A Differential Approach for Gaze Estimation with Calibration

Gang Liu[1]
gang.liu@idiap.ch
Yu Yu[1]
yu.yu@idiap.ch
Kenneth A. Funes-Mora [2]
kenneth@eyeware.tech
Jean-Marc Odobez[1]
odobez@idiap.ch

[1] Idiap Research Institute
[2] Eyeware Tech SA

Abstract

Gaze estimation methods usually regress gaze directions directly from a single face or eye image. However, due to important variabilities in eye shapes and inner eye structures amongst individuals, universal models obtain limited accuracies and their output usually exhibit high variance as well as biases which are subject dependent. Therefore, increasing accuracy is usually done through calibration, allowing gaze predictions for a subject to be mapped to his/her specific gaze. In this paper, we introduce a novel image differential method for gaze estimation. We propose to directly train a convolutional neural network to predict the gaze differences between two eye input images of the same subject. Then, given a set of subject specific calibration images, we can use the inferred differences to predict the gaze direction of a novel eye sample. The assumption is that by allowing the comparison between two eye images, annoyance factors (alignment, eyelid closing, illumination perturbations) which usually plague single image prediction methods can be much reduced, allowing better prediction altogether. Experiments on 3 public datasets validate our approach which constantly outperforms state-of-the-art methods even when followed by subject specific gaze adaptation.

1 Introduction

As a non-verbal behavior and major indicator of human attention, gaze is an important communication cue which has also been shown to be related with higher-level characteristics such as personality and mental state. It thus finds applications in many domains like Human-Robot-Interaction (HRI) [2, 16], Virtual Reality [17], social interaction analysis [10], or health care [27]. With the development of sensing function on mobile phones, gaze is also expected to be involved in a wider set of application in mobile scenarios [9, 11, 23].

Related works. Non-invasive vision based gaze estimation has been addressed using two main paradigms [8]: geometric models, and appearance. Geometric approaches rely on eye feature extraction (like glints when working with infrared systems, eye corners or iris center localization) to learn a geometric model of the eye and then infer gaze direction using these features and model [1, 22, 24, 29, 30, 31]. However, they usually require high resolution eye images for robust and accurate feature extraction, are prone to noise or illumination, and do not handle well head pose variabilities and medium to large head poses. Thus, many recent methods rely on an appearance based paradigm, directly predicting gaze from an eye (or face) image input [5, 20, 35, 37], allowing them to be robust when dealing with low to mid-resolution images and to obtain good generalization performance. Amongst them, deep neural networks (NN) have been shown to work well. They leverage on large amount of data to train a regression network capturing the essential features of the eye images under various conditions like illumination and self-shadow, glasses, impact of head pose. For instance, [35] relied on simple LeNet type of shallow network applied to eye images and first demonstrated that NNs outperform most other appearance based methods. Krafka e*t.al* [11] proposed to combine eyes and faces information together using a multi-channel network. Zhang *et.al* [33] trained a weighted network to predict gaze from a full face image. Shrivastava *et.al* [18] learned a model from simulated eye images using a generative adversarial network.

Motivation. Nevertheless, even when using deep Neural Network (NN) regressors, the accuracy of appearance-based method has been limited to around 5 to 6 degrees, with a high inter person variance [5, 18, 20, 33, 35, 37]. This is due to many factors including dependencies on head poses, large eye shape variabilities, and only very subtle eye appearance changes when looking at targets separated by such small angle differences. For instance, one factor that can explain why appearance based methods encounter limited accuracy when building person independent models is that the visual axis is not aligned with the optical axis (related to the observed iris) [7], and that such alignment differences are subject specific. Thus, in theory, images of two eyes with the same appearance but with different internal eyeball structure can correspond to different gaze directions, demonstrating that gaze can not be fully predicted from the visual appearance.

A straightforward solution to this problem is to learn person-specific models which can achieve far better accuracy [35]. However, training person-specific appearance models may require large amounts of personal data, especially for network based methods and even when conducting simple network fine tuning adaptation. This is not practical in real life applications. To solve this problem, Lu *et.al* [13] proposed an adaptive linear regression method relying on few training samples, but the eye representation (multi-grid normalized mean eye image) is not robust to environmental changes. Starting from a trained NN, Krafka *et.al.* [11] relied on eye images when looking at a grid of 13 dot sample. Feature maps from the last layer of the pretrained NN were then employed to train a Support-Vector-Regression (SVR) person specific gaze prediction model. However, SVR regression from a high dimensional feature vector input is not robust to noise. In another direction, Zhang *et.al.* [36] proposed to train person-specific gaze estimators from user interactions with multiple devices, such as mobile phone, tablet, laptop, or smart TVs, but this does not correspond to the majority of use cases.

