a fully integrated prototype of our proposed

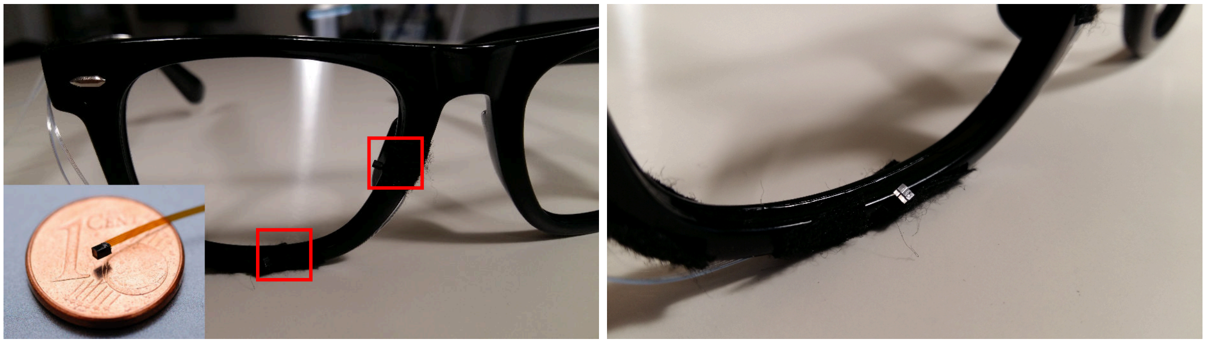


Fig. 2. Fully integrated version of *InvisibleEye* consisting of multiple, millimetre-size Awiba NanEye RGB cameras (marked in red) that are invisibly integrated into an off-the-shelf glasses frame. For our evaluations we developed two other prototypes to be able to characterise performance across key design parameters, including image resolution and number of cameras, as well as camera angle and positioning, and to compare with a state-of-the-art (high-resolution) mobile eye tracker.

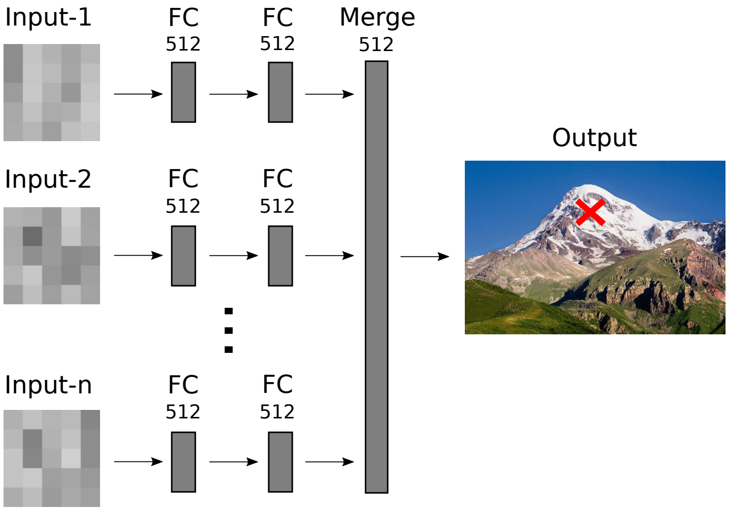


Fig. 3. Overview of the neural network used in our work for learning-based gaze estimation. The network takes multiple low-resolution eye images as input. Each image is encoded using two fully connected (FC) layers and image-specific representations are then merged to jointly predict gaze direction in scene camera coordinates (red cross).

system using an off-the-shelf glasses frame and medical-purpose millimetre-size cameras. In this prototype and one of the following experiments, we used the Awaiba NanEye camera which has a footprint of only 1 × 1 mm [51]. As can be seen, this hardware concept using tiny eye cameras enables extremely unobtrusive design. In addition, to compensate for both low image quality and limited visibility of non-adjustable embedded cameras, we further propose to use multiple low-resolution eye images as input to the gaze estimation pipeline.

3.1 Neural Network for Multi-View Gaze Estimation

As discussed above, we propose to take a machine learning-based gaze estimation approach. Using a set of training (calibration) eye images associated with ground-truth gaze positions in the scene camera, our system trains a gaze estimation function that can directly output gaze positions from arbitrary input eye images. Prior work on remote appearance-based estimation already demonstrated that, in the ideal case, only a 15-dimensional feature representation (eye image of 3×5 pixels) is sufficient to achieve less than one degree of accuracy [35]. Inspired by such prior attempts, in this work we examine the machine learning pipeline assuming low-resolution cases.

We use an artificial neural network as illustrated in Figure 3 to learn a mapping from low-resolution eye images to gaze positions. We assume the existence of training (calibration) data from the target user, and train a person-specific mapping function for each user. Unlike prior work [5], our method takes multiple eye images obtained from the tiny wearable eye cameras and learn a joint mapping function from all eye images. While there is a trade-off between the depth and performance of the neural network, our proposed network architecture is designed to be sufficiently shallow to reduce training time and inference time at run-time. Separate stacks of two fully connected layers with 512 hidden units and ReLU activation take raster-scanned image vectors from each of the N eye cameras as input. The outputs of those stacks are merged in another fully connected layer with 512 hidden units, and the output is predicted by a linear regression layer. The network is trained to jointly predict the x- and y-coordinate of the gaze positions, and the loss function is defined as the mean absolute distance between the predicted and ground-truth values. We implemented the network using Keras [19] with the Tensorflow [1] backend and chose the Adagrad algorithm [20] as optimiser with a learning rate of $lr = 0.005$. We trained our models on a modern i7-6850K CPU, on which training until convergence took about 1-2 minutes in all cases. At test time, we achieved ~ 700 frames per second (FPS) on the same CPU using a single core. When using a Nvidia GeForce GTX 1080 Ti GPU we achieved up to ~ 850 FPS. For comparison, for the gaze estimation pipeline of Pupil Labs [28], a commercial, state-of-the art mobile eye tracker, we achieved only ~ 270 FPS. These results indicate the significantly smaller amount of computation required for *InvisibleEye* and thus its potential for mobile and embedded platforms that have only limited computational power.

4 EXPERIMENTS

To systematically explore the feasibility and performance of *InvisibleEye* we conducted a series of experiments on three large-scale and increasingly difficult datasets, two of which we collected specifically for the purpose of this work. Experiment 1 was conducted in an idealised setting using synthesised eye images. Synthesising the eye images allowed us to use an arbitrary number of "virtual" cameras in different positions, which would not be possible when recording with real cameras. For Experiment 2 we implemented a first prototype to record real data in a constraint environment. This allowed us to control several of the parameters that make mobile gaze estimation difficult, in particular slippage of the headgear or changes in lighting conditions. Experiment 3 evaluated the performance of *InvisibleEye* in a challenging mobile real-world setting using a second prototype. It is important to note that, in all experiments that follow, the network was trained in a person-dependent fashion, i.e., trained for each user individually with person-specific training data. In the following, we report on each of these experiments in turn.

