Learning to Find Eye Region Landmarks for Remote Gaze Estimation in Unconstrained Settings

Seonwook Park  
ETH Zurich  
spark@inf.ethz.ch

Xucong Zhang  
MPI for Informatics  
xczhang@mpi-inf.mpg.de

Andreas Bulling  
MPI for Informatics  
bulling@mpi-inf.mpg.de

Otmar Hilliges  
ETH Zurich  
otmarh@inf.ethz.ch

ABSTRACT

Conventional feature-based and model-based gaze estimation methods have proven to perform well in settings with controlled illumination and specialized cameras. In unconstrained real-world settings, however, such methods are surpassed by recent appearance-based methods due to difficulties in modeling factors such as illumination changes and other visual artifacts. We present a novel learning-based method for eye region landmark localization that enables conventional methods to be competitive to latest appearance-based methods. Despite having been trained exclusively on synthetic data, our method exceeds the state of the art for iris localization and eye shape registration on real-world imagery. We then use the detected landmarks as input to iterative model-fitting and lightweight learning-based gaze estimation methods. Our approach outperforms existing model-fitting and appearance-based methods in the context of person-independent and personalized gaze estimation.

CCS CONCEPTS

• Human-centered computing → Pointing; • Computing methodologies → Computer vision;

KEYWORDS

Gaze Estimation; Eye Region Landmark Localization

1 INTRODUCTION

Gaze estimation using off-the-shelf cameras, including those in mobile devices, can assist users with motor-disabilities or enable crowd-sourced visual saliency estimation without the cost of specialized hardware [Xu et al. 2015]. Such methods can also improve user experience in everyday tasks, such as reading [Biedert et al. 2010; Kunze et al. 2013], or facilitate gaze-based interaction [Majaranta and Bulling 2014]. However, existing gaze estimation systems can fail when encountering issues such as low image quality or challenging illumination conditions. In this work, we provide a novel perspective on addressing the problem of gaze estimation from images taken in challenging real-world environments.

Conventional feature-based and model-based gaze estimation typically rely on accurate detection of eye region landmarks, such as the iris center or the eye corners. Many previous works have therefore focused on accurately localizing iris center and eye corners landmarks [Fuhl et al. 2016a; Li et al. 2005; Timm and Barth 2011; Valenti et al. 2012; Wood and Bulling 2014]. On the recent real-world MPIIGaze dataset [Zhang et al. 2015], however, appearance-based methods were shown to significantly outperform model or feature-based methods such as EyeTab, a state-of-the-art model-based method [Wood and Bulling 2014]. The latest appearance-based methods perform particularly well in the person-independent gaze estimation task and in unconstrained settings in which visual artifacts, such as motion blur or sensor noise, are prevalent and cannot be easily removed or modeled [Krafka et al. 2016; Zhang et al. 2015, 2017]. However, appearance based methods are not without drawbacks. First, training data is notoriously expensive and tedious to acquire and even if data is available the quality of the labels can vary. Furthermore, while effective most appearance-based methods are black-box solutions and understanding why and when they work can be a challenge. Finally, deep CNN architectures require a lot of computational power during training and adaptation once fully converged can be difficult and computationally costly.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ETRA ’18, June 14–17, 2018, Warsaw, Poland  
© 2018 Copyright held by the owner/author(s). Publication rights licensed to Association for Computing Machinery.  
ACM ISBN 978-1-4503-5706-7/18/06. . . $15.00  
https://doi.org/10.1145/3204493.3204545
In this work, we propose a new approach to gaze estimation that brings together the best of two-worlds. Similar to appearance-based methods, we leverage deep neural networks for representation learning. While prior work implicitly learns to extract features that are useful for the gaze estimation task, we explicitly learn features that are interpretable. These interpretable features then allow traditional model-based or feature-based approaches to out-perform appearance-based methods in cross-dataset and person-specific gaze estimation. In prior work, features were hand-crafted for gaze estimation using image processing techniques and model fitting. Since such approaches make assumptions about the geometry and shape of eyes, they are sensitive to appearance changes that are prevalent in unconstrained real-world images. Thus such methods suffer from the lack of robust detection of important features in such natural images. We note that the task of eye-region landmark detection bears similarities to the problem of joint detection in human hand and full body pose. Thus, we overcome the above limitation by showing that robust eye-region landmark detectors can be trained solely on high-quality synthetic eye images, providing detailed and accurate labels for the location of important landmarks in the eye region, such as eyelid-sclera border, limbus regions (iris-sclera border), and eye corners. We then show that a relatively compact (in terms of model complexity) state-of-the-art convolutional neural network (CNN) can be trained on such synthetic data to robustly and accurately estimate eye region landmarks in real-world images, without ever providing such images at training time (see Fig. 1). The key advantage of this approach is that model-based and feature-based gaze estimation methods can be applied even to eye images for which iris localization and ellipse fitting can be highly challenging with traditional methods. We experimentally show that such methods perform much better with learned eye landmarks than what has been reported in the literature previously. This allows us to combine the well-studied task of landmark localization with personalized models or parameters for accurate eye tracking for individuals in the real-world.

