Appearance-Based Gaze Estimation in the Wild

Xucong Zhang[1] Yusuke Sugano[1] Mario Fritz[2] Andreas Bulling[1]
[1]Perceptual User Interfaces Group, [2]Scalable Learning and Perception Group
Max Planck Institute for Informatics, Saarbrücken, Germany
{xczhang,sugano,mfritz,bulling}@mpi-inf.mpg.de

Abstract

Appearance-based gaze estimation is believed to work well in real-world settings, but existing datasets have been collected under controlled laboratory conditions and methods have been not evaluated across multiple datasets. In this work we study appearance-based gaze estimation in the wild. We present the MPIIGaze dataset that contains 213,659 images we collected from 15 participants during natural everyday laptop use over more than three months. Our dataset is significantly more variable than existing ones with respect to appearance and illumination. We also present a method for in-the-wild appearance-based gaze estimation using multimodal convolutional neural networks that significantly outperforms state-of-the art methods in the most challenging cross-dataset evaluation. We present an extensive evaluation of several state-of-the-art image-based gaze estimation algorithms on three current datasets, including our own. This evaluation provides clear insights and allows us to identify key research challenges of gaze estimation in the wild.

1. Introduction

Appearance-based gaze estimation is well established as a research topic in computer vision because of its relevance for several application domains, including gaze-based human-computer interaction and visual behaviour analysis [31]. Purely learning-based methods were recently proposed to learn generic gaze estimators from large amounts of person, and head pose-independent training data [10, 34, 39]. Such methods have the potential to bring appearance-based methods into settings that do not require any user- or device-specific training. Gaze estimation using monocular cameras is particularly promising given the proliferation of such cameras in hand-held and portable devices, such as mobile phones and laptops, as well as interactive displays.

While appearance-based gaze estimation is believed to perform well in everyday settings, state-of-the-art learning-based methods are still developed and evaluated on datasets collected under controlled laboratory conditions. These conditions are characterised by limited variability of eye appearances as well as the assumption of accurate head pose estimates. Current appearance-based gaze estimation methods are also not evaluated across different datasets, which bears the risk of significant dataset bias – a key problem also in object recognition [43] and salient object detection [23].

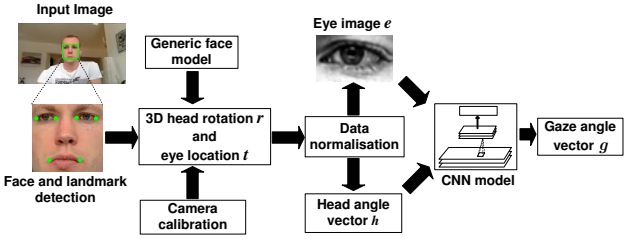


Figure 1: Overview of our method for in-the-wild appearance-based gaze estimation using multimodal convolutional neural networks.

In this work we make the first step towards appearance-based gaze estimation in the wild. Given a lack of realistic data, we created the MPIIGaze dataset that contains 213,659 images collected from 15 laptop users over several months (see Figure 2). MPIIGaze covers a realistic variability in appearance and illumination and therefore represents a significant advance over existing datasets. Laptops not only allow us to record in the wild but they also have considerable potential as an application platform, such as for gaze interaction [28]. The dataset and annotations are publicly available online.

We study two key tasks through extensive evaluations of appearance-based gaze estimation algorithms on three publicly available gaze estimation datasets:

1. Handling appearance differences between training and testing data. Since we cannot always assume a training dataset that can cover the whole test space, the important question is how robustly the estimator can handle unknown appearance conditions.

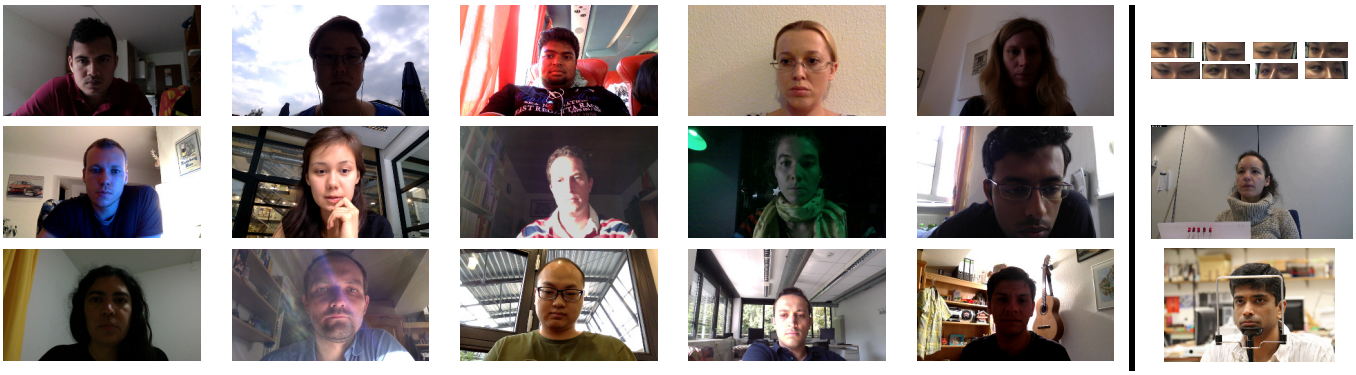


Figure 2: Sample images from our MPIIGaze dataset showing the considerable variability in terms of place and time of recording, directional light and shadows. For comparison, the last column shows sample images from other current publicly available datasets (cf. Table 1): UT Multiview [39] (top), Eyediap [8] (middle), Smith et al. [37] (bottom).