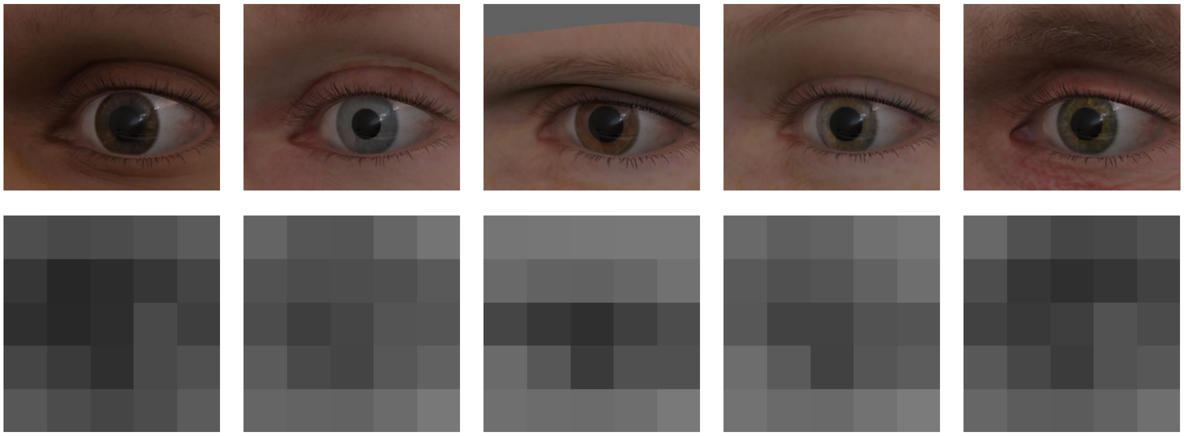


Fig. 4. (top row) Sample eye images from the original UnityEyes dataset [68] and the corresponding low-resolution grey-scale images (bottom row) that were used as input to the learning-based gaze estimation method.

4.1 Experiment 1: Evaluation on Synthetic Images

Before constructing the first hardware prototype for *InvisibleEye*, we opted to investigate the design space using synthetic eye image data. The goal of Experiment 1 on these synthetic images was to evaluate the minimum number and positions of cameras.

4.1.1 Data Synthesis. The dataset for Experiment 1 was generated using UnityEyes, a computer graphics eye region model to synthesise highly-realistic and perfectly annotated eye region images [68]. UnityEyes combines a novel generative 3D model of the human eye region with a real-time rendering framework. The model is based on high-resolution 3D face scans and uses real-time approximations for complex eyeball materials and structures as well as anatomically inspired procedural geometry methods for eyelid animation. Using UnityEyes, we synthesised images for five different eye regions as illustrated in Figure 4. We used a uniform 5×5 grid of camera angles to synthesise the images (see Figure 6a). The used camera angles span the full range of angles UnityEyes is capable of synthesising, which is a frontal view as one extreme, and views that are increasingly bottom-up or from the side. Top-down views were largely occluded by the ridge bone and were therefore not considered here. For each combination of eye region, camera angle, and lighting condition, we recorded a set of 1,600 different eyeball poses, corresponding to a uniform 40×40 grid of gaze angles. The step size in this grid was 1°, so the dataset covers a horizontal and vertical field of view of 40°. Each set was randomly split into a set of 1,280 training images and 320 test images. The images produced by UnityEyes are of high resolution and we therefore down-sampled them to resolutions below 20×20 pixels to simulate the images a low-quality sensor would yield. We also converted them to grayscale to further lower their dimension.

4.1.2 Results. To investigate the difficulty of estimating gaze with extremely low resolution images and the capabilities of *InvisibleEye* at this task, we trained different neural networks for different image resolutions. For a baseline comparison we also computed results using k-Nearest-Neighbours (kNN) with $k = 5$. Furthermore, we evaluated all approaches for different numbers of cameras. For the kNN approach we concatenated the corresponding images of different cameras before training. The results of this series of experiments are summarised in Figure 5a. As can be seen from the figure, both kNN and our approach achieve very low gaze estimation error. For example, at 10×10 image resolution, using

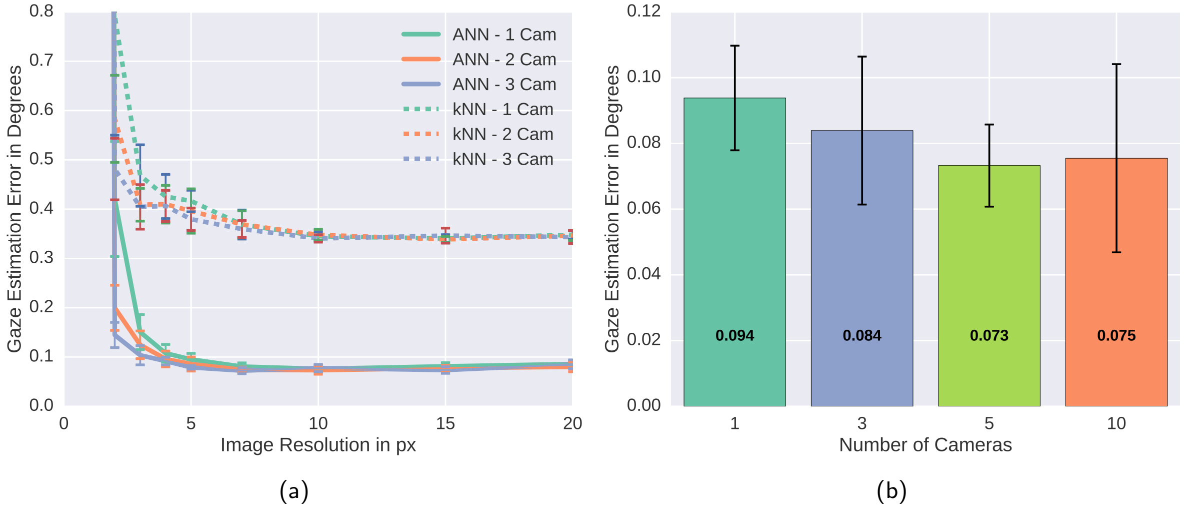


Fig. 5. (a) Average gaze estimation error for different image resolutions for a k-Nearest-Neighbours approach and our suggested neural network approach. (b) Gaze estimation error for different numbers of cameras at 5-pixel resolution. Each bar corresponds to the error of the best combination of x cameras out of 5 randomly selected sets.

a single camera, kNN achieves 0.345° error and the neural network achieved 0.078°. One can also see that there is little benefit in increasing the resolution of the input images. For both approaches, however, the figure also shows that the addition of cameras to the system helps to improve the results, especially for very low resolutions. At 3×3-pixel image resolution, for example, the result of the neural network improves from 0.15° error to 0.12° and 0.1° error for two and three cameras respectively, which is an improvement of 20% and 33%. Figure 5b shows the results for even higher numbers of cameras. As one can see, additional cameras help to improve performance slightly, but beyond four to five cameras the error does not significantly decrease any further.

Besides choosing the right number of cameras, another important parameter is the positioning of those cameras. One would like them to have an informative view but not to occlude the user's field of view. Figure 6b shows the error when using every available camera individually. As one can see, frontal views of the eye yield the lowest error, while bottom-up views are superior to side views. The worst result is achieved with the highly off-axis view from the very bottom and on the far side.

4.1.3 Discussion. Although the results achieved on synthetic data do not directly translate to the real world, since the gaze estimation task is a lot easier without real-world noise, the first set of experiments clearly demonstrates that mobile gaze estimation does not necessarily require high-resolution images. Further, we found that using multiple cameras can improve performance, but more than three to four cameras are unlikely to yield significant improvements. These results thus serve as important guidelines for designing *InvisibleEye* prototypes, which will be discussed in the following sections. We also found that frontal views of the eye yield the best results. We believe this is because frontal views have the least occluded view of the pupil and iris (e.g. with respect to the eyelashes), resulting in more distinct features for gaze estimation. However, since one of the key attributes of *InvisibleEye* should be that its cameras are in non-occluding positions, frontal views are not an option in practice. Since bottom-up