Contributions. In this paper, we first propose a simpler method than the above for adaptation. The method learns the linear relationship between the gaze predictions from a pretrained NN applied to few training samples and their groundtruth gaze, and is shown to achieve better results than the state-of-the art SVR method of [11].

Secondly, although the above methods can reduce the subject specific bias between the subject (test) data and the overall training dataset, it does this by only working with the gaze prediction or feature outputs, and does not account for the high gaze prediction variance within each subject's data. To address this issue, our main contribution is to propose a differ-

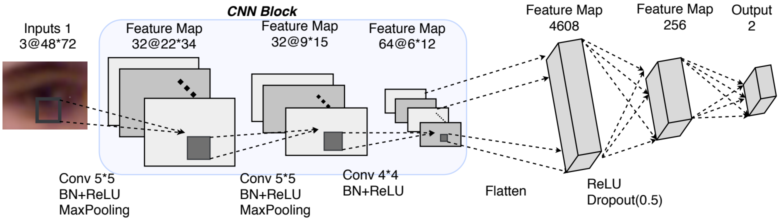


Figure 1: Baseline CNN structure for gaze estimation.

ential gaze estimation approach, by training a differential NN to predict the gaze difference between two eye images instead of predicting the gaze directly. At training time, a unified and person independent differential gaze prediction model is built which can be used at test time for person specific gaze inference relying on only a few calibration samples.

The closest work to ours is Venturelli *et.al.* [26]. However, they are addressing a different task (head pose estimation). Furthermore, inspired by the works on face identification, they trained a siamese network with two distinct depth images as input, but this was done within a multi-task approach in which both absolute head poses and head pose differences were used as loss function. Hence, at test time, the pose is directly predicted from only one the parallel structure. And furthermore, while several layers of our differential networks are used to predict the gaze difference, in their case the pose differences was only computed from the network pose prediction output.

Paper organization. First in Sec. 2, we introduce the state-of-the-art NN methods for gaze prediction. We illustrate the subject specific bias problem and propose a linear adaptation method to build subject specific gaze prediction models. In Section 3, we introduce our approach, including the proposed differential NN for differential gaze prediction. Experiments are presented in Sec. 4, while Sec. 5 concludes the work.

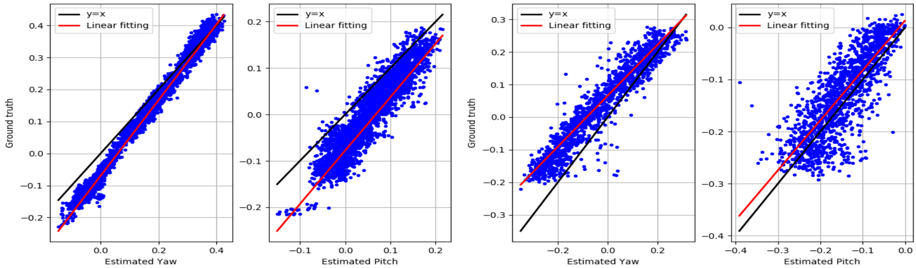
2 Baseline CNN approach and linear adaptation

In this section, we first introduce a standard convolution neural network (CNN) for person independent gaze estimation. We then show the data bias existing between the training set and the test data of individuals and present our proposed linear adaptation method.

2.1 Gaze estimation with CNN

Network structure. Fig. 1 presents the standard NN structure for gaze estimation. It consists of three convolutional layers and two fully connected layers[1]. More precisely, the input eye image $I \in R^{M \times N \times C}$, where $(M,N,C) = (48,72,3)$ denote the dimensions and number of channels of the image, is first whitened. The convolutional layers are then applied and the resulting feature maps are flattened to be fed into the fully-connected layers. The predicted gaze direction $\mathbf{g}^p(I) \in R^{2\times 1}$ is regressed at the last layer. The details of the network

[1]Note that it is slightly different from [34].



(a) Eyediap (b) MPIIGaze

Figure 2: Scatter plot of the network regression (X-axis) and labeled groundtruth (Y-axis) of the yaw (left plot) and pitch (right plot) angles for an indiviual eye taken in the (a) Eyediap dataset; and (b) MPIIGaze dataset.

parameters can be found in the figure.

