Personalization of Gaze Direction Estimation with Deep Learning

Zoltán Tősér[1], Róbert A. Rill[1,3], Kinga Faragó[1], László A. Jeni[2], and András Lőrincz[1]

[1] Faculty of Informatics, Eötvös Loránd University, Budapest, Hungary
[2] Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[3] Faculty of Mathematics and Informatics, Babeș-Bolyai University, Cluj-Napoca, Romania

Abstract. There is a growing interest in behavior based biometrics. Although biometric data has considerable variations for an individual and may be faked, yet the combination of such 'weak experts' can be rather strong. A remotely detectable component is gaze direction estimation and thus, eye movement patterns. Here, we present a novel personalization method for gaze estimation systems, which does not require a precise calibration setup, can be non-obtrusive, is fast and easy to use. We show that it improves the precision of gaze direction estimation algorithms considerably. The method is convenient; we exploit 3D face model reconstruction for the enrichment of a small number of collected data artificially.

1 Introduction

With the advance of facial expression recognition and animation technologies (see, e.g., [8] and the references therein) biometric information is becoming more and more ambiguous and imitable by computer graphics. Behavior based biometric may serve us as a rescue. It was shown more than a decade ago that facial expressions and head movements provide as relevant recognition cues as the face itself [22]. From both practical and theoretical point of views, imitation of such behavioral patterns will also be feasible in the near future, but – as argued many years ago – the more behavioral information is available, the better the chances are for the identification of anomalies and malicious episodes [18]. IoT and smart tools provide novel means for such characterization. On the other hand, remote identification of a person may not use IoT tools and only visible behavioral patterns may serve us. Eye movement pattern is one of the suggested components [1] and it may be used both in task oriented [2] and task independent settings [12, 6, 17]. Precision of the measurement is critical.

Another application field is gaze-based control, e.g., for special needs, since it may replace the need for wearable tools [24].

Here, we put forth a personalization method that can work with a small number of labeled samples, since we increase the number of samples artificially: we fit the mentioned 3D face model (i.e., [8]) to the image, rotate the model to

different head poses and increase our dataset with the 2D projections of the 3D data. Otherwise, the method would be of limited use, as we discuss it later.

The paper is organized as follows. We review the related gaze direction estimation works (Sect. 2) followed by the section on the databases and the estimation methods. The methods include deep learning, supervised descent, Support Vector Regression (SVR) that we use for the estimation of the facial mesh, positions of eye marker points, the head pose, and the gaze direction and we are searching for a good combination (Sect. 3). Results can be found in Section 4. Conclusions are drawn in the last section (Sect. 5).

2 Related Works

Gaze estimation systems are generally classified into two types: model-based and appearance-based methods. Our work is concerned with the latter.

In recent years, numerous papers have been published on appearance-based gaze estimation systems. Lu et al. [13, 14] described a method using Adaptive Linear Regression (ALR). They manually designed a feature descriptor based directly on normalized pixel intensities of the preprocessed eye region. They estimate the gaze positions by finding the best subset of the training samples, which linearly reconstructs the feature descriptor of the actual test sample. The estimated gaze position is computed with a linear regression on the selected subset.

Instead of regression, Smith et al. [20] solve a classification problem: they classify images to detect "gaze locking" i.e. direct eye contact with the camera. They also start from raw pixel intensities of a masked area on the image, but they apply principal component analysis and multiple discriminant analysis to achieve dimensionality reduction. Their classifier is a linear support vector machine.

Using the same dataset, Schneider et al. [19] compared various feature descriptors such as Histogram of Oriented Gradients (HOG) [5], Local Binary Patterns (LBP) [16] and raw pixel intensities in combination with different regression algorithms, such as k-Nearest-Neighbours and Support Vector Regression. They report that a multi-level HOG with LBP features and SVR make the best combination.

Sato et al. [23] presented a unique solution to enrich their training database for gaze estimation. They created a setup with multiple 2D cameras and reconstructed a 3D model of the subject's face. Given this 3D model they synthesised 2D images from multiple views thus increasing the variation coverage of the head pose. For regression, they used random forests on the image features combined with the data on the 3D head poses.

To our best knowledge Zhang et al. [26] were the first to use convolutional neural networks for gaze estimation. Alike Sato et al. [23], they also appended the head pose to the convolutional feature descriptor. They achieved slightly better results than Schneider et al. [19]. For more details, on this subject, the interested reader is referred to the paper of Zhang et al. [26].

3 Databases and Methods

3.1 Databases

Several datasets are publicly available for training gaze estimation systems, including the EYEDIAP [15], the MPIIGaze [26] and the UT Multi-view [23] sets. Among these, the full face is visible only in the EYEDIAP dataset. Since we extend our training dataset by fitting a 3D model of the head to images and we want to rotate the heads, our method requires the whole face. We used two datasets in our studies; the dataset from Columbia and our own dataset.