2. Pursuing the performance gain for domain-specific training. If we can assume that training data is directly collected in the target daily-life environment, the goal is to fully utilise the rich training data.

While better performances can be expected for the second domain-specific training task where both the training and testing data come from the same dataset, the ultimate goal of person-independent gaze estimation is to handle the first cross-domain training task, which leads to the most challenging but practically most important use cases.

The contribution of this work is threefold. First, we introduce the first large-scale dataset for appearance-based gaze estimation in the wild. Our dataset is one order of magnitude larger than existing datasets and significantly more variable with respect to illumination and appearance. Second, we present an extensive evaluation of state-of-the-art gaze estimation algorithms on three current datasets, including our own, and identify key research challenges of in-the-wild settings. Third, we present a method for appearance-based gaze estimation that uses multimodal convolutional neural networks and that significantly outperforms state-of-the-art methods in the most challenging cross-dataset evaluation.

2. Related Work

2.1. Gaze Estimation Methods

Gaze estimation methods can be model-based or appearance-based [12]. Model-based methods use a geometric eye model and can be further divided into corneal-reflection and shape-based methods, depending on whether they require external light sources to detect eye features. Early works on corneal reflection-based methods focused on stationary settings [36, 30, 13, 51] and were later extended to handle arbitrary head poses using multiple light sources or cameras [52, 53]. In contrast, shape-based methods [16, 4, 50, 44] directly infer gaze directions from observed eye shapes, such as pupil centre or iris edges. Although they have recently been applied to more practical application scenarios [18, 11, 41, 49], their accuracy is lower and it is unclear whether shape-based approaches can robustly handle low image quality and variable lighting conditions. Appearance-based gaze estimation methods directly use eye images as input and can therefore potentially work with low-resolution eye images. While early works assumed a fixed head pose [3, 42, 48, 35, 27, 24], recent works focused on methods for 3D head pose estimation [25, 26, 9, 6]. However, appearance-based methods require larger amounts of user-specific training data than model-based methods, and it remains unclear if the learned estimator can generalise to unknown users. Similarly, previous methods typically assumed accurate 3D head poses as input, which is a strong assumption for unconstrained in-the-wild settings.

2.2. Calibration-Free Gaze Estimation

The requirement to collect person-specific training data during a calibration step is a key limitation of both model-based and appearance-based methods. To address this limitation, several previous works used interaction events, such as mouse clicks or key presses, as a proxy for the user's on-screen gaze position [40, 15]. Alternatively, visual saliency maps [5, 38] or pre-recorded human gaze patterns of the presented visual stimuli [1] were used as bottom-up, probabilistic training data to learn the estimation function. However, all of these approaches rely on observations of a specific person and environment, which limits their applicability.

Purely data-driven approaches leverage large amounts of training data to learn gaze estimators that generalise to arbitrary users without the need for person-specific calibration [34, 10, 39] settings. These methods have significant potential to bring gaze estimation to new settings, includ-

	Participants	Head poses	On-screen gaze targets	Illumination conditions	Duration (days)	Images
McMurrough et al. [29]	20	1	16	1	1	videos
Villaneuva et al. [45]	103	1	12	1	1	1,236
Weidenbacher et al. [46]	20	19	2-9	1	1	2,220
Smith et al. [37]	56	5	21	1	1	5,880
Eyediap [8]	16	continuous	continuous	2	1	videos
UT Multiview [39]	50	8 + synthesised	160	1	1	64,000
MPIIGaze (ours)	15	continuous	continuous	daily life	45.7	213,659

Table 1: Comparison of current publicly available appearance-based gaze estimation datasets with respect to number of participants, head poses and on-screen gaze targets (discrete or continuous), number of different illumination conditions, average duration of data collection per participant, and total number of images.

ing mobile devices, public displays, and egocentric cameras. However, the generalization capability of learning-based methods has not been examined yet. Moreover, prior work used 3D input for head pose information [10, 39], while we are the first to evaluate the whole pipeline for fully automatic monocular appearance-based gaze estimation for person-independent training scenario.

2.3. Datasets

Because most existing gaze estimation datasets are designed for coarse gaze estimation, the sampling density of gaze and head pose space is not sufficient to train appearance-based gaze estimators [29, 45, 46, 37] (see Table 1 for an overview of existing datasets). More comparable to MPIIGaze, the Eyediap dataset contains 94 video sequences of 16 participants looking at three different targets (discrete and continuous markers displayed on a monitor, and floating physical targets) under both static and free head motion [8]. The UT Multiview dataset also contains dense gaze samples of 50 participants as well as 3D reconstructions of eye regions that can be used to synthesise images for arbitrary head poses [39]. However, as discussed before, both datasets have the significant limitation that they were recorded under controlled laboratory settings. Although the Eyediap dataset includes two different illumination conditions, recordings under the second condition were provided only for a subset of the participants.

3. The MPIIGaze dataset

We designed our data collection procedure with two main objectives in mind: 1) to record images of participants outside of controlled laboratory conditions, i.e during their daily routine, and 2) to record participants over several months to cover a wider range of recording locations and times, illuminations, and eye appearances. We opted for recording images on laptops not only because they are suited for long-term daily recordings but also because they are an important platform for eye tracking applications [28]. Laptops are personal devices, therefore typically remaining with a single user, and they are used throughout the day and over long periods of time. They also come with high resolution front-facing cameras that are in a fixed position relative to the screen. We further opted to use an experience sampling approach to ensure images were collected regularly throughout the data collection period [19].