Loss function. Denoting the gaze groundtruth of an eye image I by $\mathbf{g}^{gt}(I)$, we used the following L1 loss function to train our baseline NN:

$$\mathcal{L} = \frac{1}{|\mathcal{D}|} \sum_{I \in \mathcal{D}} \|\mathbf{g}^{p}(I) - \mathbf{g}^{gt}(I)\|_1, \tag{1}$$

where $\mathcal{D}$ denotes the training dataset and $|\cdot|$ denotes the cardinality operator.

Network training. The network is optimized with Adam method, with a learning rate initially set to 0.001 and then divided by 2 after each epoch. In our experiment, 10 epochs are applied and proved to be sufficient. The mini batch size is 128.

2.2 Bias analysis and linear adaptation

Because each individual eye has specific characteristics (including internal non-visible dimensions or structures), in practice, we often observe a data bias between the network regression $\mathbf{g}^{p}(I)$ and the labeled groundtruth $\mathbf{g}^{gt}(I)$ of the eye images $I \in \mathcal{D}_{Test}$ belonging to a single person. This is illustrated in Fig. 2, which provides a scatter plot of the $(\mathbf{g}^{p}(I), \mathbf{g}^{gt}(I))$ angle pairs in typical cases, which can be compared with the identity mapping (black lines).

As can be observed, there is usually a linear relationship between $\mathbf{g}^{gt}(I)$ and $\mathbf{g}^{p}(I)$, which is illustrated by the red lines in the plots. Thus, when a set $\mathcal{D}_c$ of sample calibration points of a user (usually 9 to 25 points) is available, we propose to learn this relation and obtain an adapted gaze model $\mathbf{g}^{ad}$ by fitting a linear model

$$\mathbf{g}^{ad}(I) = A\,\mathbf{g}^{p}(I) + B \tag{2}$$

where $A \in R^{2\times 2}$ and $B \in R^{1\times 2}$ are the linear parameters of the model which can be estimated through least mean square error (LMSE) optimisation using the calibration data.

3 Proposed differential approach

Approach overview. Fig. 3 presents our proposed framework. Its main part is a differential network designed and trained to predict the differences in gaze direction between two images

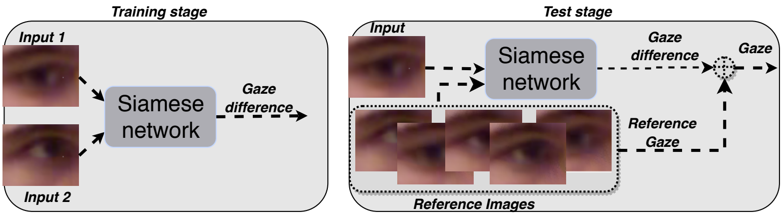


Figure 3: Approach overview. During training, random pairs of samples from the same eye are used to train a differential network. A test time, given a set of reference samples, gaze differences are computed and used to infer the gaze of the input image.

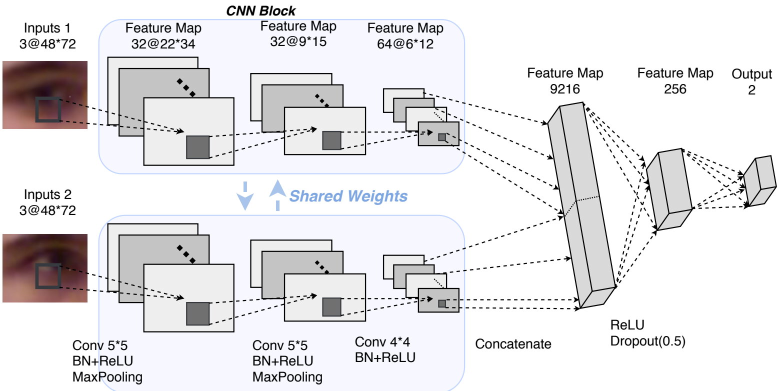


Figure 4: The designed differential network for predicting gaze differences.

of the same eye. At test time, the gaze differences between the input eye image and a set of reference images are computed first. Then the gaze of the eye image is estimated by adding these gaze differences to the reference gazes. The details of the different components are introduced in the following paragraphs.