The 'Columbia Gaze Data Set' (CGDS) [20] consists of 5880 images from 56 subjects. The head of each subject was stabilized with a chin rest. The authors used multiple, carefully aligned cameras and gaze targets to record various head poses and gaze directions. The resolution of the images is high: they use 5184 × 3456 pixels. A sample image is shown in Fig. 1(a).

Our dataset (ELTE dataset) consists of recorded videos of 19 subjects (4 females and 15 males) taken in more realistic scenarios. Subjects were instructed to gaze directly into the camera and rotate their heads in different directions while keeping their gaze locked at the camera. We used a HD webcam, uniform lighting conditions, and a white canvas as background during data collection (Fig. 1(a)).

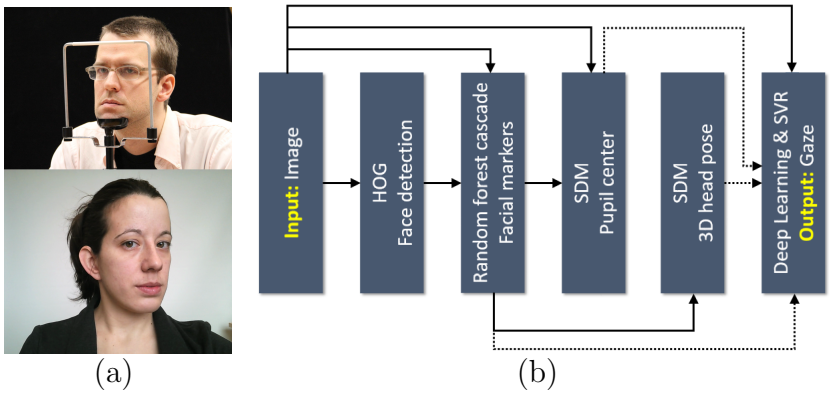

Fig. 1. (a) Datasets. Top: chins are stabilized and subjects look at predefined gaze targets [20]. Bottom: gazes are locked at the camera and head poses are changed. (b) Gaze estimation pipeline. Solid lines: used in all cases, dotted lines: used in some of the experiments. Marker positions are used for image normalization (not shown).

3.2 Methods

In our gaze estimation pipeline presented on Fig. 1(b) we use non-linear regressions at multiple stages that we review below.

As the first step of our pipeline, we estimate a bounding box around the face. The method we used is a linear support vector machine on Histogram of Oriented Gradients (HOG) features, similar to [5]. We used the open source implementation together with the available trained model from the *dlib library* [10]. Once the face has been detected, we estimate 2D facial landmarks from a small subset of pixel intensities, using a cascade of random forests [9].

The visual features used for gaze estimation were extended with the 3D head pose data. We used 49 pieces of 2D facial markers as the inputs to the Supervised Descent Method (SDM) [25] and estimated the head pose as follows: we constructed a 3D mean shape, rotated it, and successively minimized the angular error using the 2D reprojection error.

The position of the pupil center was also estimated and served as an optional additional feature. We used the facial marker positions of the SDM regressor, normalized our training images by converting them to grayscale, scaling them to a predefined intercanthal distance (ICD) and by rotating them in 2D to horizontal intercanthal direction. We also flipped each training image horizontally to increase the variance of our training data. The initial estimation was the centroid of the eye corners. We used HOG features with 9 signed bins.

The last step in our pipeline is the gaze estimation. We compared two variants: a Support Vector Regressor-based estimator (SVR) with HOG features as a baseline and a convolutional neural network (CNN) as the state-of-the-art. In both cases, we tried if additional features can improve the quality of the estimation, such as (i) the 3D head pose and (ii) the position of the 2D eye and pupil markers. In both cases, images were scaled to a predefined ICD.

We used both LIBLINEAR [7] and the LIBSVM [3] libraries for SVR estimations, both of which are publicly available. Details of these well-performing algorithms are well described in the literature [4].

We implemented a convolutional neural network similar to [26] in Lasagne. There are two main differences between the original implementation and ours: (i) we use dropout [21] and (ii) rectified linear units in all layers except for the output. Image patches were cut for both eyes with the centroid of the eye corners at the center. Adamax [11] and early stopping were used for network training.

Our architecture was composed of two convolutional and max pooling layer pairs, and 2 dense layers. The first four layers had 2×2 filters, except for the first convolutional one, which had 3×3. The optional head pose and pupil position were concatenated to the convolutional features. We used 1024 units in both fully connected layers. The dropout probability was 10% for the last pooling layer and for the first dense layer.