3.1. Collection Procedure

We implemented custom software running as a background service on participants' laptops. Every 10 minutes the software automatically asked participants to look at a random sequence of 20 on-screen positions (a recording session), visualised as a grey circle shrinking in size and with a white dot in the middle. Participants were asked to fixate on these dots and confirm each by pressing the spacebar once the circle was about to disappear. This was to ensure participants concentrated on the task and fixated exactly at the intended on-screen positions. No other instructions were given to them, in particular no constraints as to how and where to use their laptops. Because our dataset covers different laptop models with varying screen size and resolution, on-screen gaze positions were converted to physical 3D positions in a camera coordinate system. We obtained the intrinsic parameters from each camera beforehand. 3D positions of each screen plane were estimated using a mirror-based calibration method [33].

We also asked human annotators to provide face annotations for a random subset of 10,848 images to increase the value of the dataset for other tasks, such as face detection and alignment. They annotated these images with a total of 12 facial landmarks, following an extended LFW style [14], that additionally contained a face bounding box and two eye bounding boxes, as well as the left and right pupil position.

3.2. Dataset Characteristics

We collected a total of 213,659 images from 15 participants. The number of images collected by each participant varied from 34,745 to 1,498. Figure 3 (left) shows a histogram of times of the recording sessions. Although there is a certain bias towards working hours, the figure shows the high variation in recording times. Consequently, our

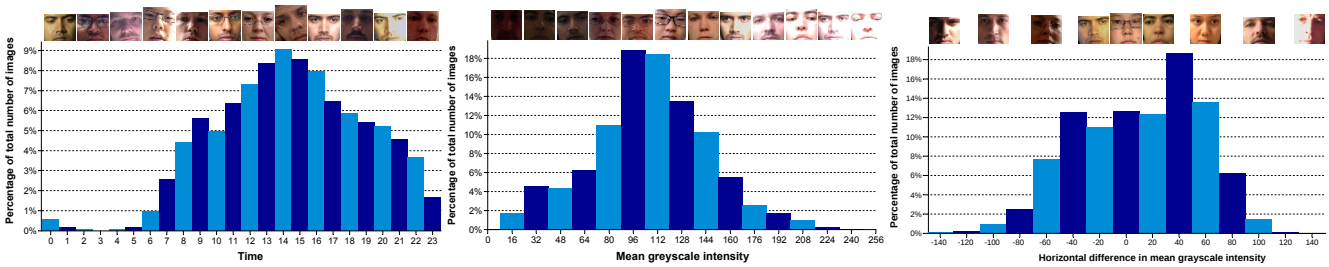
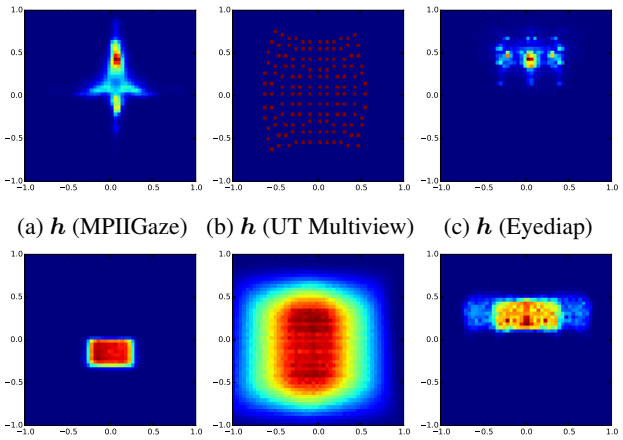


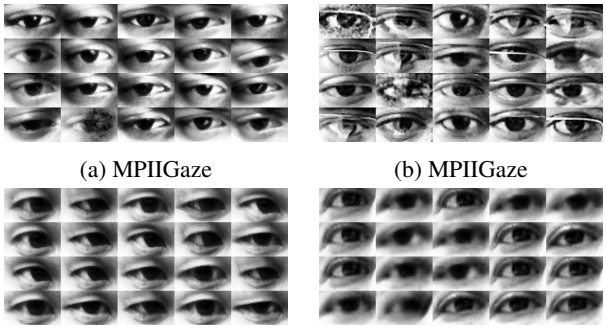
Figure 3: Key characteristics of our dataset. Percentage of images collected at different times over the day (left), having different mean grey-scale intensities within the face region (middle), and having horizontally different mean grey-scale intensities from the left to right half of the face region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



(a) $\boldsymbol{h}$ (MPIIGaze) (b) $\boldsymbol{h}$ (UT Multiview) (c) $\boldsymbol{h}$ (Eyediap)

(d) $\boldsymbol{g}$ (MPIIGaze) (e) $\boldsymbol{g}$ (UT Multiview) (f) $\boldsymbol{g}$ (Eyediap)

Figure 4: Distributions of head angle ($\boldsymbol{h}$) and gaze angle ($\boldsymbol{g}$) for the MPIIGaze, UT Multiview, and Eyediap datasets (cf. Table 1).



(a) MPIIGaze (b) MPIIGaze

(c) UT Multiview (d) Eyediap

Figure 5: Example images from the MPIIGaze (non-eyeglasses and wearing eyeglasses), UT Multiview, Eyediap datasets.

dataset also contains larger variability in illumination. To visualise the different illumination conditions, Figure 3 (bottom) shows a histogram of mean grey-scale intensities inside the face region. Figure 3 (right) further shows a histogram of the mean intensity differences from the right side to the left side of the face region, which approximates the statistics of directional light sources. These figures underline the complexity of our dataset in terms of appearance variations.