In summary, the key contributions in our work are: (a) learning of a robust and accurate landmark detector on synthetic images only, (b) improvements to iris localisation and eyelid registration tasks on single eye images in real-world settings, and (c) increased gaze estimation accuracies in cross-dataset evaluations as well as with as few as 10 calibration samples for person-specific cases.

2 RELATED WORK

Our method connects to a wide range of work in gaze estimation. We provide a brief overview in this section specifically in the context of gaze estimation using off-the-shelf cameras.

2.1 Feature-based Gaze Estimation

Feature-based gaze estimation uses geometric considerations to hand-craft feature vectors which map the shape of an eye and other auxiliary information, such as head pose, to estimate gaze direction. Huang et al. [2014b] formulate a feature vector from estimated head pose and distance between 6 landmarks detected on a single eye. A simple and common approach, as introduced by Sesma et al. [2012], is the pupil-center-eye-corner vector or PC-EC vector that has later been adapted and successfully used for estimating horizontal gaze direction on public displays [Zhang et al. 2013, 2014]. The specific claim of Sesma et al. [2012] is that the PC-EC vector can replace the traditional corneal reflections that are traditionally used in eye tracking but that cannot be determined without IR illumination. In addition, methods have been proposed that use Laplacian of Gaussian [Huang et al. 2017], Histogram of Gaussian features [Funes-Mora and Odobez 2016; Huang et al. 2017] or Local Binary Patterns [Huang et al. 2017] among other features.

We show that using rich eye region landmarks detected by our model, when combined with features such as the PC-EC vector, significantly improves gaze estimation accuracy.

2.2 Model-based Gaze Estimation

The eyeball can generally be regarded as two intersecting spheres with deformations. This is exploited in 3D model-based gaze estimation methods in which the center and radius of the eyeball as well as the angular offset between visual and optical axes are determined during user calibration procedures [Sun et al. 2015; Wang and Ji 2017; Wood et al. 2016a; Xiong et al. 2014]. The eyeball center can be determined relative to a facial landmark (such as tip of nose) [Xiong et al. 2014] or by fitting deformable eye region models [Wang and Ji 2017; Wood et al. 2016a]. In contrast, 2D model-based methods can observe the deformation of the circular iris due to perspective [Wang et al. 2003; Wood and Bulling 2014].

While previous works rely on accurate models of the face or eye region, our approach uses a neural network to fit an eyeball to an eye image. This simple approach out-performs all prior works.

2.3 Cross-ratio based Gaze Estimation

In contrast to feature-based and model-based methods, cross-ratio methods utilise just a few IR illumination sources and the detection of their corneal reflections to achieve gaze estimation robust to head pose changes [Yoo et al. 2002]. Recent extensions have improved this approach further via learning from simulation [Huang et al. 2014a] and using multiple viewpoints [Arar and Thiran 2017]. While these methods are promising, additional illumination sources may not be available on unmodified devices or settings for applications such as crowd-sourced saliency estimation using commodity devices.

2.4 Appearance-based Gaze Estimation

While appearance-based gaze estimation is a well-established research area [Baluja and Pomerau 1994; Tan et al. 2002], it has only recently become possible to benchmark gaze estimation methods for in-the-wild settings with MPIIGaze [Zhang et al. 2015] and GazeCapture [Kraikla et al. 2016] datasets. Unlike early works which directly use image intensities as features to variants of linear regression [Funes Mora et al. 2014; Lu et al. 2011a,b] or random forests [Sugano et al. 2014], and k-NN [Wood et al. 2016b], recent works employ complex models such as CNNs [Kraikla et al. 2016; Zhang et al. 2015, 2017, 2018] or GANs [Shrivastava et al. 2017].