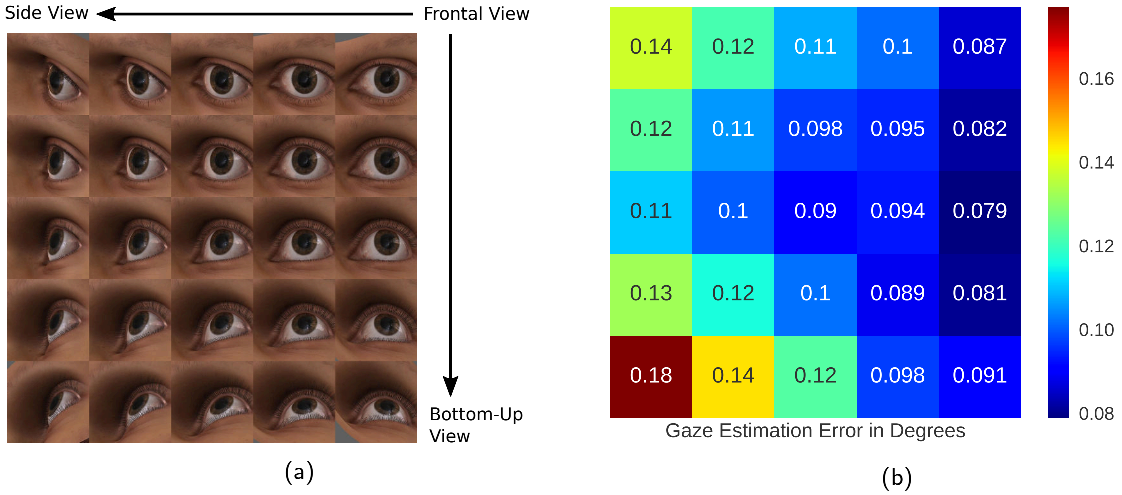


Fig. 6. (a) The use of synthetic images allows us to explore a wide range of camera angles (25 in this work) in an efficient and principled manner. (b) Average gaze estimation error in degrees when evaluating for each individual angle.

views and pure side views were the next best options according to these first experiments, we positioned the cameras in corresponding positions in our prototypes.

4.2 Experiment 2: Evaluation in a Controlled Laboratory Setting

Based on our experiments on synthetic eye images, we built a hardware prototype of *InvisibleEye* to evaluate its performance on real images. We conducted the second experiment using this prototype in a controlled laboratory environment. As discussed earlier, we used Awaiba NanEye cameras to achieve the small footprint of 1 × 1 mm. The NanEye cameras have an image resolution of 250 × 250 pixels and can capture images at 44 frames per second. Although the form factor of this medium-resolution camera is already sufficient to realise fully invisible mobile eye tracking (see Figure 2), we wanted to explore even lower image resolutions, i.e. below 20 × 20 pixels, which also promises further decreased bandwidth and computational requirements. We therefore opted to simulate this setting by artificially degrading the image resolution further.

The prototype was built by attaching four NanEye cameras to a pair of safety glasses. The NanEye cameras are very fragile and, since they are so small, also difficult to work with. We therefore opted to use safety glasses as the basis of our prototype, because it allowed us to carefully attach the cameras to the glass. The number of cameras and their positioning was motivated by the results of Experiment 1, i.e. two cameras were positioned with bottom-up views of the eye and one camera each was positioned on the far left and right side of the eye. The cameras were attached using “Blu-Tack”, a reusable putty-like pressure-sensitive adhesive. Since we attached the cameras to a pair of panoramic safety glasses, their angles are similar to what they would be in a regular glasses frame. The main difference in the angles is, that they are further away from the eye than they would be in a regular frame. We compensated for this by cropping the image by 25% from the center in each direction, which has a similar effect on the image as moving the camera closer to the eye while reducing the resolution.

4.2.1 Data Collection. We used this first hardware prototype to record a dataset of more than 280,000 close-up eye images with ground truth annotation of the gaze location. Figure 8 shows a few example

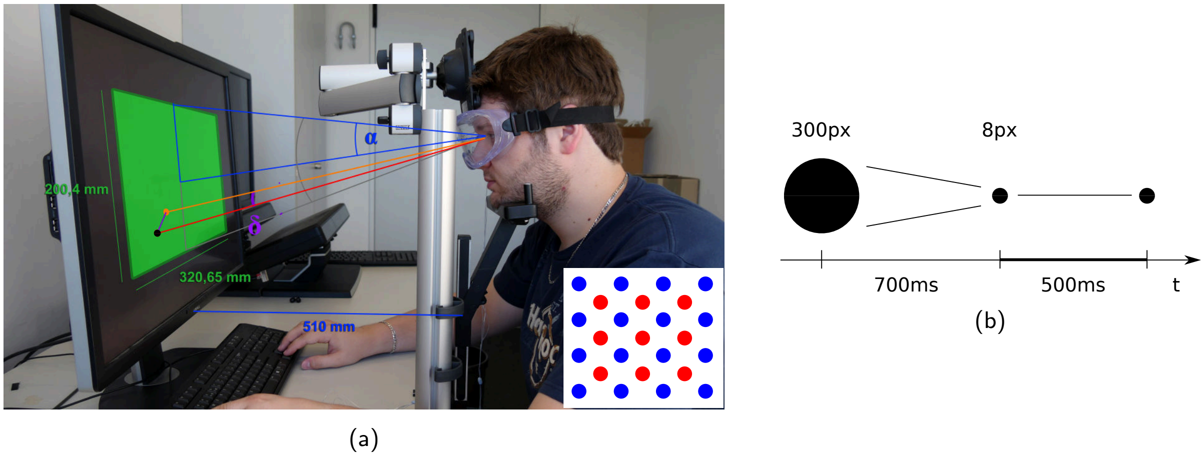


Fig. 7. (a) Overview of the recording setup used in Experiment 2 with a participant wearing the first prototype and resting his head on a chin rest. Ground truth gaze targets (marked in black) were shown in the central area of the screen covering $2 \cdot \alpha$ of the participant's visual field (marked in green). The angular error δ (purple) could then be calculated as the distance between true and predicted gaze targets (marked in orange). On-screen gaze targets were distributed in a grid and split into training (blue) and test data (red). (b) To increase ground truth accuracy, gaze targets were shown with a shrinking animation for 700 ms, and then for another 500 ms at the smallest size. Data was only recorded during the latter 500 ms.

images indicating the positional differences between the cameras and the impacts of cropping and down-sampling the images. A total of 17 participants were recorded, covering a wide range of appearances:

- **Gender**: Five (29%) female and 12 (71%) male
- **Nationality**: Seven (41%) German, seven (41%) Indian, one (6%) Bangladeshi, one (6%) Iranian, and one (6%) Greek
- **Eye Color**: 12 (70%) brown, four (23%) blue, and one (5%) green
- **Glasses**: Four participants (23%) wore regular glasses and one (6%) wore contact lenses

For each participant, two sets of data were recorded: one set of of training data and a separate set of test data. For each set, a series of gaze targets was shown on a display that participants were instructed to look at. For both training and test data the gaze targets covered a uniform grid in a random order, where the grid corresponding to the test data was positioned to lie in between the training points (see Figure 7a). Since the NanEye cameras record at about 44 FPS, we gathered approximately 22 frames per camera and gaze target. The training data was recorded using a uniform 24 × 17 grid of points, with an angular distance in gaze angle of 1.45° horizontally and 1.30° vertically between the points. In total the training set contained about 8,800 images per camera and participant. The test set's points belonged to a 23 × 16 grid of points and it contains about 8,000 images per camera and participant. This way, the gaze targets covered a field of view of 35° horizontally and 22° vertically.