To further characterise our MPIIGaze dataset in comparison with the other recent datasets [8, 39], Figure 4 summarises distributions of the head and gaze angles $\boldsymbol{h}, \boldsymbol{g}$ in the normalised space. The normalisation was done as described in Section 4.2. Each figure corresponds to a 2D histogram of either pose or gaze, colour-coded from blue (minimum) to red (maximum). Although the UT Multiview dataset (see Figures 4b and 4e) is recorded under a controlled lighting condition, it contains synthesised eye images which largely cover both gaze and pose spaces. Although the Eyediap dataset has mainly two different gaze targets, Figures 4c and 4f show distributions of 2D screen targets, which is closer to our setting. Our MPIIGaze dataset covers a 2D screen space as in the Eyediap dataset; however, the gaze angle distributions are not overlapping, due to the difference in camera positions (see Figures 4a and 4d). This indicates that the Eyediap dataset does not cover the range of gaze directions that can occur during laptop interactions and that our MPIIGaze can serve as a more appropriate basis for training and testing gaze estimators.

Figure 5 shows sample eye images from each dataset after the normalisation (see Section 4.2). Each group of images was randomly selected from a single person for roughly the same gaze directions. Compared to the UT Multiview and Eyediap datasets (see Figures 5c and 5d), our MPIIGaze dataset contains larger appearance variations even inside the eye region (see Figure 5a). The variation becomes more significant in the case of a person wearing eyeglasses (see Figure 5b), and they depict the complexity of the daily-life setting in terms of appearance changes.

4. Method

Figure 1 provides an overview of our proposed method for in-the-wild appearance-based gaze estimation using multimodal convolutional neural networks (CNN). We first employ state-of-the-art face detection and facial landmark detection methods to locate landmarks in the input image obtained from the calibrated monocular RGB camera. We then fit a generic 3D facial shape model to estimate 3D poses of the detected faces and apply the space normalisation technique proposed in [39] to crop and warp the head pose and eye images to the normalised training space. The CNN is used to learn the mapping from the head poses and eye images to gaze directions in the camera coordinate system.

4.1. Face Alignment and 3D Head Pose Estimation

Our method first detects the user's face in the image using Li et al.'s SURF cascade method [22]. We assume a single face in the images and take the largest bounding box if the detector returns multiple face proposals. We discard all images in which the detector fails to find any face, which happened in about 5% of all cases. Afterwards, we use Baltrušaitis et al.'s constrained local mode framework to detect facial landmarks [2].

We use the same definition of the face model and head coordinate system as [39]. The face model consists of 3D positions of six facial landmarks (eye and mouth corners, cf. Figure 1). The head coordinate system is defined according to the triangle connecting three midpoints of the eyes and mouth. We fit the model by estimating the initial solution using the EPnP algorithm [21], and further refining the pose via non-linear optimisation. 3D head rotation $\boldsymbol{r}$ is defined as the rotation from the head coordinate system to the camera coordinate system, and the eye position $\boldsymbol{t}$ is defined as the midpoint of eye corners for each eye.

While previous works assumed accurate head poses, we use a generic mean facial shape model for the 3D pose estimation to evaluate the whole gaze estimation pipeline in a practical setting. 3D positions of the six landmarks are recorded from all of the participants using an external stereo camera prior to the data collection, and the generic shape is built as the mean shape across all participants.

4.2. Data Normalisation

Similar to [39], we normalise the image and head pose space into a polar-coordinate angle space. Fundamentally speaking, object pose has six degrees of freedom, and in the simplest case the gaze estimator has to handle eye appearance changes in this 6D space. However, since arbitrary scaling and rotation of the camera can be compensated for by its corresponding perspective image warping, the appearance variation that needs to be handled inside the appearance-based estimation function has only two degrees of freedom. The task of pose-independent appearance-based gaze estimation is to learn the mapping between gaze directions and eye appearances, which cannot be compensated for by virtually rotating and scaling the camera.

Briefly, the normalisation is done by scaling and rotating the camera so that: 1) the camera looks at the midpoint of the eye corners from a fixed distance d, and 2) x axes of the head coordinate system and camera coordinate system become parallel. Eye images were cropped at a fixed resolution $W \times H$ with a fixed focal length f in the normalised camera space, and histogram-equalised to form the input eye image. This results in a set of fixed-resolution eye images $\boldsymbol{e}$ and 2D head angle vectors $\boldsymbol{h}$, and the ground-truth gaze positions are also converted to the normalised camera space to give 2D gaze angle (yaw and pitch) vectors $\boldsymbol{g}$. In order to reduce the effect of different lighting conditions, eye images $\boldsymbol{e}$ are histogram-equalised after the normalisation process. We used the same setting for camera distance d, focal length f and the resolution $W \times H$ as in [39]. In this manner, the normalised eye images are compatible between different datasets and we can evaluate the cross-dataset performance of appearance-based methods.

4.3. Gaze Estimation With Multimodal CNNs

The task for the CNN is to learn the mapping from the input features (2D head angle $\boldsymbol{h}$ and eye image $\boldsymbol{e}$) to gaze angles $\boldsymbol{g}$ in the normalised space. As pointed out in [39], the difference between the left and right eyes is irrelevant in the person-independent training scenario. By flipping eye images horizontally and mirroring $\boldsymbol{h}$ and $\boldsymbol{g}$ around the y axis, we handle both eyes by a single regression function.

