Head pose-free gaze estimation using domain adaptation

Byungtae Ahn,1 Minseok Seo,2 and Dong-Geol Choi1✉
1Department of AI Machinery, Mechanical Systems Safety Research Division, Korea Institute of Machinery & Materials, Daejeon, Korea
2Department of Information and Communication Engineering, Hanbat National University, Daejeon, Korea
✉Email: dgchoi@hanbat.ac.kr

Human gaze information has been widely used in various areas, such as medical diagnosis and human–computer interactions (HCI). This study proposes a head pose-free 3D gaze estimation method using a deep convolutional neural network (DCNN). To infer gaze direction, only a small grayscale image is required without any special devices such as an infrared (IR) illuminator and RGBD sensor. A domain adaptation method to reduce the feature gap between real and synthetic image data is also proposed here. Moreover, a novel synthetic dataset (SynFace) that contains head poses, gaze directions, and facial landmarks is established and released. The proposed method outperforms state-of-the-art methods and achieves a mean error of less than 4°.

Introduction: Gaze information has long been used in medical diagnosis [1], psychological analysis [2], human–computer interactions (HCI) [3], marketing [4], and driver state estimation [5]. Gaze estimation has been studied for decades. The approaches can largely be classified into model-based methods and appearance-based methods. Early studies focused on model-based methods [6]; such methods use geometric features such as the pupil centres and corneal reflections from infrared light sources. However, these methods apparently have drawbacks due to the following limiting factors: (1) they require expensive equipment; (2) sunlight interferes with infrared sensors, which means that the systems can operate only indoors; and (3) there are strict distance limitations (usually 60 cm) because of the need to obtain a high-resolution eye image. In contrast, the appearance-based methods directly learn mapping functions from image to gaze vectors [6]. These methods require only a camera and can be used both in outdoor environments and at a variety of distances.

Early studies assumed a fixed head pose, but they were extended to allow some head movement by adding complementary techniques [7, 8]. Recent studies [4, 9] combine head pose and eye information by a feature concatenation method. However, they use flexible face models to detect face and eye regions. Since these model-based methods depend on facial landmark detection, their performances are sensitive to various environments where facial landmarks are difficult to detect (e.g. head pose changes, occlusions, and low-resolution images).

Gaze estimation methods can be divided into 2D and 3D methods depending on the type of results. Methods that output two-dimensional coordinates (x and y) are used in devices such as mobile phones and tablets. It is relatively easy to obtain ground truth data for a 2D gaze estimation method that estimates gaze points only on a screen for a specific device. These methods require only the user’s face image in front of the device and the corresponding pixel position on the screen. In contrast, methods that estimate 3D gaze vectors in free space are widely used in various fields such as psychological analysis, driver state estimation, and marketing. It is relatively difficult to obtain ground truth data because not only the user’s face image but also the 3D gaze vector with respect to the camera coordinate system in 3D space are required. To address appearance variations of eyes corresponding to head pose, previous works have used bias compensation [7] and warping methods [8, 10]. Although those methods simplify the problem somewhat, the estimated model does not represent all aspects of appearance changes. To address these limitations, recent studies [11, 12] have gathered large datasets that cover various head poses and eye appearances. In this study, we follow this approach by building a large synthetic dataset that contains the most possible head poses and eye movements. We adopt an appearance-based approach and propose a 3D gaze estimation method using a deep convolutional neural network (DCNN). We train the network using both synthetic data and real image data. The synthetic data supplement the real training data, which is difficult to obtain in sufficient quantities for the 3D methods. In addition, the learned features are not biased toward a specific head pose and eye movement. However, the insufficient realism of the synthetic data may cause a gap between the features learned from the synthetic and real image domains. To address this problem, we introduce a domain adaptation method suitable for the proposed deep neural network (DNN). Moreover, we rendered a synthetic dataset (SynFace) that represents the most possible head poses and eye movements. To mimic a variety of environmental factors, various head poses, eye movements, and types of illumination were added. We will release this dataset to assist future researchers in this area. Our contributions can be summarized as follows:

1. We propose a DNN structure for domain adaptation in 3D gaze estimation.
2. We propose a domain adaptation method that uses both real and synthetic data.
3. We construct and release a large synthetic head pose and gaze dataset (SynFace) that includes various head poses and eye movements.