Differential network architecture. Differential networks have been first proposed in [4] for signature verification using image matching. Following the deep learning revival, they have been widely considered for tasks like feature extraction [12, 19, 32], image matching and retrieval [14, 28], person re-identification [15, 25], or object tracking [3]. They usually consist of two parallel networks with shared weights, in which a pair of distinct images is used as input, one for each parallel channel, and the distance between the outputs of each parallel network is computed as differential network output. Implicitly, when dealing with discrete category problems, the goal of such differential networks is to learn (usually using a hinge-loss function) a mapping from the image space to a new feature space such that samples from the same class are close, while samples from different classes are far. In the regression case (our case), the loss function is usually defined by comparing the output distance with the labelled groundtruth.

The network we use is illustrated in Fig. 4, and is slightly modified from the traditional siamese approach. Each branch in the parallel structure is composed of three convolutional neural layers, all of them followed by batch normalization and ReLU units. Max pooling is applied after the first and second layers for reducing the image dimensions. After the third layer, the feature maps of the two input images are flatted and concatenated into a new tensor. Then two fully-connected layers are applied on the tensor to predict the gaze difference between the two input images. Thus, where traditional siamese approaches would predict the gaze for each image, and compute the differences from these predictions, our approach uses neural network layers to predict this difference from an intermediate eye feature representation.

Loss function. The differential network is trained using a set of random image pairs (I, J) coming from the same eye in the training data. Denoting by $\mathbf{d}^p(I, J)$ the gaze difference predicted by the differential network, we can define the loss function as:

$$\mathcal{L} = \sum_{(I,J)\in\mathcal{D}^k\times\mathcal{D}^k} \|\mathbf{d}^p(I,J) - \left(\mathbf{g}^{gt}(I) - \mathbf{g}^{gt}(J)\right)\|_1, \tag{3}$$

where $\mathcal{D}^k$ is the subset of $\mathcal{D}$ that only contains images of the same eye[2] of person k.

Network training. The network is optimized with the Adam method, with an initial learning rate of 0.001 which is divided by 2 after each epoch. In experiments, 20 epochs are applied. The mini batch size is 128. Note that as the number of possible image pairs is too large, we have reduced it by using each of the image $I \in \mathcal{D}^k$ as the first image of a pair, and randomly selecting the second image $J \in \mathcal{D}^k$ of the pair. So we have $|\mathcal{D}^k|$ pairs for the subject k.

Gaze inference. As the network predicts gaze differences only, the method requires at least one reference image to predict an absolute gaze vector. In practice, we rely on a small calibration set $\mathcal{D}_c$ of images of the same eye. We then predict the gaze difference between the test image I and the reference images F, and combining these gaze difference with the gaze groundtruth of the reference images, we can infer the gaze direction of the test image. More formally, we have:

$$\mathbf{g}^{sm}(I) = \frac{1}{|\mathcal{D}_c|} \sum_{F\in\mathcal{D}_c} \left(\mathbf{g}^{gt}(F) + \mathbf{d}^p(I,F)\right). \tag{4}$$

4 Experimental results and analysis

4.1 Datasets

We validated our algorithm on three public datasets.

Eyediap. This dataset contains 94 videos associated with 16 subjects [6]. Videos belong to three categories: continuous screen (CS) target, discrete screen (DS) target or floating target (FT). The CS videos were used in our experiments, which comprises static pose recordings where subjects approximately maintain the same pose while looking at targets, and dynamic poses in which subjects perform additional important head movements while looking. From this data, we cropped around $80K$ images of the left and right eyes and frontalized them according to [5]. The labelled world gaze groundtruth was converted accordingly.

MPIIGaze. This dataset [35] contains 1500 left and right eye images of 15 subjects, which were recorded under various conditions in head pose or illuminations and contains people with glasses. The provided images are approximately of size 36×60 pixels, and are already frontalized relying on the head pose yaw and pitch. Note that although in [35] the head pose

[2]Note that we learn a differential model for the left eye, and one for the right eye.

is used as input for gaze prediction, this did not improve our results in experiments so it was not used for the experiments reported below.

UT-Multiview. This dataset [21] comprises 23040 (1280 real and 21760 synthesized) left and right eye samples for each of the 50 subjects (15 female and 35 male). It was collected under strict laboratory control condition, with various head pose. Importantly, eye images are not frontalized. Thus, in experiments, we concatenated the head pose in the network as described in [35]. More precisely, we concatenated the head pose $\mathbf{h}(I) \in R^{1\times 2}$ of the input I image with the last fully-connected layers for the baseline CNN (Fig. 1), and did the same for the differential network, i.e. we concatenated the two head pose $\mathbf{h}(I)$ and $\mathbf{h}(J)$ of the differential input pair (I,J) with the last fully-connected layer in Fig. 4.