Beyond our pipeline, we included the ZFace tool [8], an SDM application. ZFace starts with an SDM based generic tracker that locates the 2D and 3D coordinates of main fiducial landmarks in each image. It then reconstructs a high resolution 3D mesh of 512 points. We generated new, realistic 2D projections of the face by mapping the texture to the 3D mesh and rotating it. Although ZFace could be used in the gaze estimation pipeline, due to time considerations, we inserted the cascade of random forests into that.

3.3 Personalization

The personalization method requires only a handful of images, yet it may decrease gaze estimation error by more than 40%. Our algorithm works as follows.

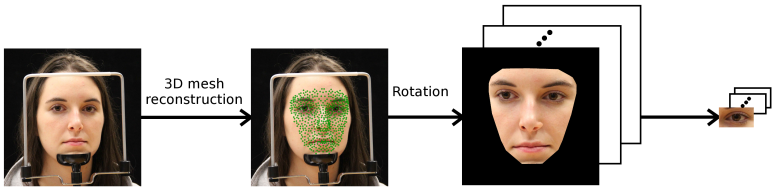

Fig. 2. Personalization pipeline on a CGDS image: input → reconstruct the 3D mesh → generate new training samples by rotating it.

1. Get 'personalization images' of the subject with known gaze vectors and various head positions, e.g., when both components of the gaze vector is 0, i.e. the subject is looking directly into the camera.
2. Fit a 3D mesh to each personalization image. We used ZFace [8] in this step.
3. Rotate the estimated 3D meshes in random directions and generate 2D projections. Calculate the gaze vectors in accordance with these rotations.
4. Improve the gaze estimation model with the generated 2D projections.

The method is sketched in Fig. 2. We explored different algorithmic combinations to be detailed in the following section (Sect. 4).

4 Results

We studied the performance of three algorithms, namely, LIBLINEAR, LIBSVM and CNN in the absence of personalization. We evaluated these algorithms on our own database and on the CGDS database. After reconstructing the 3D meshes on both of them, we first rotated them back to a frontal view, then we rotated them with angles drawn randomly from the uniform distributions in the ranges $[-30°, 30°]$ and $[-15°, 15°]$ for the yaw and pitch angles, respectively. We evaluated the gaze algorithms with leave-one-subject-out cross-validation. LIBSVM is somewhat better than LIBLINEAR, but in some cases we used the latter as our baseline due to computer time requirements; scaling characteristics of LIBSVM can be prohibitive for large sample sizes. In LIBLINEAR we used the solver for the dual problem and also employed a bias term. Results are shown in Table 1.

We evaluated the performance of the personalization pipeline for different algorithms and ICDs. We extended the visual features both with the head pose, the eye and pupil marker positions in all cases. Images used for personalization were randomly selected from the samples of each subject. The images were preprocessed the same way as in the evaluation of gaze estimation algorithms. We

Additional features	ELTE LSVM 32	ELTE LSVM 96	ELTE LLIN 32	ELTE LLIN 96	ELTE CNN 32	ELTE CNN 96	CGDS CNN 32	CGDS CNN 96
None	6.89	7.11	6.69	7.06	5.98	5.06	10.37	8.62
Pupil	4.88	7.69	5.20	5.36	5.64	5.07	10.11	8.53
Head pose	6.17	**6.35**	6.05	6.06	3.92	**3.85**	**8.07**	**6.97**
Pupil + h.pose	**3.78**	7.96	**5.07**	**5.27**	**3.82**	3.86	8.28	7.12

Table 1. Comparisons of performances for the two databases ELTE and CGDS [20] and for the three algorithms LIBSVM (LSVM), LIBLINEAR (LLIN), and CNN. 32 and 96 in the table header denote the ICD we used for scaling. The table shows mean angular errors in degrees.

Pers. images	ELTE LSVM	ELTE LLIN	ELTE CNN 32 ICD (a)	ELTE CNN 32 ICD (b)	ELTE CNN 96 ICD (a)	ELTE CNN 96 ICD (b)	CGDS CNN 96 ICD (a)	CGDS CNN 96 ICD (b)
0	3.78	5.07	3.82	3.83	3.86	3.89	7.12	7.09
5	2.81	3.71	2.91	3.06	2.45	3.29	6.13	6.47
10	2.56	3.45	2.61	2.75	2.24	3.24	5.59	6.27
15	2.36	3.13	2.34	2.26	2.02	2.28	4.98	5.33
20	2.22	3.08	2.21	2.04	1.80	1.93	4.61	4.59

Table 2. Comparisons of personalization performances for the two databases ELTE and CGDS [20], for the the three algorithms LIBSVM (LSVM), LIBLINEAR (LLIN), CNN, and for different number of personalization images. All runs had both pupil and head pose data as inputs. For each personalization image 10 rotated samples were generated. Notation: augmented database is (a): trained from scratch, (b): added as a single mini batch at the end of training. The displayed values are mean angular errors in degrees.

show our results on Table 2. By using the personalization pipeline, performace increases gradually. For 20 personalization images rotated to 10 different head poses, the mean gaze error fell down to less than two third of its original value (from 100% to 58%) on the average.