Our model uses the LeNet network architecture that consists of one convolutional layer followed by a max-pooling layer, a second convolution layer followed by a max-pooling layer, and a final fully connected layer [20, 17]. We train a linear regression layer on top of the fully connected layer to predict gaze angle vectors $\boldsymbol{g}$. We use a multimodal CNN model to take advantage of both eye image and head pose information [32]. We encode head pose information into our CNN model by concatenating $\boldsymbol{h}$ with the output of the fully connected layer (see Figure 6). Input to the network are the grey-scale eye images $\boldsymbol{e}$ with a fixed size of 60×36 pixels. For the two convolutional layers, the feature size is 5×5 pixels, while the number of features is 20 for the first layer and 50 for the second layer. The number of hidden units in the fully connected layer is 500, where each unit connects to all the feature maps of the previous convolutional layer, and is calculated by summing up all activation values. The output of the network is a 2D gaze angle vector $\hat{\boldsymbol{g}}$ that consists of two gaze angles, yaw $\hat{g}_\phi$ and pitch $\hat{g}_\theta$. As a loss function we use the sum of the individual L_2 losses that measure the distance between the predicted $\hat{\boldsymbol{g}}$ and actual gaze angle vectors $\boldsymbol{g}$.

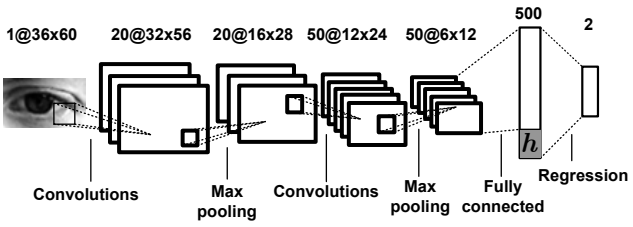


Figure 6: Architecture of the proposed multimodal CNN. Head angle vectors $\boldsymbol{h}$ are added to the output of the fully connected layer.

5. Experiments

In this section, we discuss the person-independent gaze estimation task and validate the effectiveness of the proposed CNN-based gaze estimation approach. We conduct both cross-dataset and within-dataset conditions to compare our method with state-of-the-art methods on the MPIIGaze dataset. To account for the sample number bias among participants in our dataset, in the following experiments we use a random subset for both training and testing. Specifically, we randomly pick 1,500 left eye samples and 1,500 right eye samples from each person[1].

In addition to our CNN-based method, we evaluate the following baseline methods using the same facial landmark detection, head pose estimation, and input features.

Random Forests (RF) Random forests were recently demonstrated to outperform existing methods for person-independent appearance-based gaze estimation [39]. We use the implementation provided by the authors, which first clusters training samples according to head angles and query test samples to their nearest clusters. We used the same parameters as in [39], and also resized input eye images to 16×9 pixels.

k-Nearest Neighbours (kNN) As shown in [39], a simple kNN regression estimator can perform well in scenarios that offer a large amount of dense training samples. We use the same kNN implementation and also incorporate a training sample clustering in head angle space.

Adaptive Linear Regression (ALR) Because it was originally designed for a person-specific and sparse set of training samples [27], ALR does not scale to large datasets. We therefore use the same approximation as in [10], i.e. we select five training persons for each test person by evaluating the interpolation weights. We further select random subsets of samples from the test sample's neighbours in head pose space. We use the same image resolution as for RF.

[1]Since one participant has only 1,448 images, we randomly oversampled the data to get 3,000.

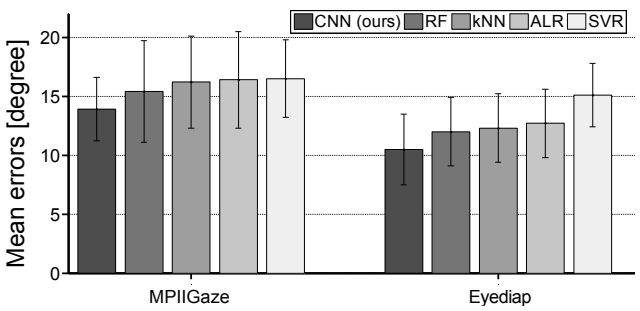


Figure 7: Cross-dataset evaluation with training data from the UT Multiview dataset. Bars correspond to mean error across all participants in the MPIIGaze (left) and screen-target sequences of Eyediap (right) datasets. Error bars indicate standard deviations.

Support Vector Regression (SVR) Schneider et al. [34] used SVR with a polynomial kernel under a fixed head pose. We use a linear SVR [7] given the large amount of training data. We also use a concatenated vector of HOG and LBP features (6×4 blocks, 2×2 cells for HOG) as suggested in [34]. However, we do not use manifold alignment, since it does not support pose-independent training.

Shape-Based Approach (EyeTab) Finally, in addition to these appearance-based methods, we evaluate one state-of-the-art shape-based method [49] on the MPIIGaze dataset. We use the implementation provided by the authors. In their method gaze estimation is performed by fitting a limbus model (a fixed-diameter disc) to detected iris edges.

5.1. Cross-Dataset Evaluation

We first present a comparative experimental validation for the cross-dataset evaluation condition. We selected the UT Multiview dataset as the training dataset because it covers the largest area in head and gaze angle space (see Figure 4). In addition to our MPIIGaze dataset, we also show results using the Eyediap dataset as test data.

For the Eyediap dataset we used the 3D head poses provided with the dataset. These were estimated by fitting personal 3D shape models to depth images [8]. Since their floating target sequences contain many extreme gaze directions that are not covered by the UT Multiview dataset, we only used the screen target sequences.

Figure 7 summarises mean angular errors of all methods on both MPIIGaze and Eyediap. Bars correspond to mean error across all participants in each dataset, and error bars indicate standard deviations across persons. The mean prediction error of a naive predictor that always outputs the average gaze direction of all training samples is 42.4 degrees on Eyediap and 34.2 degrees on MPIIGaze. The shape-based EyeTab method performs poorly on the MPIIGaze dataset (47.1 degrees mean error and 7% mis-

detection rate), and this supports the advantage of the appearance-based approaches in challenging conditions. In this setting, our CNN-based approach shows the best accuracy on both datasets (13.9 degrees on MPIIGaze, 10.5 degrees on Eyediap), with a significant performance gain (10% on MPIIGaze, 12% on Eyediap, paired Wilcoxon test [47], $p < 0.05$) over the state-of-the-art RF method. However, performance on MPIIGaze is generally worse than on the Eyediap dataset, which indicates the fundamental difficulty of the in-the-wild setting.