4.2 Experimental protocol

Cross-Validation. For the Eyediap and MPIIGaze datasets, we applied a leave-one-subject-out protocol, while due to its size, we used a 3-fold cross-validation protocol for the UT-Multiview dataset. Note that for this dataset, we train with real and synthesis data, but only test on real data. Note that these protocols for MPIIGaze and UT-Multiview are the standard ones used in the original paper and followed by other researchers.

Performance measure. We trained and tested models for the left and right eyes separately, as we noticed that the left and right eyes may have different structures, and importantly, the labeled gaze might follow different distributions.

Following the above protocols, the error was defined as the average of the average gaze angular error computed for each fold. More precisely, if $\mathcal{D}_{Test}$ denotes the test data (for a single subject) of a given fold, the trained model for that fold is evaluated by computing:

$$\mathcal{E}(\mathcal{D}_{Test}) = \frac{1}{|\mathcal{D}_{Test}|} \sum_{I\in\mathcal{D}_{Test}} \arccos\left(\bar{v}(\mathbf{g}^{p}(I)) \cdot \bar{v}\left(\mathbf{g}^{gt}(I)\right)\right), \tag{5}$$

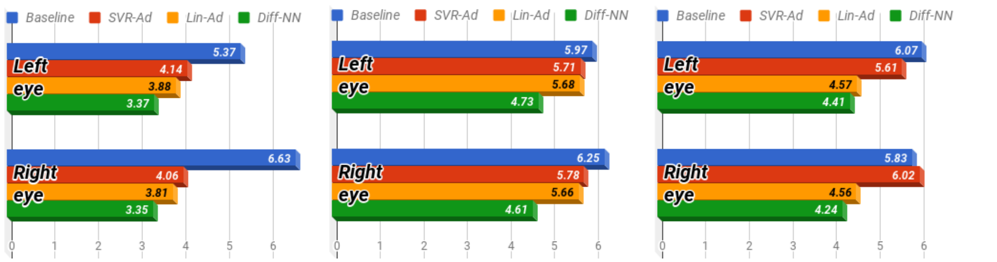
where $\bar{v}(\theta_1,\theta_2)$ denotes the unitary 3D gaze vector associated with the gaze angles (θ_1,θ_2). Note that for the linear adaptation and the differential NN methods, reference images are required to predict the gaze for the given subject. In this case, we randomly selected 9 points in the test set $\mathcal{D}_{Test}$ for 200 times, and reported the average error computed for each random selection as defined above.

Tested models. Several methods were tested for comparison. ***Baseline*** corresponds to the generic model introduced in Section 2. ***Lin-Ad*** corresponds to the *Baseline* model followed by linear adaptation (Section 2.2). ***SVR-Ad*** is our implementation of the SVR adaptation method of [11] built upon the *Baseline* model above. ***Diff-NN*** is the method we propose.

4.3 Experimental results

The experimental results are presented in Fig. 5, in which the left, mid and right plots are the results on Eyediap, MPIIGaze and UT-multiview datasets. In each sub-figure, the upper bars indicate the results for the left eye, and the bottom ones for the right eye. The colors correspond to the different approaches: *Baseline* (blue), *Lin-Ad* (orange), *SVR-Ad* [11] (red), and our *Diff-NN* proposed method (green).

Baseline model. First, let us note that under the same protocol, our *Baseline* model works slightly better than [35], which reported an error of 6.3° on MPIIGaze, and of 5.9° on UT-



(a) Eyediap (b) MPIIGaze (c) UT-multiview

Figure 5: Average angular error on three public datasets. Note that the *Baseline* method does not require calibration data.