Of particular interest are CNN-based approaches that have been demonstrated to yield high accuracy and have benefited from adopting novel architectures such as shown in [Zhang et al. 2018] where a VGG-16 network yielded an improvement of 0.8° (6.3° → 5.5°)
compared to the MnistNet architecture for within-dataset leave-one-person-out evaluation. On the architectural side, other works investigated multi-modal training, such as with head pose information [Zhang et al. 2015], full-face images [Zhang et al. 2017], or an additional “face-grid” modality for direct estimation of point of regard [Krafka et al. 2016]. The use of a lightly modified AlexNet with face images [Zhang et al. 2017] and the drastic improvement in accuracy (6.7° → 4.8°, within-dataset) highlight the need to further study what a CNN learns for the task of gaze estimation.

For the first time, we show that features implicitly learned by CNNs can be used for personalized gaze estimation using as few as 10 calibration samples. We also show that our explicitly learned landmarks features out-perform AlexNet features in this setting.

2.5 Human Pose Estimation and Facial Landmark Localization

The detection of so-called 2D landmarks from images is a well-studied topic in computer vision. In particular, there exists a wealth of research on facial landmark detection and skeletal joint detection for the task of human pose estimation using deep convolutional neural networks [Sun et al. 2013; Toshev and Szegedy 2014]. Facial landmarks and human joint positions can often be occluded but long-range dependencies and global context (i.e., other body parts) can make up for lack of local appearance based information. Thus, recent work has attempted to learn spatial relations between joint positions [Tompson et al. 2014]. Of particular note is the stacked hourglass architecture by Newell et al. [2016]. The stacked multi-scale architecture is simple and has been shown to out-perform other state-of-the-art methods while having low model complexity (few number of parameters). Originally developed for pose estimation, the architecture has been successfully adapted to the task of facial landmark localization in the new Menpo Facial Landmark Localisation Challenge [Yang et al. 2017; Zafeiriou et al. 2017].

3 OVERVIEW

The goal of our work is to provide a robust and accurate landmark detector for eye-gaze estimation and related tasks. For this purpose we leverage a high quality synthetic eye image dataset [Wood et al. 2016b]. Based on recent progress in human pose estimation, we show how this data can be used to train such a detector and apply it directly to real images without fine-tuning or domain adaptation. The extracted landmarks can be directly used in feature-based and model-based gaze estimation yielding improved accuracy in the cross-person and person-specific cases. Thus, we bring such feature-based and model-based approaches back into the forefront of gaze estimation research in unconstrained settings.
In our work we adapt the original architecture to the task of landmark detection in eye-images (see Fig. 2). While eye images contain fewer global structuring elements than in pose estimation, there is still significant spatial context that can be exploited by large receptive field models. We take advantage of this property to detect eyeball center and occluded iris edge landmarks to reasonable accuracy, sometimes even under total occlusion.

The 64 refined feature maps can be combined via a 1 × 1 convolutional layer to produce 18 heatmaps (or confidence maps), each representing the estimated location of a particular eye region landmark. Intermediate supervision is carried out by calculating a pixel-wise sum of squared differences loss on each predicted heatmap. The original paper [Newell et al. 2016] demonstrates that using 8 hourglass modules with intermediate supervision yields significantly improved landmark localization accuracy compared to 2-stack or 4-stack models with the same number of model parameters.

In our work we use only 3 hourglass modules (with 1 residual module per stage), training on eye images and annotations provided by UnityEyes. Though this model consists of less than 1 million model parameters, it is sufficient to demonstrate our approach and allows for a real-time implementation (~ 20Hz). We use single eye images (150 × 90) as input and generate 18 heatmaps (75 × 45): 8 in the limbus region, 8 on the iris edge, 1 at the iris center, and 1 at the eyebrow center. Fig. 2 shows our pipeline at inference time, where the network produces 18 heatmaps which are further processed via a soft-argmax layer [Honari et al. 2018] to find sub-pixel landmark coordinates. These coordinates are then passed on to gaze estimation methods detailed in the following sections.

We find that by applying training data augmentation, a robust model can be learned even when training purely on synthetic eye images. UnityEyes is effectively infinite in size and was designed to exhibit good variations in iris colour, eye region shape, head pose and illumination conditions. Though the appearance variations do not include visual artifacts common in webcam images nor common eye decorations such as eyeglasses or make-up, we later show that our model applies directly to real-world imagery in the tasks of iris localization, eyelid registration, and gaze estimation.