While our CNN-based approach expanded the feasibility of the generalisation task, these results at the same time reveal the critical limitation of the UT Multiview dataset and their learning-by-synthesis approach, whose variation of training data is limited in terms of eye appearances. This indicates the importance of the training data, and that we need to address this goal from the both standpoints of data and methodology to bridge the gap from the domain-restricted training scenario.

5.2. Within-Dataset Evaluation

Although the previous cross-dataset evaluation showed the advantage of our CNN-based gaze estimation approach, there is still a huge performance gap compared to the performance reported in [39]. To discuss the limits of person-independent performance on the MPIIGaze dataset, we performed leave-one-person-out evaluation on the MPIIGaze dataset.

With the same baseline methods as in Section 5.1, Figure 8 shows mean angular errors of the within-dataset evaluation. Since the model-based EyeTab method has been shown to perform relatively poorly in our setting, we alternatively show a learning-based result using the detected pupil (iris centre) positions. More specifically, we used the pupil positions detected using [49] in the normalised eye image space as a feature for kNN regression, and performed the same leave-one-person-out test.

In this case there is domain-specific prior knowledge about gaze distribution, and the mean prediction error becomes 13.9 degrees. The pupil position-based approach works better than the original EyeTab method but its performance is still worse than appearance-based gaze estimation methods. All appearance-based methods showed better performances than in Section 5.1, and this indicates the importance of dataset- or domain-specific training data for appearance-based gaze estimation methods. Although its performance gain over the other baseline methods becomes smaller in this setting, our CNN-based method still performed the best among them with 6.3 degrees mean error.

In order to illustrate the difference on handling appearance variations between cross-dataset and within-dataset scenarios, Figure 9 shows estimation errors with respect to different illumination conditions. Similarly to Figure 3, we

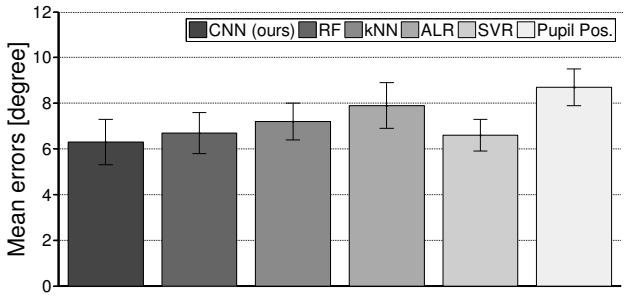


Figure 8: Within-dataset leave-one-person-out evaluation on MPIIGaze. Mean estimation errors of the proposed method and other appearance-based methods. Error bars indicate standard deviations.

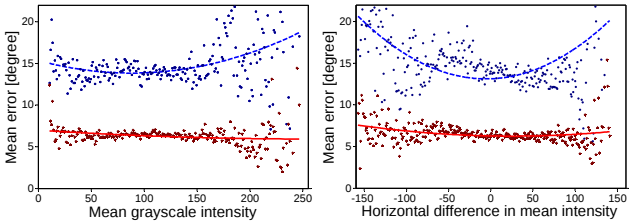


Figure 9: Estimation error distribution w.r.t. mean grey-scale intensity within the face region (left) and horizontal mean grey-scale intensity difference between the left and right half of the face region (right). The blue dots and curve from cross-dataset evaluation and the red dots and curve from within-MPIIGaze evaluation.

evaluate the error distribution with respect to mean grey-scale intensity of the face region and horizontal mean intensity difference between the left and right face regions. Compared to the model pre-trained on the UT Multiview dataset (blue dots and curve), the model trained on the MPIIGaze dataset (red dots and curve) shows better performance across different lighting conditions. This clearly illustrates the effect of different lighting conditions and the importance of the appearance variation in the training data.

5.3. Performance Validation of the Multimodal CNN

While previous results show the potential of appearance-based gaze estimation methods in a challenging daily-life condition, there still exists a large performance gap compared to person-specific training results reported in prior work. To further discuss the performance limits of the CNN-based approach, we also show more detailed comparisons between RF and CNN models.

We first show a comparison between different architectures of the CNN on the UT Multiview dataset with the same three-fold cross-validation setting as reported in [39]

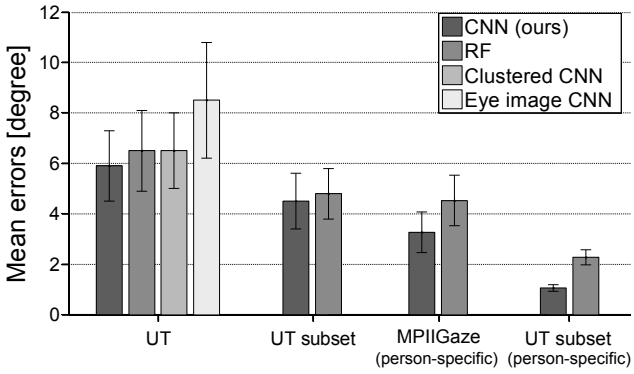


Figure 10: Comparison of the different CNN models and RF on (from left to right): UT Multiview dataset, subset of the UT Multiview dataset which has the same head and gaze angle ranges as the MPIIGaze dataset, using person-specific training on the MPIIGaze dataset, and using person-specific training on the UT Multiview subset. Error bars indicate standard deviations.