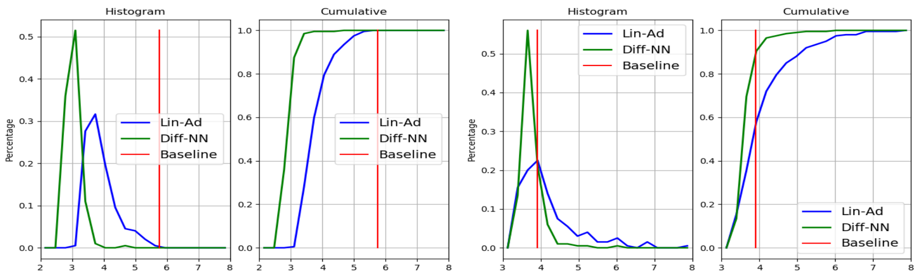
Multiview, compared to 6.11° and 5.95° on average in our case. This is probably due to our network architecture being slightly more complex, while still avoiding over-fitting.

Linear and SVR Adaptation. Results demonstrate that, as expected, calibration helps and that our linear adaption method can greatly improve the results. The improvements are for the left and right eyes: 27.7% and 43.3% on Eyediap, 24.7% and 21.8% on UT-Multiview, and 4.9% and 9.4% on MPIIGaze. The difference in gain is most probably due to the recording protocols. While the Eyediap and UT-Multiview datasets were mainly recorded over the course of one session, the MPIIGaze dataset was collected in the wild, over a much longer period of time, and with much more lighting variability (but less head pose variability). This can be observed in Fig. 2 showing typical scattering plots of the Eyediap and MPIIGaze datasets. The Eyediap plots follow a more straight and compact linear relationship than those on the MPIIGaze dataset, reflecting the higher variability within the last dataset. Seen differently, we can interpret the results as having a session-based adaptation in the Eyediap and UT-Multiview cases, whereas in MPIIGaze, the adaptation is more subject-based.

Results also show that our linear adaptation *Lin-Ad* method is working better than the *SVR-Ad* adaption approach [], with an average gain of 6.3%, 1.4% and 21.5% on the Eyediap, MPIIGaze, and UT-Multiview datasets, respectively. The main reason might be that in *SVR-Ad*, the regression weights after the last fully-connected (FC) layer are not exploited, in spite of their importance regarding the gaze prediction. In addition, finding an appropriate kernel in the 256 dimensional space of the last FC output might not be so easy, and 9 points might not be sufficient for regression within such a space.

Differential method. Our approach *Diff-NN* performs much better than the other two adaptation methods which, on average over the 3 datasets, have an error 14.0% (*Lin-Ad*) and 26.8% (*SVR-Ad*) higher than ours. In particular, we can note that the gain is particularly important on the MPIIGaze dataset (21.4% compared to *Lin-Ad*), demonstrating that our strategy of directly predicting the gaze differences from pairs of images -hence allowing to implicitly match and compare these images- using our modified Siamese network is more powerful, and more robust against eye appearance variations across time, places, or illumination, than adaptation methods relying on gaze predictions only (*Lin-Ad*), or on compact eye image representations (*SVR-Ad*). On other more 'session-based' datasets, our linear adaptation method is already doing well, so that the gain is lower (around 10% on average).

Calibration data variability. The performance of the adaptation methods are computed as the average over 200 random selection of 9 calibration samples. Depending on the selection (samples might be noisy, or not distributed well on the gaze grid), results may differ. Fig. 6



(a) MPIIGaze (b) Eyediap

Figure 6: Distribution and cumulative distribution of angular errors due to random selection of the calibration images, for two subjects (one from the MPIIGaze dataset, one from the Eyediap dataset), and for different methods: *Diff-NN* (green curve), *Lin-Ad* (blue curve), *Baseline* (red; note that this method does not rely on calibration data).

illustrates the variabilities of *Lin-Ad* and *Diff-NN* for the different trials of two subjects.

The example on the left shows a typical example where there is a relatively large bias for the given subject. In that case, whatever the selection of the calibration samples, the results of both *Lin-Ad* and *Diff-NN* are better than the baseline. The example on the right shows one of the few cases where the baseline is already good, with little bias but nevertheless quite noisy samples. In that case, there is only around 60% chances to obtain a better result with the linear adaptation, but still over 80% chances with our approach. Also, importantly, our *Diff-NN* approach is less sensitive to the choice of calibration points than *Lin-Ad*, as can be seen from the slope of the cumulative curves (steeper for *Diff-NN*).