The application of our model to an image from MPIIGaze is shown in Fig. 3 where it can be seen that heatmaps become more accurate and confident as training progresses.

### 4.2 Learning

We now detail the training and data augmentation procedures.

#### 4.2.1 Loss Functions.

The network performs the task of predicting heatmaps, one per eye region landmark. The heatmaps encode the per-pixel confidence on a specific landmark’s location. As such, the highest confidence value is at the pixel nearest to the actual landmark, with confidence quickly dropping off with distance and most of the map containing values set to 0. We place 2-dimensional Gaussians centered at the sub-pixel landmark positions such that the peak value is 1. The neural network then minimizes the $l_2$ distance between the predicted and ground-truth heatmaps per landmark via the following loss term:

$$
\mathcal{L}_{\text{heatmaps}} = \alpha \sum_{i=1}^{18} \sum_{p} \left\| \hat{h}_i(p) - h_i(p) \right\|_2^2,
$$

where $h_i(p)$ is the confidence at pixel $p$ and $\hat{h}_i$ is a heatmap predicted by the network. We empirically set the weight coefficient $\alpha = 1$.

For our model-based method, we additionally predict an eyeball radius value $\hat{r}_{uv}$. This is done by first appending a soft-argmax layer [Honari et al. 2018] to calculate landmark coordinates from heatmaps, then further appending 3 linear fully-connected layers with 100 neurons each (with batch normalization [Ioffe and Szegedy 2015] and ReLU activation) and one final regression layer with 1 neuron. The loss term for the eyeball radius output is:

$$
\mathcal{L}_{\text{radius}} = \beta \left\| \hat{r}_{uv} - r_{uv} \right\|_2^2,
$$

where we set $\beta = 10^{-7}$ and use ground-truth radius $r_{uv}$.

#### 4.2.2 Training Data Augmentation.

We found that employing strong data augmentation during training improves the performance of the model in the context of gaze estimation. We apply the following augmentations (range in brackets are scaling coefficients of value sampled from $N(0, 1)$): translation (2–10 px), rotation (0.1–2.0 rad), intensity (0.5–20.0), blur (0.1–1.0 std. dev. on 7 × 7 Gaussian filter), scale (1.01–1.1), downscale-then-upscale (1x–5x), and addition of lines (0–2) for artificial occlusions. We do not perform any image flipping during training but simply assure that the inner eye corner is on the left side of the input image at test time.

Our final training scheme applies curriculum learning, increasing noise as training progresses [Bengio et al. 2009]. To make this easier to control, we implement a difficulty-measure which ranges from 0 to 1. We begin training with difficulty 0 and linearly increase difficulty until $10^6$ training steps have passed. Thereafter, difficulty is kept at 1. Sample input images are shown in Fig. 4.

We further verify the utility of strong and extensive data augmentation by performing cross-dataset gaze estimation on MPIIGaze with manual eye corner annotations, using our model-fitting method (see Sec. 5.2). Fig. 4c shows a significant decrease in gaze estimation error with higher amounts of training data augmentation (after 1M training steps with batch size of 32). We also observe that with more training steps, a model trained with weaker data augmentation can perform similarly to that trained with stronger...
augmentation. Thus data augmentation not only improves the robustness of the model but speeds up training.

4.2.3 Further Details. We use the ADAM optimizer [Kingma and Ba 2014], with a learning rate of $5 \times 10^{-4}$, batch size of 16, $l_2$-regularization coefficient of $10^{-4}$, and ReLU activation. Our reference model was trained for 6.8M steps on an Nvidia 1080 Ti GPU. During test time, we use statistics computed per-batch for batch normalization, instead of using the population statistics computed during training, from synthetic data.

5 GAZE ESTIMATION

In this section we discuss how we make use of our eye region landmarks in feature-based and model-based gaze estimation.

5.1 Feature-based gaze estimation

To create our features, we first consider the inner and outer eye corners, $c_1$ and $c_2$ respectively. We normalize all detected landmark coordinates by the eye width $c_2 - c_1$, and center the coordinate system on $c_1$. In addition, we provide a 2D gaze prior by subtracting eyeball center $(u_c, v_c)$ from iris center $(u_{ii}, v_{ii})$. Despite being a crude estimate, this prior improves performance significantly where very low number of training samples are available (such as in person-specific calibration). Our final feature vector is formed of 17 normalized coordinates (8 from limbus, 8 from iris edge, 1 from iris center) and a 2D gaze direction prior, resulting in 36 features.