(see Figure 10 left). As can be seen, our proposed multimodal CNN model outperformed the RF method with 5.9 degrees mean error. Although [39] reported that their pose-clustered structure improved their RF performance, the performance of the CNN became worse if the same clustering structure was introduced. This indicates the higher learning flexibility of the CNN, which contributes to the large performance gain in the cross-dataset case (Section 5.1). The performance can be degraded further if there is no head pose input, and this shows the fundamental importance of the head pose information in this pose-independent gaze estimation task.

The performance within the UT Multiview dataset is almost in the same range as the performance within the MPIIGaze dataset (see Figure 8). However, these two cases are expected to have different difficulty levels. To investigate the difference within these results in more detail, we further show a three fold evaluation using a subset (3,000 samples per person) of the UT Multiview dataset selected so as to have the same pose and gaze angle distributions as the MPIIGaze dataset. The result is shown in the next part of Figure 10, and the performance gap compared to Figure 8 indicates the error that arises from the in-the-wild setting, including appearance variations and eye alignment errors.

Although this is not always a practical assumption, estimators trained on person-specific training data show the upper limit of the performance we can achieve. The rest of Figure 10 shows mean errors of person-specific models on both the MPIIGaze and UT Multiview datasets. For MPIIGaze, the last quarter of the data from each person was used as test data, and the rest of the data was used as training data. For UT Multiview, 500 test samples were randomly selected for each person from the above subset, and the other 2,500 samples were used as training data. These results further show the potential performance of the appearance-based estimator, and clearly depict the performance gap to be investigated.

6. Conclusion

Despite a large body of previous work on the topic, appearance-based gaze estimation methods have so far been evaluated exclusively under controlled laboratory conditions. In this work, we presented the first extensive study on appearance-based gaze estimation in the unconstrained daily-life setting. We built a novel in-the-wild gaze dataset through a long-term data collection using laptops, which shows significantly larger variations in eye appearance than existing datasets. Throughout the comprehensive benchmarking of image-based monocular gaze estimation methods, our study clearly revealed the potential and remaining technical challenges of appearance-based gaze estimation. Our CNN-based estimation model significantly outperforms state-of-the-art methods in the most challenging person- and pose-independent training scenario. This work and our dataset provide a critical insight on addressing grand challenges in daily-life gaze interaction.

Acknowledgements

We would like to thank Laura Sesma for her help with the dataset handling and normalisation. This work was funded in part by the Cluster of Excellence on Multimodal Computing and Interaction (MMCI) at Saarland University as well as an Alexander von Humboldt Research Fellowship. We would further like to thank the NVIDIA Corporation for donating the GPU used in this research.

References

[1] F. Alnajar, T. Gevers, R. Valenti, and S. Ghebreab. Calibration-free gaze estimation using human gaze patterns. In *Proc. ICCV*, 2013. 2

[2] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Continuous conditional neural fields for structured regression. In *Proc. ECCV*, pages 593–608, 2014. 5

[3] S. Baluja and D. Pomerleau. Non-intrusive gaze tracking using artificial neural networks. Technical report, DTIC Document, 1994. 2

[4] J. Chen and Q. Ji. 3d gaze estimation with a single camera without ir illumination. In *Proc. ICPR*, pages 1–4, 2008. 2

[5] J. Chen and Q. Ji. Probabilistic gaze estimation without active personal calibration. In *Proc. CVPR*, pages 609–616, 2011. 2

[6] J. Choi, B. Ahn, J. Parl, and I. S. Kweon. Appearance-based gaze estimation using kinect. In *Proc. URAI*, pages 260–261, 2013. 2

[7] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008. 6

[8] K. A. Funes Mora, F. Monay, and J.-M. Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proc. ETRA*, pages 255–258, 2014. 2, 3, 4, 6

[9] K. A. Funes Mora and J.-M. Odobez. Gaze estimation from multimodal kinect data. In *Proc. CVPRW*, pages 25–30, 2012. 2

[10] K. A. Funes Mora and J.-M. Odobez. Person independent 3d gaze estimation from remote rgb-d cameras. In *Proc. ICIP*, 2013. 1, 2, 3, 6

[11] K. A. Funes Mora and J.-M. Odobez. Geometric generative gaze estimation (g3e) for remote rgb-d cameras. In *Proc. CVPR*, pages 1773–1780, 2014. 2

[12] D. W. Hansen and Q. Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):478–500, 2010. 2

[13] C. Hennessey, B. Noureddin, and P. Lawrence. A single camera eye-gaze tracking system with free head motion. In *Proc. ETRA*, pages 87–94, 2006. 2

[14] G. B. Huang, M. Mattar, T. Berg, E. Learned-Miller, et al. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition*, 2008. 3

[15] M. X. Huang, T. C. Kwok, G. Ngai, H. V. Leong, and S. C. Chan. Building a self-learning eye gaze model from user interaction data. In *Proc. MM*, pages 1017–1020, 2014. 2

[16] T. Ishikawa, S. Baker, I. Matthews, and T. Kanade. Passive driver gaze tracking with active appearance models. In *Proc. 11th World Congress on Intelligent Transportation Systems*, 2004. 2

[17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 5

[18] L. Jianfeng and L. Shigang. Eye-model-based gaze estimation by rgb-d camera. In *Proc. CVPRW*, pages 606–610, 2014. 2

[19] R. Larson and M. Csikszentmihalyi. The experience sampling method. *New Directions for Methodology of Social & Behavioral Science*, 1983. 3

[20] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 5

[21] V. Lepetit, F. Moreno-Noguer, and P. Fua. EPnP: An accurate o(n) solution to the PnP problem. *International Journal of Computer Vision*, 81(2):155–166, 2009. 5