5 Summary

We have presented a non-obtrusive method together with a learning architecture for gaze direction estimation in a considerable range of head pose angles. Such estimations have a number of applications from the medical field, to remote surveillance systems and also computer assisted education. The special feature of our method is the personalization capability that does not require a complicated calibration setup, yet improves precision considerably.

6 Acknowledgements

This work was supported in part by EIT Digital under grant No. 16257. The authors thank the contributions of Tamás Nyíri who ran the numerical studies.

References

1. Bednarik, R., Kinnunen, T., Mihaila, A., Fränti, P.: Eye-movements as a biometric. In: Image analysis, pp. 780–789. Springer (2005)
2. Cantoni, V., Galdi, C., Nappi, M., Porta, M., Riccio, D.: Gant: Gaze analysis technique for human identification. Pattern Recognition 48(4), 1027–1038 (2015)
3. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2(3), 27 (2011)
4. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning 20(3), 273–297 (1995)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. vol. 1, pp. 886–893. IEEE (2005)
6. Eberz, S., Rasmussen, K.B., Lenders, V., Martinovic, I.: Preventing lunchtime attacks: Fighting insider threats with eye movement biometrics. In: NDSS (2015)
7. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. The Journal of Machine Learning Research 9, 1871–1874 (2008)
8. Jeni, L.A., Cohn, J.F., Kanade, T.: Dense 3d face alignment from 2d videos in real-time. In: Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on. vol. 1, pp. 1–8. IEEE (2015)
9. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1867–1874 (2014)
10. King, D.E.: Dlib-ml: A machine learning toolkit. The Journal of Machine Learning Research 10, 1755–1758 (2009)
11. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Kinnunen, T., Sedlak, F., Bednarik, R.: Towards task-independent person authentication using eye movement signals. In: Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications. pp. 187–190. ACM (2010)
13. Lu, F., Sugano, Y., Okabe, T., Sato, Y.: Inferring human gaze from appearance via adaptive linear regression. In: Computer Vision (ICCV), 2011 IEEE International Conference on. pp. 153–160. IEEE (2011)
14. Lu, F., Sugano, Y., Okabe, T., Sato, Y.: Adaptive linear regression for appearance-based gaze estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence 36(10), 2033–2046 (2014)
15. Mora, K.A.F., Monay, F., Odobez, J.M.: 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. pp. 255–258. ACM (2014)
16. Ojala, T., Pietikainen, M., Harwood, D.: Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In: Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on. vol. 1, pp. 582–585. IEEE (1994)
17. Rigas, I., Komogortsev, O., Shadmehr, R.: Biometric recognition via eye movements: Saccadic vigor and acceleration cues. ACM Transactions on Applied Perception (TAP) 13(2), 6 (2016)

18. Ross, A., Jain, A.: Information fusion in biometrics. Pattern recognition letters 24(13), 2115–2125 (2003)
19. Schneider, T., Schauerte, B., Stiefelhagen, R.: Manifold alignment for person independent appearance-based gaze estimation. In: 2014 22nd International Conference on Pattern Recognition (ICPR). pp. 1167–1172. IEEE (2014)
20. Smith, B.A., Yin, Q., Feiner, S.K., Nayar, S.K.: Gaze locking: passive eye contact detection for human-object interaction. In: Proceedings of the 26th annual ACM symposium on User interface software and technology. pp. 271–280. ACM (2013)
21. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 15(1), 1929–1958 (2014)
22. Stone, J.: Face recognition: When a nod is better than a wink. Current Biology 11(16), R663–R664 (2001)
23. Sugano, Y., Matsushita, Y., Sato, Y.: Learning-by-synthesis for appearance-based 3d gaze estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1821–1828 (2014)
24. Vörös, G., Verő, A., Pintér, B., Miksztai-Réthey, B., Toyama, T., Lőrincz, A., Sonntag, D.: Towards a smart wearable tool to enable people with sspi to communicate by sentence fragments. In: Pervasive Computing Paradigms for Mental Health, pp. 90–99. Springer (2014)
25. Xiong, X., Torre, F.: Supervised descent method and its applications to face alignment. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 532–539 (2013)
26. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4511–4520 (2015)