The 36 landmarks-based features are then used to train a support vector regressor (SVR) which directly estimates a gaze direction in 3D, $(\theta, \phi)$ representing eyeball pitch and yaw respectively. The SVR can be trained to be person-independent with a large number of images from different people, or from a small set of person-specific images for personalized gaze estimation. Where possible, we perform leave-one-out cross validation to determine hyper parameters.

5.2 Model-based gaze estimation

As done commonly for remote gaze estimation [Sun et al. 2015; Wood et al. 2015], we use a simple model of the human eyeball, depicting it as one large sphere and a smaller intersecting sphere to represent the corneal bulge (Fig. 5a). Let us denote the predicted 8 iris edge landmarks in a given eye image as $(u_{i1}, v_{i1}), \ldots, (u_{i8}, v_{i8})$. In addition, we detect 1 landmark for the eyeball center $(u_c, v_c)$ and another for the iris center $(u_{ii}, v_{ii})$. Furthermore we estimate

the eyeball radius in pixels, $r_{uv}$. Knowing the eyeball and iris center coordinates and eyeball radius in pixels makes it possible to fit a 3D model without access to any camera intrinsic parameters, and thus without the need for camera calibration.

As we assume no known camera parameters, our coordinates can only be unprojected into 3D space in pixel units. Thus, the radius remains $r_{xy} = r_{uv}$ in 3D space and $(x_c, y_c) = (u_c, v_c)$. If we assume a gaze direction $(\theta, \phi)$, we can now write the iris center coordinates as:

$$
\begin{align*}
    u_{i0} &= x_{i0} = x_c - r_{xy} \cos \theta \sin \phi \\
    v_{i0} &= y_{i0} = y_c + r_{xy} \sin \theta
\end{align*}
$$

To write similar expressions for the 8 iris edge landmarks, we must jointly estimate angular iris radius $\delta$ and an angular offset $\gamma$ which is equivalent to eye roll. This offset between template and actual eye roll or iris rotation is shown in Fig. 5b. With the new variables we can now write for the $j$-th iris edge landmark (with $j = 1 \ldots 8$):

$$
\begin{align*}
    u_{ij} &= x_{ij} = x_c - r_{xy} \cos \theta_j \sin \phi_j \\
    v_{ij} &= y_{ij} = y_c + r_{xy} \sin \theta_j
\end{align*}
$$

where,

$$
\begin{align*}
    \theta_j &= \theta + \delta \sin \left( \frac{\pi}{4} j + \gamma \right) \\
    \phi_j &= \phi + \delta \cos \left( \frac{\pi}{4} j + \gamma \right)
\end{align*}
$$

For model-based gaze estimation, $\theta$, $\phi$, $\gamma$, and $\delta$ are unknown whereas other variables are provided by the eye region landmark localization step of our system. We solve this problem using an iterative optimization method such as the conjugate gradient method where the minimized loss function is represented as:

$$
\sum_{0 \leq j \leq 8} \left( u_{ij} - u_{ij}^* \right)^2 + \left( v_{ij} - v_{ij}^* \right)^2
$$

where $(u_{ij}^*, v_{ij}^*)$ is the estimated pixel coordinates of the $j$-th iris landmark at each iteration.

To adapt this model to a specific person, we calculate person-specific parameters based on calibration samples. Gaze correction can be applied with $(\hat{\theta}, \hat{\phi}) = (\theta + \Delta\hat{\theta}, \phi + \Delta\hat{\phi})$ where $(\Delta\hat{\theta}, \Delta\hat{\phi})$ is the person-specific angular offset between optical and visual axes.

6 EVALUATIONS

We now provide comprehensive evaluations of our method. Since our method can output many more eye landmarks, there is no directly comparable baselines from previous works. Therefore, we first evaluate how well our method can estimate eye-shape by assessing eyelid registration performance, and then address the problem of iris center localization on remote eye images. Finally, we perform various gaze estimation evaluations where we: (a) evaluate the accuracy of our model-fitting approach, (b) evaluate cross-dataset performance of our model-based and feature-based methods against an appearance-based method, and (c) compare feature-based, model-based, and appearance-based approaches for the case of training a gaze estimation model on few calibration samples.