[22] J. Li and Y. Zhang. Learning surf cascade for fast and accurate object detection. In *Proc. CVPR*, pages 3468–3475, 2013. 5

[23] Y. Li, X. Hou, C. Koch, J. Rehg, and A. Yuille. The secrets of salient object segmentation. In *Proc. CVPR*, 2014. 1

[24] K. Liang, Y. Chahir, M. Molina, C. Tijus, and F. Jouen. Appearance-based gaze tracking with spectral clustering and semi-supervised gaussian process regression. In *Proc. ETSA*, pages 17–23, 2013. 2

[25] F. Lu, T. Okabe, Y. Sugano, and Y. Sato. Learning gaze biases with head motion for head pose-free gaze estimation. *Image and Vision Computing*, 32(3):169 – 179, 2014. 2

[26] F. Lu, Y. Sugano, T. Okabe, and Y. Sato. Head pose-free appearance-based gaze sensing via eye image synthesis. In *Proc. ICPR*, pages 1008–1011, 2012. 2

[27] F. Lu, Y. Sugano, T. Okabe, and Y. Sato. Adaptive linear regression for appearance-based gaze estimation. *IEEE Trans. PAMI*, 36(10):2033–2046, Oct 2014. 2, 6

[28] P. Majaranta and A. Bulling. Eye tracking and eye-based human–computer interaction. In *Advances in Physiological Computing*, pages 39–65. Springer, 2014. 1, 3

[29] C. D. McMurrough, V. Metsis, J. Rich, and F. Makedon. An eye tracking dataset for point of gaze detection. In *Proc. ETRA*, pages 305–308, 2012. 3

[30] C. H. Morimoto, A. Amir, and M. Flickner. Detecting eye position and gaze from a single camera and 2 light sources. In *Proc. ICPR*, pages 314–317, 2002. 2

[31] C. H. Morimoto and M. R. Mimica. Eye gaze tracking techniques for interactive applications. *Computer Vision and Image Understanding*, 98(1):4–24, 2005. 1

[32] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng. Multimodal deep learning. In *Proc. ICML*, pages 689–696, 2011. 5

[33] R. Rodrigues, J. a. Barreto, and U. Nunes. Camera pose estimation using images of planar mirror reflections. In *Proc. ECCV*, pages 382–395, 2010. 3

[34] T. Schneider, B. Schauerte, and R. Stiefelhagen. Manifold alignment for person independent appearance-based gaze estimation. In *Proc. ICPR*, 2014. 1, 2, 6

[35] W. Sewell and O. Komogortsev. Real-time eye gaze tracking with an unmodified commodity webcam employing a neural network. In *Ext. Abstracts CHI*, pages 3739–3744, 2010. 2

[36] S.-W. Shih and J. Liu. A novel approach to 3-d gaze tracking using stereo cameras. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 34(1):234–245, 2004. 2

[37] B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar. Gaze locking: passive eye contact detection for human-object interaction. In *Proc. UIST*, pages 271–280, 2013. 2, 3

[38] Y. Sugano, Y. Matsushita, and Y. Sato. Appearance-based gaze estimation using visual saliency. *IEEE Trans. on PAMI*, 35(2):329–341, Feb 2013. 2

[39] Y. Sugano, Y. Matsushita, and Y. Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *Proc. CVPR*, pages 1821–1828, 2014. 1, 2, 3, 4, 5, 6, 7, 8

[40] Y. Sugano, Y. Matsushita, Y. Sato, and H. Koike. An incremental learning method for unconstrained gaze estimation. In *Proc. ECCV*, pages 656–667, 2008. 2

[41] L. Sun, M. Song, Z. Liu, and M.-T. Sun. Realtime gaze estimation with online calibration. In *Proc. ICME*, pages 1–6, 2014. 2

[42] K.-H. Tan, D. J. Kriegman, and N. Ahuja. Appearance-based eye gaze estimation. In *Proc. WACV*, pages 191–195, 2002. 2

[43] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *Proc. CVPR*, pages 1521–1528. IEEE, 2011. 1

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

[45] A. Villanueva, V. Ponz, L. Sesma-Sanchez, M. Ariz, S. Porta, and R. Cabeza. Hybrid method based on topography for robust detection of iris center and eye corners. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 9(4):25, 2013. 3

[46] U. Weidenbacher, G. Layher, P.-M. Strauss, and H. Neumann. A comprehensive head pose and gaze database. In *Proc. IET*, pages 455–458, 2007. 3

[47] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, pages 80–83, 1945. 7

[48] O. Williams, A. Blake, and R. Cipolla. Sparse and semi-supervised visual mapping with the Sˆ 3GP. In *Proc. CVPR*, pages 230–237, 2006. 2

[49] E. Wood and A. Bulling. Eyetab: Model-based gaze estimation on unmodified tablet computers. In *Proc. ETRA*, pages 207–210, 2014. 2, 6, 7

[50] H. Yamazoe, A. Utsumi, T. Yonezawa, and S. Abe. Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions. In *Proc. ETRA*, pages 245–250, 2008. 2

[51] D. H. Yoo and M. J. Chung. A novel non-intrusive eye gaze estimation using cross-ratio under large head motion. *Computer Vision and Image Understanding*, 98(1):25–51, 2005. 2

[52] Z. Zhu and Q. Ji. Eye gaze tracking under natural head movements. In *Proc. CVPR*, pages 918–923, 2005. 2

[53] Z. Zhu, Q. Ji, and K. P. Bennett. Nonlinear eye gaze mapping function estimation via support vector regression. In *Proc. ICPR*, pages 1132–1135, 2006. 2