4.4 Algorithm complexity

The two adaptation methods do not have the same complexity. Compared to the CNN *Baseline*, the linear adaptation only requires the computation of Eq.2, which has negligable computational cost. Our *Diff-NN* approach, however, requires to predict the gaze differences between the test sample and N_c reference images, so that we could think the complexity being of the order of N_c that of *Baseline*. Fortunately, thanks to our differential architecture (see Fig. 4), the extra-computation is not as high. Indeed, first we can pre-compute and save the feature maps at the last convolutional neural layer of all the reference images, so that the computation of one gaze difference requires mainly the forward pass of one image. Secondly, the feature maps of the test image also need to be computed only once, which can be achieved by stacking the feature maps of the reference images in a mini-batch, and compute all gaze differences in parallel.

Tab. 1 compares the running time (in ms) for the *Baseline* and the different *Diff-NN* options (and $N_c = 9$ as in reported experiments). They have been obtained by computing the average run-time of processing 5000 images. The CPU is an Intel(R) Core(TM) i7-5930K with 6 kernels and 3.50GHz per kernel. The GPU is an Nvidia Tesla K40. The program is written in Python and Pytorch. Note that as Pytorch library will call multiple kernels for the computation, the CPU-based run-time is also short. From this Table, we can see that our *Diff-NN* method and architecture has a computational complexity close to the baseline.

Table 1: Run-times (in ms) between the *Baseline* and our proposed *Diff-NN* method, using mini-batch (*Diff-NN* *) computation or not.

	CPU			GPU		
	Baseline	*Diff-NN*	*Diff-NN* *	*Baseline*	*Diff-NN*	*Diff-NN* *
Run-time	2.5	7.6	3.5	1.4	4.0	1.5

5 Conclusion

This paper aims to improve appearance-based gaze estimation using subject specific models built from few calibration images. Our main contributions are to propose (1) a linear adaptation method based on these reference images; (2) a differential NN for predicting gaze differences instead of gaze direction to alleviate the impact of annoyance factors like illumination, cropping variability, variabilities in eye shapes. Experimental results on three public and commonly used datasets prove the efficacy of the proposed methods. More precisely, while at very little extra computation cost the linear adaptation method can already boost the results on single session like situations, the differential NN method produces even more robust and stable results across different sessions of the same user, but costs some more run-time compared to a baseline CNN.

Acknowledgement

The current work was co-funded by the Innosuisse, the Swiss Innovation agency, through the REGENN (Robust Eye-Gaze Estimation Deep Network) grant 26041.1.

References

[1] Kenneth Alberto, Funes Mora, Jean-marc Odobez, and De Lausanne. 3D Gaze Tracking and Automatic Gaze Coding from RGB-D Cameras. *IEEE Conference in Computer Vision and Pattern Recognition, Vision Meets Cognition Workshop*, pages 4321–4322, 2014.

[2] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. Conversational gaze aversion for humanlike robots. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction*, HRI '14, pages 25–32, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2658-2. doi: 10.1145/2559636.2559666. URL http://doi.acm.org/10.1145/2559636.2559666.

[3] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016.

[4] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. In *Advances in Neural Information Processing Systems*, pages 737–744, 1994.

[5] Kenneth A. Funes-Mora and Jean-Marc Odobez. Gaze estimation in the 3d space using rgb-d sensors. *International Journal of Computer Vision*, 118(2):194–216, 2016. ISSN 1573-1405.

[6] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 255–258. ACM, 2014.

[7] Elias Daniel Guestrin and Moshe Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on biomedical engineering*, 53(6):1124–1133, 2006.

[8] Dan Witzner Hansen and Qiang 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.

[9] Qiong Huang, Ashok Veeraraghavan, and Ashutosh Sabharwal. Tabletgaze: unconstrained appearance-based gaze estimation in mobile tablets. *arXiv preprint arXiv:1508.01244*, 2015.

[10] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. Prediction of who will be the next speaker and when using gaze behavior in multiparty meetings. *ACM Trans. Interact. Intell. Syst.*, 6(1):4:1–4:31, May 2016. ISSN 2160-6455. doi: 10.1145/2757284. URL http://doi.acm.org/10.1145/2757284.

[11] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, and Harini Kannan. Eye Tracking for Everyone. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2176–2184, 2016. ISSN 10636919. doi: 10.1109/CVPR.2016.239.

[12] BG Kumar, Gustavo Carneiro, Ian Reid, et al. Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5385–5394, 2016.

[13] Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato. Adaptive Linear Regression for Appearance-Based Gaze Estimation. *Pami*, 36(10):2033–2046, 2014. ISSN 0162-8828. doi: 10.1109/TPAMI.2014.2313123.