For our evaluations of gaze estimation error, we select the EYE-DIAP [Funes Mora et al. 2014], MPIIGaze [Zhang et al. 2015], UT Multiview [Sugano et al. 2014] and Columbia Gaze [Smith et al.
2013] datasets. These datasets are often used to evaluate gaze estimation methods from remote camera imagery. For the EYEDIAP dataset, we evaluate on VGA images with static head pose for fair comparison with similar evaluations from [Wang and Ji 2017]. Please note that we do not discard the challenging floating target sequences from EYEDIAP. For MPIIGaze, we use manually annotated eye corners to produce segmented eye images to ensure that we can detect eye region landmarks within image bounds. The input eye image dimensions to the hourglass network are $150 \times 90$, unless stated otherwise. Specific challenges represented in the selected datasets include far distance from camera to eye (EYEDIAP), low image quality (MPIIGaze, UT Multiview), and high variations in head poses (UT Multiview) or gaze directions (EYEDIAP).

6.1 Eyelid Registration

![Figure 6: Eyelid registration performance compared against Wood et al. [2015]. Success rate is thresholded on euclidean distance error normalized by interocular distance.](image)

The method proposed by Wood et al. [2015] is the current state-of-the-art for eyelid registration on challenging remote eye images collected in real-world environments. We first perform eyelid registration on the subset of the 300-W dataset [Sagonas et al. 2013] as was tested in [Wood et al. 2015]. The eyelid registration error for a single eye is defined as the mean Euclidean distance to the annotated eyelid, normalized by interocular distance. The ground-truth eyelid annotation is generated from the 6 annotated eyelid landmarks and interocular distance is approximated by the distance between the outer corners of the left and right eyes. Fig. 6 shows our results in comparison to the previous constrained local neural field (CLNF) approach where it can be seen that our method is more accurate and robust. This clearly demonstrates the advantage of the proposed CNN-based eye landmark localization method.

6.2 Iris Localization

Fuhl et al. [2016a] show that pupil detection algorithms developed for the head-mounted case can excel at iris center localization on eye images captured by remotely located cameras. In such scenarios, no glint (from active illumination) can be perceived, and the pupil is often indistinguishable from iris regions. ElSe [Fuhl et al. 2016b] is reported to perform best on BioID, GI4E [Villanueva et al. 2013], and a newly captured dataset exhibiting challenging head poses.

We compare our method, ElSe, and ExCuSe [Fuhl et al. 2015] on the mentioned datasets and report our results in Fig. 7. Input eye images are created based on eye corner annotations such that the input resolution for our method is $150 \times 90$ and for ExCuSe and ElSe are $384 \times 288$ (as done in [Fuhl et al. 2016a]). The figure reports iris localization success rates determined based on distance thresholds to ground-truth, normalized by horizontal eye width. It can be seen that our method consistently out-performs the two state-of-the-art methods across the whole threshold range for all three datasets.

6.3 Model-Based Gaze Estimation

![Figure 7: Iris localization success rate thresholded on euclidean distance error normalized by horizontal eye width. We compare our method with the state-of-the-art ExCuSe and ElSe algorithms which perform well on webcam images.](image)

Our model-fitting algorithm assumes no knowledge about camera intrinsic parameters or accurate 3D models of the face or eye region of specific persons. It rather relies on an estimation of eyeball center, iris center landmarks, and eyeball radius. We compare our model-fitting approach with the other state-of-the-art methods on both Columbia Gaze and EYEDIAP datasets, using 20 images from each person for calibration. For testing on EYEDIAP, we use all available frames to provide a comprehensive evaluation. It can be seen from Tab. 1 that while our results are similar to Wang and Ji [2017] for the Columbia Gaze dataset, we achieve significant improvements on EYEDIAP. Considering that the VGA images from the EYEDIAP dataset are very low-resolution and of low quality, our results further demonstrate the robustness of our approach. In contrast to previous model-fitting approaches which use full face or eye region images as input, we solve the more challenging problem of model-based gaze estimation from single eye images. In this case eye rotation is ambiguous, in that it is not easily possible to define an “up” direction based on the eye shape of an unseen person.