The recording procedure was split into two parts for training and test data. For both parts, participants were instructed to put on the prototype and rest their head on a chin rest positioned exactly 510 mm in front of a display. The display was a 30-inch LED monitor with a pixel pitch of 0.25 mm and viewable image dimensions of 641.3×400.8 mm, set to 2560×1600-pixel resolution. On the display, the grid of gaze targets was shown, which the participants were instructed to look at. Each point appeared as a big circle

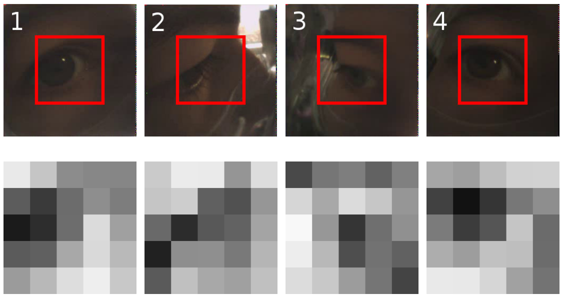
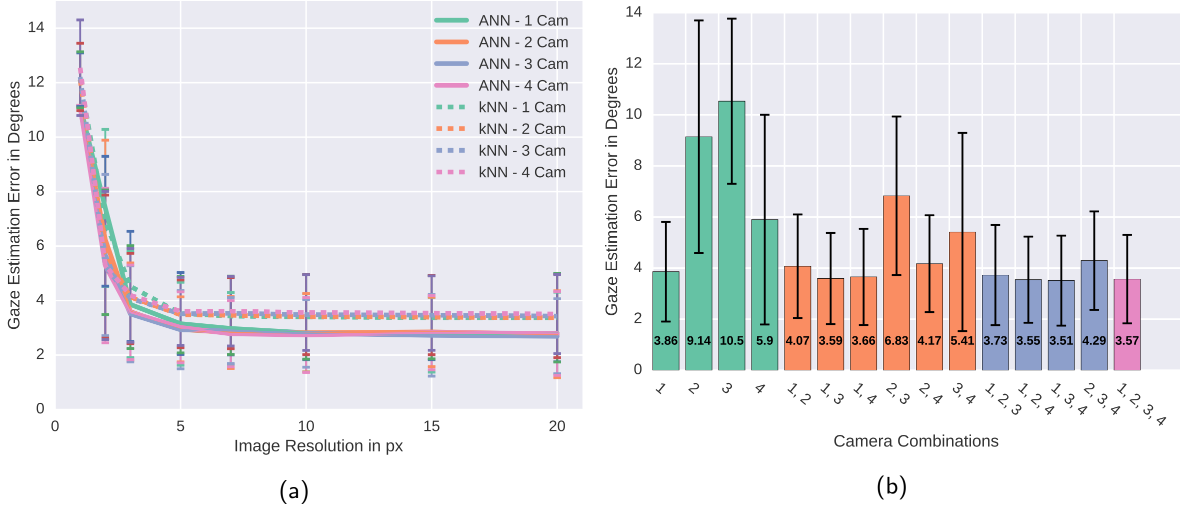


Fig. 8. (top) Sample eye images from one participant recorded using four NanEye4 cameras. We identify each of the cameras by the number at the top-left. (bottom) Corresponding cropped low-resolution versions of these images.



(a)

(b)

Fig. 9. (a) Average gaze estimation error for different image resolutions for a k-Nearest-Neighbours approach and our suggested neural network approach on the controlled laboratory data. (b) Average gaze estimation error for different numbers of cameras at 3-pixel resolution. Please refer to Figure 8 for the camera-label assignment.

300 pixels in diameter and shrunk to a circle of 8 pixels diameter over the course of 700 ms. The small circle was then displayed for another 500 ms, until the display of the next point started. Data was only recorded during the latter 500 ms, i.e. while the small circle was shown (see Figure 7a). It is important to note that the chin rest did not fully restrain participants and we noticed that their head sometimes moved noticeably, thus resulting in a certain amount of label noise. Using the shrinking animation for the circle helps the participants to locate the circle on the screen and gives them time to relocate their gaze. Similar to [30], we also showed an “L” or an “R” in between every 20th pair of points in the sequence. The letter was displayed for 500 ms at the position of the last point. Participants were asked to confirm the letter they had seen by pressing the corresponding left or right arrow-key. This was done to ensure participants focused on the gaze targets and task at hand throughout the recording.

The data is publicly available at *http://www.mpi-inf.mpg.de/invisibleeye*.

4.2.2 Results. We again computed the performance of *InvisibleEye* for different resolutions and camera combinations. Figure 9a shows the performance for different resolutions and up to three cameras. Compared to the synthetic case, one can see that the gaze estimation error is now considerably higher but still follows a similar distribution as before. Specifically, for resolutions above 5×5 pixels, the error remains stable with for example 2.9° error for the ANN and 3.52° error for kNN with one camera at exactly 5×5 pixels. These error values are in a range that is low enough for many practical applications like activity recognition [16] or attention analysis [56]. However, if we consider Figure 9b we can see that additional cameras do not help for every combination of cameras. Instead, combining cameras that perform worse individually achieves the biggest increase in performance.

4.2.3 Discussion. We have seen that even for very low image resolutions of only 3×3 pixels, *InvisibleEye* is capable of estimating gaze at a low error of 3.86° with a single camera and 3.57° when combining four cameras. This shows that gaze estimation at these low resolutions is possible with real-world data at an accuracy that is practically relevant. These error values further represent an upper bound to what *InvisibleEye* can achieve in this setting, due to the label noise in the data. In the following experiment we will see that, although we move into a more difficult setting, the achieved errors will be even lower since we do not have as much label noise in the data.

Furthermore, the results suggest that combining multiple cameras does not yield a benefit in every case but can improve performance markedly when combining cameras that perform badly individually. Since, in practice, one will always have design constraints on the hardware and a different fit of the device on every user, one runs the risk of positioning the cameras badly for at least some participants. The possibility of combining the information from multiple bad cameras is therefore highly relevant in practice.

4.3 Experiment 3: Evaluation in a Mobile Setting

In the controlled setting, we assumed a display at a fixed distance in front of the user and predicted gaze in the screen coordinate system. In practice, however, we want to allow users to move around freely and still be able to track gaze on all kinds of objects, not only displays. Bridging this gap between the controlled laboratory setting and the real world requires adding a scene camera to the system that records the user's field of view and allows us to estimate gaze in scene camera coordinates.

We built a second hardware prototype featuring such a scene camera to test *InvisibleEye* in a mobile setting. We also explicitly allowed gaze targets at arbitrary depths. The depth at which a gaze target lies directly correlates with the location of the target projected into the camera image. From only the view of one eye, this location in the image is, however, in general not inferable. If, for example, the target is moved along the gaze ray projected from the recorded eye into the world, the image of the eye will not change at all if it keeps gazing at the target, while the location of the target in the scene camera image might change considerably [6]. It is therefore necessary to use views from both eyes to resolve this ambiguity, which we do by using symmetric pairs of cameras recording both eyes. Further, we explicitly allow slippage of the headset, which is a problem frequently occurring in practice.

For this second prototype we decided against using NanEye cameras mainly because comparison with state-of-the-art mobile gaze estimation methods is impossible due to the lower image resolution. We instead used Pupil Labs cameras [28] to record the eyes and the scene using a custom-built, 3D printed frame (see Figure 10). Note that, unlike NanEye cameras, the Pupil Labs eye cameras record infrared images of the eye similarly as most cameras in commercially available eye trackers. The field of view of the scene camera was approximately 80° × 60°. Please note that although these cameras are slightly