[14] Wenjie Luo, Alexander G Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5695–5703, 2016.

[15] Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 1325–1334. IEEE, 2016.

[16] AJung Moon, Daniel M. Troniak, Brian Gleeson, Matthew K.X.J. Pan, Minhua Zheng, Benjamin A. Blumer, Karon MacLean, and Elizabeth A. Croft. Meet me where i'm gazing: How shared attention gaze affects human-robot handover timing. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction*, HRI '14, pages 334–341, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2658-2. doi:

10.1145/2559636.2559656. URL http://doi.acm.org/10.1145/2559636.2559656.

[17] Thies Pfeiffer. Towards gaze interaction in immersive virtual reality: Evaluation of a monocular eye tracking set-up. 2007.

[18] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Josh Susskind, Wenda Wang, and Russ Webb. Learning from simulated and unsupervised images through adversarial training. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 3, page 6, 2017.

[19] Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, and Francesc Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 118–126. IEEE, 2015.

[20] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. Learning-by-synthesis for appearance-based 3D gaze estimation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1821–1828, 2014. ISSN 10636919. doi: 10.1109/CVPR.2014.235.

[21] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1821–1828. IEEE, 2014.

[22] Li Sun, Mingli Song, Zicheng Liu, and Ming-Ting Sun. Real-time gaze estimation with online calibration. *IEEE MultiMedia*, 21(4):28–37, 2014.

[23] Marc Tonsen, Julian Steil, Yusuke Sugano, and Andreas Bulling. Invisibleeye: Mobile eye tracking using multiple low-resolution cameras and learning-based gaze estimation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, 1(3), 2017. doi: 10.1145/3130971. URL https://perceptual.mpi-inf.mpg.de/files/2017/08/tonsen17_imwut.pdf.

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

[25] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In *European Conference on Computer Vision*, pages 791–808. Springer, 2016.

[26] Marco Venturelli, Guido Borghi, Roberto Vezzani, and Rita Cucchiara. From depth data to head pose estimation: a siamese approach. *arXiv preprint arXiv:1703.03624*, 2017.

[27] MÃl'lodie Vidal, Jayson Turner, Andreas Bulling, and Hans Gellersen. Wearable eye tracking for mental health monitoring. *Computer Communications*, 35(11):1306–1311, 2012. URL http://dx.doi.org/10.1016/j.comcom.2011.11.002https://perceptual.mpi-inf.mpg.de/files/2013/03/vidal12_comcom.pdf.

[28] Fang Wang, Le Kang, and Yi Li. Sketch-based 3d shape retrieval using convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 1875–1883. IEEE, 2015.

[29] Kang Wang and Qiang Ji. Real Time Eye Gaze Tracking with Kinect. *iccv*, pages 1003–1011, 2017.

[30] Erroll Wood and A Bulling. Eyetab: Model-based gaze estimation on unmodified tablet computers. *Etra*, pages 3–6, 2014. doi: 10.1145/2578153.2578185. URL http://dl.acm.org/citation.cfm?id=2578185.

[31] Erroll Wood, Tadas Baltrušaitis, Louis Philippe Morency, Peter Robinson, and Andreas Bulling. A 3D morphable eye region model for gaze estimation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9905 LNCS:297–313, 2016. ISSN 16113349. doi: 10.1007/978-3-319-46448-0_18.

[32] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 4353–4361. IEEE, 2015.

[33] Chi Zhang, Rui Yao, and Jinpeng Cai. Efficient eye typing with 9-direction gaze estimation. *Multimedia Tools and Applications*, pages 1–18, 2017. ISSN 15737721. doi: 10.1007/s11042-017-5426-y.

[34] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4511–4520. IEEE, jun 2015. ISBN 978-1-4673-6964-0. doi: 10.1109/CVPR.2015.7299081. URL http://ieeexplore.ieee.org/document/7299081/.

[35] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation. 2016. URL http://arxiv.org/abs/1611.08860.

[36] Xucong Zhang, Michael Xuelin Huang, Yusuke Sugano, and Andreas Bulling. Training Person-Specific Gaze Estimators from User Interactions with Multiple Devices. 2018. doi: 10.1145/3173574.3174198. URL https://perceptual.mpi-inf.mpg.de/files/2018/01/zhang18{_}chi.pdf.

[37] Wangjiang Zhu and Haoping Deng. Monocular free-head 3d gaze tracking with deep learning and geometry constraints. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.