6.4 Feature-Based Gaze Estimation

In Sec. 5.1, we argued for using all landmark coordinates and an iris-center-eyeball-center vector for feature-based gaze estimation. Tab. 2 shows an evaluation of different input features when training an SVR with 20 person-specific samples. In each case, the coordinates or vectors are normalized by eye width, defined as the Euclidean distance between the detected eye corners. It can be seen...
that naive features, such as pupil center or iris landmark coordinates, do not result in good gaze estimation performance. It is only when both eyelid and iris landmarks are used that feature-based gaze estimation can improve significantly, in particular for the UT Multiview dataset that contains large variability in head pose and gaze direction. Performance is further improved by adding our gaze direction prior based on our eyeball center estimations. For subsequent evaluations, we select eight iris edge landmarks, eight eyelid landmarks, one iris center landmark, and one iris-center-eyeball-center vector as features for training our feature-based method.

6.5 Cross-Dataset Gaze Estimation

Table 3: Cross-dataset evaluation of gaze estimation error when trained on UT Multiview (150k entries).

<table>
<thead>
<tr>
<th>Features</th>
<th>EYEDIAP</th>
<th>MPIIGaze+</th>
<th>UT</th>
<th>Columbia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pupil Center</td>
<td>8.4</td>
<td>5.3</td>
<td>17.2</td>
<td>8.0</td>
</tr>
<tr>
<td>PC-EC vectors</td>
<td>8.0</td>
<td>4.9</td>
<td>13.0</td>
<td>7.7</td>
</tr>
<tr>
<td>Iris landmarks</td>
<td>8.3</td>
<td>5.0</td>
<td>17.8</td>
<td>7.5</td>
</tr>
<tr>
<td>Eyelid-Iris landmarks</td>
<td>7.4</td>
<td>4.6</td>
<td>12.4</td>
<td>6.5</td>
</tr>
<tr>
<td>Ours</td>
<td>7.5</td>
<td>4.6</td>
<td>11.5</td>
<td>6.2</td>
</tr>
</tbody>
</table>

6.6 Personalized Gaze Estimation

Figure 8: Mean angular error for personalized gaze estimation where we compare our model-based and feature-based approaches against an appearance-based approach.

While appearance-based gaze estimation has recently demonstrated progress towards person-independent gaze estimation in unconstrained settings, it is an open question whether an appearance-based approach can easily adapt to individual people (i.e., provide better than average performance on a specific user).

One advantage of our approach is that we introduce an eye region landmark localization method that does not require any calibration. The detected landmarks can be directly used to train a simple regression method that in turn is amenable to personalization via few calibration samples. It has been demonstrated that such calibration samples can be collected using a conventional grid of 9-points and that this can be further boosted by recording short video clips [Lu et al. 2011a]. In this evaluation, we explore personalization in this sense for our model-based and feature-based methods.

For a fair comparison to appearance-based methods, we train an AlexNet that directly regresses gaze direction using training data augmentation and UnityEyes images. We then take the output of the 7th layer and train an SVR (SVR_{FC7}) on the 4,096-dimensional feature vector. Thus we directly compare the utility of our explicit, low-dimensional feature representation versus an implicit, high-dimensional feature representation. Note that this is a fair comparison since the last fully connected layer of the AlexNet directly sits under the regression layer during training, hence the learned representation is optimized for the task of gaze estimation.

For MPIIGaze, we select 3,000 samples per person as done in [Zhang et al. 2018] while for EYEDIAP and UT Multiview, we select 1,000 samples per person. We sample our calibration entries from these sets and use the remaining entries for evaluation. For sampling calibration entries, we perform a variance-maximizing sampling based on ground-truth gaze pitch and yaw angles.
We then use a SVR trained on MPIIGaze and UT Multiview datasets.

We demonstrate robustness to challenging illumination conditions, and we implement a real-time proof-of-concept system. The application begins by grabbing a 1280 × 720 RGB frame from a Creative Senz3D webcam (where depth information is not used). We use dlib [King 2009] for face detection and a 5-point facial landmark detection. Two eye images are segmented using the detected eye corners, then each are used as input to our eye landmark localization model. The estimated iris and eyelid landmarks are visualized by blue and yellow outlines respectively in the top-left corner of Fig. 9. We then use a SVR trained on MPIIGaze and UT Multiview datasets to provide an estimation of gaze direction for each eye individually. The unoptimized system runs at up to 26Hz on a desktop PC (Intel i7-4770 and Nvidia GeForce 1080Ti). In the accompanying video, we demonstrate robustness to challenging illumination conditions, occlusion due to eyeglasses, large head pose changes, as well as noise and blur due to the large distance between camera and person.