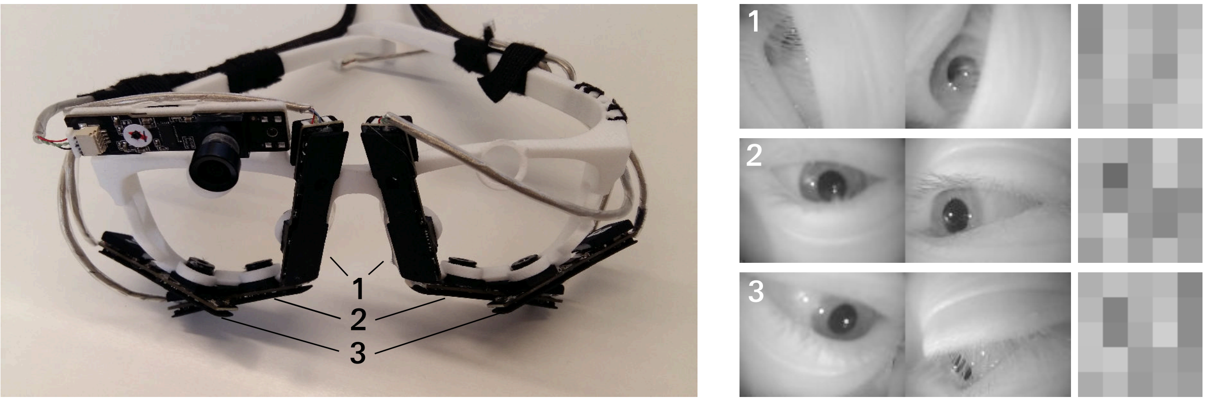


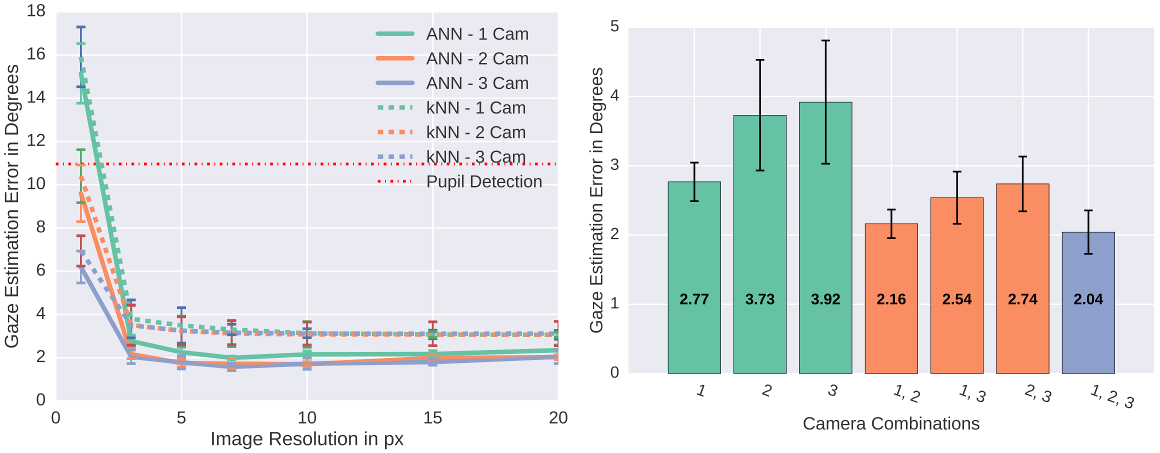
Fig. 10. (left) Second prototype consisting of a custom 3D-printed glasses frame that can hold up to six Pupil Labs [28] eye cameras with an additional scene camera. (right) Sample images recorded with the prototype with original image on the left and corresponding low-resolution counterparts on the right. We identify camera pairs by the number in the top-left corner.

bigger, they are now located directly in the frame of a pair of glasses, i.e. their viewing angles are exactly as they should be.

4.3.1 Data Collection. Using this prototype, we recorded another dataset of 240,000 eye images with four participants (four male, aged between 24 and 38 years). To record gaze data at varying distances in a mobile setting, a calibration marker was attached to a wall in front of the participants. Participants were asked to position themselves at an arbitrary distance of up to 3 meters in front of the marker and to perform a series of head movements while gazing at the marker. The head movements consisted of continuously moving the head upwards and downwards while rotating it from the far left to the far right within approximately 10 seconds. Participants were asked to perform the movement such that the marker would move to the edge of their field of view but always remain visible, so they could gaze at it. After performing the head movements, participants were asked to position themselves at a new randomly selected distance for another recording. We repeated this procedure for the whole duration of the recording session. Additionally, to simulate slippage of the headset that is pervasive in mobile settings [54], participants were asked to take off the headset and to put it back on after every 6th recording. Each recording session lasted for about 15 minutes and every participant performed a total of four sessions. This way we were able to efficiently gather images for gaze angles of a large field of view of roughly 70 × 60°.

Each eye image was automatically labelled with the position of the calibration marker in the scene camera. All cameras were set to record images of 640 × 480 pixels resolution at 120 Hz. Per session, approximately 30,000 images were recorded by each camera. To reduce the required time for training, we reduced the training set to a random subset of 15,000 images. The data of the first three sessions was used as training data, while the data of the fourth session was used for testing. Given that the data was recorded indoors, the images recorded by the infrared cameras were not subject to any significant changes in lighting conditions.

As before, the images we recorded with this second prototype were of much higher quality than what we required for *InvisibleEye*. We therefore down-sampled the images to a lower resolution. We did not crop the images this time because the cameras were sufficiently close to the eye in this second prototype.



(a)

(b)

Fig. 11. (a) Average gaze estimation error for different image resolutions and number of cameras. Also shown is the performance of a state-of-the-art (high-resolution) mobile eye tracking method. (b) Average gaze estimation error for every possible combination of cameras averaged across all participants for an image resolution of 3×3 pixels. Please refer to Figure 10 for the camera-label assignment.

The images recorded from each camera pair, i.e. one camera from the left side and its symmetrical counterpart from the right side, were concatenated before the process. Sample images and corresponding low-resolution versions are shown in Figure 10. For gaze estimation, we used the same neural network architecture as before.

4.3.2 Results. Because the data recorded with our prototype for the mobile setting was recorded with high-quality cameras, we first computed a baseline performance using a state-of-the-art gaze estimation approach based on pupil detection on the original high-resolution images. For this we used the previously mentioned publicly available pipeline from Pupil Labs [28]. We randomly picked 200 images out of the first 1,650 images recorded for every participant as calibration data. Due to the continuous and random head movement during recording sessions, the first 1,650 images already cover the entire field of view of the participant and thus represent a realistic set of calibration images. We picked only images recorded by the left camera of pair number two, since this camera position is the closet to that of traditional eye trackers. After detecting the pupil positions in all calibration images, the next step in the Pupil Labs pipeline is to fit a 7th order polynomial to map the pupil positions to the ground truth gaze positions. Using this polynomial, we then estimated gaze positions for all other images from the same participant using the detected pupil position in each image. This baseline method achieved an error of 10.96°. This high error is due to the strong slippage of the headset that is present in the data but not being compensated for. By comparing the positions of one eye corner in a random subset of images, which can be interpreted as an estimate of this slippage, we found that the average distance to the centroid of all eye corner positions was 36.3 pixels.

Similar as before, we evaluated the average gaze estimation performance of *InvisibleEye* for increasingly lower resolutions as well as the number of used cameras. The results of this analysis are shown in Figure 11a. As we can see, the curves look similar to corresponding ones in the constraint setting, i.e. for resolutions larger than 5×5 pixels the performance remains stable, whereas it drops for lower resolutions. At 5×5-pixel resolution the average error when using all three camera pairs was 1.79°. The average error when using only a single camera was 2.25°. Thus, the use of multiple views of the eye led to a performance increase of approximately 20%. Figure 11b shows the average performance of every possible combination of camera pairs across all participants. Here we can see that, in all cases, the addition of a second camera pair improved the results on average.

4.3.3 Discussion. Experiment 3 has shown that for images recorded using cameras positioned around the frame, even if using high-resolution images classical approaches based on pupil detection perform badly. We showed that in contrast, for the same camera positions, *InvisibleEye* achieves a better performance, even for image resolutions as low as 3×3 pixels (corresponding to an error of 2.04°) using three cameras. This result shows that *InvisibleEye* is a viable option even in difficult settings. In this setting we have also seen that adding more cameras can improve performance. This might indicate that the apparent camera angles are difficult enough by themselves and that they can complement each other well, as was the case for cameras 2 and 3 in the controlled laboratory setting.

5 DISCUSSION

In this work we introduced *InvisibleEye*, a new approach that addresses several key challenges of current mobile eye trackers. The key novelty of our approach is the combination of small and low-quality cameras with an image resolution as low as 3×3 pixels with a method for learning-based gaze estimation. Our experiments show that despite the very low image resolution, *InvisibleEye* can still achieve an accuracy of 2.25° at 5×5-pixel image resolution when using a single pair of cameras in a mobile setting. We have also shown that using three pairs of cameras capturing different views of the eye can further improve performance to 1.79°. The hardware requirements for an embedded system to run *InvisibleEye* at test time are also very low. While model training might be feasible on a mobile device, it could be outsourced to a standard desktop machine or a cloud service too, making *InvisibleEye* easy to deploy in practice. These findings are highly encouraging given that they not only demonstrate the feasibility of our approach but, more importantly, underline its potential for finally realising the vision of invisible mobile eye tracking. Despite these promising results, our *InvisibleEye* prototypes still have several limitations. First, all evaluations shown here are based on person-specific training, i.e. every user needs to record training data with the device prior to first use. It is important to note, however, that while highly undesirable from a usability point of view, the requirement for person-specific training or calibration also applies to state-of-the-art mobile eye trackers that use a classic gaze estimation method and person-specific calibration. Nonetheless, the amount of person-specific training data currently required for our method is still significantly larger than the one for standard calibration approaches. Methods from transfer learning could, for example, be used to reduce the amount of required training data, and it is also promising to investigate implicit calibration approaches.

The ability to robustly estimate gaze across the large variability in eye appearances of different people is a significantly more challenging task and thus represents the most important direction for future work. A less challenging yet still highly practical solution could be eye tracker self-calibration in which gaze positions are inferred, for example, from saliency maps calculated from the scene camera images [54]. This has the potential to allow the user to gather training data naturally just by wearing the device for an extended amount of time, thereby continuously improving performance during everyday use.

Second, in this work we have not yet evaluated the performance of *InvisibleEye* in an outdoor environment, nor during long-term recordings. Usually, mobile eye tracking systems perform a lot worse outdoors because the sun can create intense reflections and shadows on the eye image [60]. It remains to be explored if a learning-based approach can improve the robustness in such challenging environments.

Finally, while the two prototype systems of *InvisibleEye* that we have built were sufficient to investigate its performance in both stationary controlled and mobile real-world settings, a fully integrated mobile eye tracker that can be used robustly in daily life is still highly desirable. Currently, such full integration is not possible with the NanEye cameras used in this work, given that they have to be connected to a desktop computer using a special-purpose USB breakout board. The cameras do use a standard video interface, however, which makes us confident that fully embedded integration of both hardware and software will soon be feasible.

6 CONCLUSION

In this work we presented *InvisibleEye* – a novel approach that, in contrast to a long line of work on mobile eye tracking, relies on tiny cameras that can be nearly invisibly integrated into a normal glasses frame. To compensate for the cameras' low image resolution of only a few pixels, we showed how to combine multiple of them using a learning-based gaze estimation method that directly regresses from eye images to gaze directions. We evaluated our system on three increasingly challenging datasets to study its performance across key design parameters including image resolution, number of cameras, as well as camera angle and positioning. Our approach achieved a person-specific gaze estimation accuracy of 1.79° using four cameras with a resolution of only 5×5 pixels. These findings are promising and not only underline the potential of this new approach but mark an important step towards realising the vision of fully unobtrusive, comfortable, and socially acceptable mobile eye tracking.

7 ACKNOWLEDGEMENTS

This work was supported by the Cluster of Excellence on Multimodal Computing and Interaction at Saarland University, Germany, and a Fellowship for Postdoctoral Researchers from the Alexander von Humboldt Foundation.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, and others. 2016. Tensorflow: Large-scale Machine Learning on Heterogeneous Distributed Systems. *arXiv preprint arXiv:1603.04467* (2016).

[2] Evgeniy Abdulin, Ioannis Rigas, and Oleg Komogortsev. 2016. Eye Movement Biometrics on Wearable Devices: What Are the Limits?. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1503–1509.

[3] Deepak Akkil, Jari Kangas, Jussi Rantala, Poika Isokoski, Oleg Spakov, and Roope Raisamo. 2015. Glance Awareness and Gaze Interaction in Smartwatches. In *Proc. of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (HCI)*. ACM, 1271–1276. DOI:http://dx.doi.org/10.1145/2702613.2732816

[4] Nuri Murat Arar, Hua Gao, and Jean-Philippe Thiran. 2015. Robust Gaze Estimation Based on Adaptive Fusion of Multiple Cameras. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, Vol. 1. IEEE, 1–7. DOI:http://dx.doi.org/10.1109/FG.2015.7163121

[5] Shumeet Baluja and Dean Pomerleau. 1994. *Non-intrusive Gaze Tracking Using Artificial Neural Networks.* Technical Report. DTIC Document.

[6] Michael Barz, Florian Daiber, and Andreas Bulling. 2016. Prediction of Gaze Estimation Error for Error-Aware Gaze-Based Interfaces. In *Proc. International Symposium on Eye Tracking Research and Applications (ETRA)*. 275–278. DOI:http://dx.doi.org/10.1145/2857491.2857493

[7] Roman Bednarik, Hana Vrzakova, and Michal Hradis. 2012. What Do You Want to Do Next: A Novel Approach for Intent Prediction in Gaze-based Interaction. In *Proc. International Symposium on Eye Tracking Research and Applications (ETRA)*. ACM, 83–90. DOI:http://dx.doi.org/10.1145/2168556.2168569

[8] Frank H Borsato and Carlos H Morimoto. 2016. Episcleral Surface Tracking: Challenges and Possibilities for Using Mice Sensors for Wearable Eye Tracking. In *Proc. International Symposium on Eye Tracking Research and Applications (ETRA)*. ACM, 39–46. DOI:http://dx.doi.org/10.1145/2857491.2857496

[9] Andreas Bulling and Hans Gellersen. 2010. Toward Mobile Eye-Based Human-Computer Interaction. *IEEE Pervasive Computing* 9, 4 (2010), 8–12. DOI:http://dx.doi.org/10.1109/MPRV.2010.86

[10] Andreas Bulling and Daniel Roggen. 2011. Recognition of Visual Memory Recall Processes Using Eye Movement Analysis. In *Proc. ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. 455–464. DOI:http://dx.doi.org/10.1145/2030112.2030172

[11] Andreas Bulling, Daniel Roggen, and Gerhard Tröster. 2008. It's in Your Eyes: Towards Context-awareness and Mobile HCI Using Wearable EOG Goggles. In *Proc. of the 10th International Conference on Ubiquitous Computing (UbiComp)*. ACM, 84–93. DOI:http://dx.doi.org/10.1145/1409635.1409647

[12] Andreas Bulling, Daniel Roggen, and Gerhard Tröster. 2009. Wearable EOG Goggles: Seamless Sensing and Context-awareness in Everyday Environments. *Journal of Ambient Intelligence and Smart Environments* 1, 2 (2009), 157–171. DOI:http://dx.doi.org/10.3233/AIS-2009-0020

[13] Andreas Bulling, Daniel Roggen, and Gerhard Tröster. 2011. What's in the Eyes for Context-Awareness? *IEEE Pervasive Computing* 10, 2 (April 2011), 48 – 57. DOI:http://dx.doi.org/10.1109/MPRV.2010.49

[14] Andreas Bulling, Jamie A. Ward, Hans Gellersen, and Gerhard Tröster. 2008. Robust Recognition of Reading Activity in Transit Using Wearable Electrooculography. In *Proc. International Conference on Pervasive Computing (Pervasive)*. 19–37. DOI:http://dx.doi.org/10.1007/978-3-540-79576-6_2

[15] Andreas Bulling, Jamie A. Ward, Hans Gellersen, and Gerhard Tröster. 2009. Eye Movement Analysis for Activity Recognition. In *Proc. ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. 41–50. DOI:http://dx.doi.org/10.1145/1620545.1620552

[16] Andreas Bulling, Jamie A Ward, Hans Gellersen, and Gerhard Troster. 2011. Eye movement analysis for activity recognition using electrooculography. *IEEE transactions on pattern analysis and machine intelligence* 33, 4 (2011), 741–753.