Our results in Fig. 8 show that the model-fitting approach only benefits initially and only marginally from calibration. We note that our calibration procedure for this case is simple and only estimates the offset between visual and optical axes. However, for both SVR landmarks and SVR FC, we see significant improvements in accuracy with increasing calibration samples. When comparing SVR landmarks and SVR FC, it can be seen that our landmarks-based method performs better across the whole range and accuracies only converge to similar values after more than 100 calibration samples. In particular in the case when there are only a low number of calibration samples ours provides significant accuracy gains. This suggests that personalized gaze estimation is indeed feasible with our feature-based method. Furthermore, even with as few samples as 20, our method improves significantly with 16.4% over the state-of-the-art person-independent gaze estimation error as reported in [Zhang et al. 2018] (5.5° → 4.6°). Finally, SVR landmarks shows improvements of up to 2° in terms of gaze estimation error compared to SVR FC on EYEDIAP and UT Multiview which exhibit large variations in gaze direction and head pose respectively.

6.7 Real-time Implementation

To provide an initial qualitative evaluation on how well our method works we implement a real-time proof-of-concept system. The application begins by grabbing a 1280 × 720 RGB frame from a Creative Senz3D webcam (where depth information is not used). We use dlib [King 2009] for face detection and a 5-point facial landmark detection. Two eye images are segmented using the detected eye corners, then each are used as input to our eye landmark localization model. The estimated iris and eyelid landmarks are visualized by blue and yellow outlines respectively in the top-left corner of Fig. 9. We then use a SVR trained on MPIIGaze and UT Multiview datasets to provide an estimation of gaze direction for each eye individually. The unoptimized system runs at up to 26Hz on a desktop PC (Intel i7-4770 and Nvidia GeForce 1080Ti). In the accompanying video, we demonstrate robustness to challenging illumination conditions, occlusion due to eyeglasses, large head pose changes, as well as noise and blur due to the large distance between camera and person.

7 CONCLUSIONS AND FUTURE WORK

Summary. In this paper we introduced a new approach to gaze estimation from webcam images in unconstrained settings. We showed that eye region landmarks around the iris and eyelid edges can be found in real, unconstrained images using a model trained exclusively on synthetic input. Furthermore, the same neural network can fit an eyeball to the eye image, yielding improvements in gaze estimation. We demonstrate that our method improves upon the state-of-the-art in a number of tasks including eyelid registration and iris localization, in cross-dataset (person-independent) and personalized gaze estimation. We show that both our model-based and feature-based methods out-perform an AlexNet baseline which has significantly more model parameters. We also show that the compact model can be used for robust real-time gaze estimation. For the case of personalization we show that our landmarks-based SVR prevails especially in the case of few calibration samples.

Directions for future work. In our work, we provided initial insights on how explicitly learned landmark-based features can outperform implicitly learned features in certain cases. However, more research is necessary to understand which representation yields the most accurate and robust gaze estimation method. We showed that our neural network is capable of detecting irises and eyelids to a higher accuracy than prior work. However, direct evaluation of eye region landmark localization on current real-world datasets is challenging due to the absence of high-quality labels. We experimentally showed encouraging performance gains for feature-based and model-based methods when leveraging our learned landmarks-based representation such as in the most challenging cross-dataset evaluation, contradicting previous comparisons [Zhang et al. 2018]. This implies that more research into feature-based and model-based methods could further improve gaze-estimation when coupled with powerful feature representations.

One particularly interesting result presented here is the capability to personalize deep-learning based gaze estimation methods, with our explicit landmark-based feature performing very well for low number of calibration samples. In future work this insight could be coupled with an appropriate online calibration procedure to produce highly accurate gaze estimates even in situations where only a user-facing camera is available and under difficult environmental conditions. However, much future research is necessary to fully understand the best ways of garnering and utilizing calibration samples while improving gaze estimation accuracy significantly. The hope is that our work will eventually enable conducting experiments at large scale by leveraging the commonly available cameras on devices such as smartphones, tablets, and laptops only.

ACKNOWLEDGEMENTS

We would like to thank Erroll Wood, Tadas Baltrusaitis, and Wolfgang Fuhl for their help. This work was supported in part by ERC Grant OPTINT (StG-2016-717054), the Cluster of Excellence on Multimodal Computing and Interaction at Saarland University, Germany, and by a JST CREST research grant (JPMJCR14E1), Japan.

1https://youtu.be/cLUIHYZ5N4s