[17] Andreas Bulling, Christian Weichel, and Hans Gellersen. 2013. EyeContext: Recognition of High-level Contextual Cues from Human Visual Behaviour. In *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 305–308. DOI:http://dx.doi.org/10.1145/2470654.2470697

[18] Andreas Bulling and Thorsten O. Zander. 2014. Cognition-Aware Computing. *IEEE Pervasive Computing* 13, 3 (2014), 80–83. DOI:http://dx.doi.org/10.1109/mprv.2014.42

[19] François Chollet. 2015. Keras. (2015).

[20] John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research* 12, Jul (2011), 2121–2159.

[21] Augusto Esteves, Eduardo Velloso, Andreas Bulling, and Hans Gellersen. 2015. Orbits: Enabling Gaze Interaction in Smart Watches using Moving Targets. In *Proc. ACM Symposium on User Interface Software and Technology (UIST)*. 457–466. DOI:http://dx.doi.org/10.1145/2807442.2807499

[22] Wolfgang Fuhl, Thomas Kübler, Katrin Sippel, Wolfgang Rosenstiel, and Enkelejda Kasneci. 2015. ExCuSe: Robust Pupil Detection in Real-World Scenarios. In *International Conference on Computer Analysis of Images and Patterns (CAIP)*. Springer, 39–51.

[23] Wolfgang Fuhl, Thiago Santini, Gjergji Kasneci, and Enkelejda Kasneci. 2016. PupilNet: Convolutional Neural Networks for Robust Pupil Detection. *arXiv preprint arXiv:1601.04902* (2016).

[24] Wolfgang Fuhl, Thiago C Santini, Thomas Kübler, and Enkelejda Kasneci. 2016. ElSe: Ellipse Selection for Robust Pupil Detection in Real-World Environments. In *Proc. of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*. ACM, 123–130. DOI:http://dx.doi.org/10.1145/2857491.2857505

[25] Sabrina Hoppe, Tobias Loetscher, Stephanie Morey, and Andreas Bulling. 2015. Recognition of Curiosity Using Eye Movement Analysis. In *Adj. Proc. ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. 185–188. DOI:http://dx.doi.org/10.1145/2800835.2800910

[26] Yoshio Ishiguro, Adiyan Mujibiya, Takashi Miyaki, and Jun Rekimoto. 2010. Aided Eyes: Eye Activity Sensing for Daily Life. In *Proc. of the 1st Augmented Human International Conference*. ACM, 25. DOI:http://dx.doi.org/10.1145/1785455.1785480

[27] Amir-Homayoun Javadi, Zahra Hakimi, Morteza Barati, Vincent Walsh, and Lili Tcheang. 2015. SET: A Pupil Detection Method Using Sinusoidal Approximation. *Frontiers in neuroengineering* 8 (2015). DOI:http://dx.doi.org/10.3389/fneng.2015.00004

[28] Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction. In *Adj. Proc. ACM International Joint Conference on Pervasive and*

Ubiquitous Computing (UbiComp). 1151–1160. DOI:http://dx.doi.org/10.1145/2638728.2641695

[29] Elizabeth S. Kim, Adam Naples, Giuliana Vaccarino Gearty, Quan Wang, Seth Wallace, Carla Wall, Michael Perlmutter, Jennifer Kowitt, Linda Friedlaender, Brian Reichow, Fred Volkmar, and Frederick Shic. 2014. Development of an Untethered, Mobile, Low-cost Head-mounted Eye Tracker. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '14)*. ACM, New York, NY, USA, 247–250. DOI:http://dx.doi.org/10.1145/2578153.2578209

[30] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. 2016. Eye Tracking for Everyone. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2176–2184. DOI:http://dx.doi.org/10.1109/CVPR.2016.239

[31] Christian Lander, Sven Gehring, Antonio Krüger, Sebastian Boring, and Andreas Bulling. 2015. GazeProjector: Accurate Gaze Estimation and Seamless Gaze Interaction Across Multiple Displays. In *Proc. ACM Symposium on User Interface Software and Technology (UIST)*. DOI:http://dx.doi.org/10.1145/2807442.2807479

[32] Dongheng Li and Derrick Parkhurst. 2006. Open-Source Software for Real-Time Visible-Spectrum Eye Tracking. In *Proc. of the COGAIN Conference*, Vol. 17.

[33] Dongheng Li, David Winfield, and Derrick J Parkhurst. 2005. Starburst: A Hybrid Algorithm for Video-based Eye Tracking Combining Feature-based and Model-based Approaches. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*. IEEE, 79–79. DOI:http://dx.doi.org/10.1109/CVPR.2005.531

[34] Xindian Long, Ozan K Tonguz, and Alex Kiderman. 2007. A High Speed Eye Tracking System with Robust Pupil Center Estimation Algorithm. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 3331–3334. DOI:http://dx.doi.org/10.1109/IEMBS.2007.4353043

[35] 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 (TPAMI)* 36, 10 (2014), 2033–2046. DOI:http://dx.doi.org/10.1109/TPAMI.2014.2313123

[36] Päivi Majaranta and Andreas Bulling. 2014. *Eye Tracking and Eye-Based Human-Computer Interaction*. Springer Publishing London, 39–65. DOI:http://dx.doi.org/10.1007/978-1-4471-6392-3_3

[37] Hiroyuki Manabe and Masaaki Fukumoto. 2006. Full-time Wearable Headphone-type Gaze Detector. In *CHI'06 Extended Abstracts on Human Factors in Computing Systems*. ACM, 1073–1078. DOI:http://dx.doi.org/10.1145/1125451.1125655

[38] Mohsen Mansouryar, Julian Steil, Yusuke Sugano, and Andreas Bulling. 2016. 3D Gaze Estimation from 2D Pupil Positions on Monocular Head-Mounted Eye Trackers. In *Proc. International Symposium on Eye Tracking Research and Applications (ETRA)*. 197–200. DOI:http://dx.doi.org/10.1145/2857491.2857530

[39] Addison Mayberry, Pan Hu, Benjamin Marlin, Christopher Salthouse, and Deepak Ganesan. 2014. iShadow: Design of a Wearable, Real-Time Mobile Gaze Tracker. In *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*. ACM, 82–94. DOI:http://dx.doi.org/10.1145/2594368.2594388

[40] Atsushi Nakazawa and Christian Nitschke. 2012. Point of Gaze Estimation through Corneal Surface Reflection in an Active Illumination Environment. *Computer Vision–ECCV 2012* (2012), 159–172. DOI:http://dx.doi.org/10.1007/978-3-642-33709-3_12

[41] Eleni Nasiopoulos, Evan F Risko, Tom Foulsham, and Alan Kingstone. 2015. Wearable Computing: Will It Make People Prosocial? *British Journal of Psychology* 106, 2 (2015), 209–216. DOI:http://dx.doi.org/10.1111/bjop.12080

[42] Basilio Noris, Jean-Baptiste Keller, and Aude Billard. 2011. A Wearable Gaze Tracking System for Children in Unconstrained Environments. *Computer Vision and Image Understanding* 115, 4 (2011), 476–486. DOI:http://dx.doi.org/10.1016/j.cviu.2010.11.013

[43] Bernardo Rodrigues Pires, Michäel Devyver, Akihiro Tsukada, and Takeo Kanade. 2013. Unwrapping the Eye for Visible-spectrum Gaze Tracking on Wearable Devices. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*. IEEE, 369–376. DOI:http://dx.doi.org/10.1109/WACV.2013.6475042

[44] Alexander Plopski, Christian Nitschke, Kiyoshi Kiyokawa, Dieter Schmalstieg, and Haruo Takemura. 2015. Hybrid Eye Tracking: Combining Iris Contour and Corneal Imaging. In *ICAT-EGVE*. 183–190. DOI:http://dx.doi.org/10.2312/egve.20151327

[45] Evan F Risko and Alan Kingstone. 2011. Eyes Wide Shut: Implied Social Presence, Eye Tracking and Attention. *Attention, Perception, & Psychophysics* 73, 2 (2011), 291–296. DOI:http://dx.doi.org/10.3758/s13414-010-0042-1

[46] Ravikrishna Ruddarraju, Antonio Haro, and Irfan Essa. 2003. Fast Multiple Camera Head Pose Tracking. In *Vision Interface*, Vol. 2. Citeseer.

[47] Ravikrishna Ruddarraju, Antonio Haro, Kris Nagel, Quan T Tran, Irfan A Essa, Gregory Abowd, and Elizabeth D Mynatt. 2003. Perceptual User Interfaces Using Vision-based Eye Tracking. In *Proc. of the 5th international conference on Multimodal interfaces*. ACM, 227–233. DOI:http://dx.doi.org/10.1145/958432.958475

[48] Javier San Agustin, Henrik Skovsgaard, Emilie Mollenbach, Maria Barret, Martin Tall, Dan Witzner Hansen, and John Paulin Hansen. 2010. Evaluation of a Low-cost Open-source Gaze Tracker. In *Proc. of the 2010 Symposium on Eye-Tracking Research & Applications*. ACM, 77–80. DOI:http://dx.doi.org/10.1145/1743666.1743685

[49] Hosnieh Sattar, Mario Fritz, and Andreas Bulling. 2017. *Visual Decoding of Targets During Visual Search From Human Eye Fixations*. arXiv:1706.05993.

[50] Hosnieh Sattar, Sabine Müller, Mario Fritz, and Andreas Bulling. 2015. Prediction of Search Targets From Fixations in Open-world Settings. In *Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. 981–990. DOI:http://dx.doi.org/10.1109/CVPR.2015.7298700

[51] Ricardo Sousa, Martin Wäny, Pedro Santos, and Fernando Morgado-Dias. 2017. NanEye–An Endoscopy Sensor with 3D Image Synchronization. *IEEE Sensors Journal* 17 (2017), 623–631. Issue 3. DOI:http://dx.doi.org/10.1109/JSEN.2016.2631582

[52] Julian Steil and Andreas Bulling. 2015. Discovery of Everyday Human Activities From Long-term Visual Behaviour Using Topic Models. In *Proc. ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. 75–85. DOI:http://dx.doi.org/10.1145/2750858.2807520

[53] Sophie Stellmach and Raimund Dachselt. 2013. Still Looking: Investigating Seamless Gaze-supported Selection, Positioning, and Manipulation of Distant Targets. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 285–294. DOI:http://dx.doi.org/10.1145/2470654.2470695

[54] Yusuke Sugano and Andreas Bulling. 2015. Self-Calibrating Head-Mounted Eye Trackers Using Egocentric Visual Saliency. In *Proc. ACM Symposium on User Interface Software and Technology (UIST)*. 363–372. DOI:http://dx.doi.org/10.1145/2807442.2807445

[55] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. 2014. Learning-by-synthesis for appearance-based 3d gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1821–1828.

[56] Yusuke Sugano, Xucong Zhang, and Andreas Bulling. 2016. Aggregaze: Collective estimation of audience attention on public displays. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 821–831.

[57] Lech Świrski, Andreas Bulling, and Neil Dodgson. 2012. Robust real-time pupil tracking in highly off-axis images. In *Proceedings of the Symposium on Eye Tracking Research and Applications*. ACM, 173–176. DOI:http://dx.doi.org/10.1145/2168556.2168585

[58] Lech Świrski and Neil Dodgson. 2013. A Fully-automatic, Temporal Approach to Single Camera, Glint-free 3d Eye Model Fitting. *Proc. PETMEI* (2013).

[59] Lech Świrski and Neil Dodgson. 2014. Rendering Synthetic Ground Truth Images for Eye Tracker Evaluation. In *Proc. of the Symposium on Eye Tracking Research and Applications*. ACM, 219–222. DOI:http://dx.doi.org/10.1145/2578153.2578188

[60] Marc Tonsen, Xucong Zhang, Yusuke Sugano, and Andreas Bulling. 2016. Labelled pupils in the wild: a dataset for studying pupil detection in unconstrained environments. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*. ACM, 139–142. DOI:http://dx.doi.org/10.1145/2857491.2857520

[61] Cihan Topal, Serkan Gunal, Onur Koçdeviren, Atakan Doğan, and Ömer N Gerek. 2014. A Low-Computational Approach on Gaze Estimation With Eye Touch System. *IEEE transactions on Cybernetics* 44, 2 (2014), 228–239. DOI:http://dx.doi.org/10.1109/TCYB.2013.2252792

[62] Akihiro Tsukada, Motoki Shino, Michael Devyver, and Takeo Kanade. 2011. Illumination-Free Gaze Estimation Method for First-Person Vision Wearable Device. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2084–2091. DOI:http://dx.doi.org/10.1109/ICCVW.2011.6130505

[63] Jayson Turner, Andreas Bulling, Jason Alexander, and Hans Gellersen. 2014. Cross-Device Gaze-Supported Point-to-Point Content Transfer. In *Proc. International Symposium on Eye Tracking Research and Applications (ETRA)*. 19–26. DOI:http://dx.doi.org/10.1145/2578153.2578155

[64] Akira Utsumi, Kotaro Okamoto, Norihiro Hagita, and Kazuhiro Takahashi. 2012. Gaze Tracking in Wide Area Using Multiple Camera Observations. In *Proc. of the Symposium on Eye Tracking Research and Applications*. ACM, 273–276. DOI:http://dx.doi.org/10.1145/2168556.2168614

[65] Mélodie Vidal, Ken Pfeuffer, Andreas Bulling, and Hans Gellersen. 2013. Pursuits: eye-based interaction with moving targets. In *Ext. Abstr. ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 3147–3150. DOI:http://dx.doi.org/10.1145/2468356.2479632

[66] Michael Voit and Rainer Stiefelhagen. 2006. Tracking Head Pose and Focus of Attention with Multiple Far-field Cameras. In *Proc. of the 8th international conference on Multimodal interfaces*. ACM, 281–286. DOI:http://dx.doi.org/10.1145/1180995.1181050

[67] Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. 2016. A 3D Morphable Eye Region Model for Gaze Estimation. In *Proc. European Conference on Computer Vision (ECCV).* DOI:http://dx.doi.org/10.2312/egp.20161054

[68] 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 *Proc. International Symposium on Eye Tracking Research and Applications (ETRA).* ACM, 131–138. DOI:http://dx.doi.org/10.1145/2857491.2857492

[69] 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 *Proc. of the IEEE International Conference on Computer Vision (ICCV).* 3756–3764. DOI:http://dx.doi.org/10.1109/ICCV.2015.428

[70] Gregory J Zelinsky, Hossein Adeli, Yifan Peng, and Dimitris Samaras. 2013. Modelling eye movements in a categorical search task. *Philosophical Transactions of the Royal Society B: Biological Sciences* 368, 1628 (2013), 20130058.

[71] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2015. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 4511–4520. DOI:http://dx.doi.org/10.1109/CVPR.2015.7299081

[72] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2017. It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation. In *Proc. IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).*

[73] Yanxia Zhang, Hans Jörg Müller, Ming Ki Chong, Andreas Bulling, and Hans Gellersen. 2014. GazeHorizon: Enabling Passers-by to Interact with Public Displays by Gaze. In *Proc. ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp).* 559–563. DOI:http://dx.doi.org/10.1145/2632048.2636071