Model-based approaches: Chen and Ji [13] used detected rigid points from 3D generic face models. Mora et al. [10] used the 3D eyeball centre, 3D pupil centre, and some personal parameters calculated from detected 2D landmarks by a supervised-descent-method facial landmark tracker (SDM tracker).

Xiong et al. [14] used the iris centre and a set of 2D facial landmarks whose 3D locations are provided by an RGBD camera. Wood et al. [15] used a 3D morphable eye region model, and Wang et al. [16] used a 3D deformable eye-face model. Our method outperforms all state-of-the-art model-based methods because their features are calculated from detected 2D facial landmarks that are highly sensitive to noise, low resolution, and head pose changes.

Appearance-based approaches: All state-of-the-art appearance-based methods are based on convolutional neural networks (CNNs). Shrivastava et al. [17] used 1 million synthetic images for training. They trained their CNN using refined synthetic images similar to real images by employing generative adversarial networks [18]. Zhang et al. [12] proposed GazeNet, the first deep appearance-based gaze estimation method. They also provided a gaze dataset (MPIIGaze) consisting of face images and corresponding ground truth gaze positions. Zhang et al. [19] used the MPIIGaze dataset [12] for training and encoded full-face images using a CNN with spatial weights. Note that Shrivastava et al. [17] used only synthetic images, while Zhang et al. [12, 19] used only real images. Our domain adaptation method, which uses both real and synthetic images, outperformed all those methods.

Domain adaptation approaches: Park et al. [20] and Guo et al. [21] like our motivation, gathering the gaze estimation data set is a labour-intensive task, and thus proposed a solution to solve this problem. In particular, Guo et al. [21] effectively applied unsupervised domain adaptation by utilizing the prediction consistency between the output of the source domain and the target domain, and achieved state-of-the-art results in two datasets, EYEDIAP [22] and MPIIGaze [12]. However, because both the source domain and the target domain are real data, the experimental results between synthetic data and real data with a relatively large domain gap are missing. In this work, we propose a method of collecting and using synthetic data, which is the most intuitive way to solve the problem that it is difficult to collect data sets. Also, we propose to apply domain adaptation to synthetic data and real data in the field of gaze estimation. Finally, the possibility of using synthetic data is shown through the experimental results.

Methodology: This section describes in detail our gaze regression and domain adaptation method, the dataset used in the training process, and the SynFace dataset we proposed.

Gaze regression and domain adaptation: Recent studies [9, 12] have used concatenated features from head pose and gaze. However, the approach for connecting the individual features ignores the correlation between head pose and eye movement. In this study, the correlation between head pose and eye movements is implicitly learned by inferring the gaze direction based on changes in eye appearance in most possible head poses from synthetic image data. The left and right eye appearances are converted to phi and theta angles (ϕl, θl, ϕr, and θr), respectively.
We utilize a DNN model proposed by Ahn et al. [23, 24] to detect face and head pose and extend it to gaze estimation. However, to use combined data from two different domains (real and synthetic images), a specific learning technique is necessary. In this study, we propose a domain adaptation method that uses real and synthetic data for gaze estimation (see Figure 1) and a two-step learning method. The face and eye regions are extracted using a deep-learning-based real-time face detector [24]. The eye image patch is used as the input to the proposed DNN for gaze estimation. The input to the network is a 64 × 64 pixel grayscale image. The feature extraction stage is composed of four convolution layers, three pooling layers, one shared fully connected layer, and one fully connected layer for each task (gaze regression and domain adaptation). A rectified linear unit (ReLU) [25] is used as the activation function, and max-pooling is used in the pooling layer. The fully connected layer, which follows the four convolution layers, creates a 64-dimensional feature vector that is shared by the multitask estimation stage. The loss function of the proposed network is formulated as follows:

\[ L = L_{\text{gaze}} + L_{\text{domain}} \]

where \( L \) is the total loss of the proposed network. The gaze regression uses Euclidean distance as follows:

\[ L_{\text{gaze}}(X'; W) = \sum_{i} \| g_{i} - f(X'; W) \|^2_2 \]

where \( X \) is a 64 × 64 px input image, \( i \) is the index of a training sample, and \( f \) represents the estimated gaze angles (pitch and yaw) in normalized degree units. \( W \) is a set of features, the weights of the convolution filters, and \( g_{i} \) is the ground truth gaze angle.

The proposed network is composed of three components: features, gaze regressor, and domain adapter. The feature extractor learns filters independent of the domain. The extracted features are transmitted to the gaze regressor and the domain adapter in the last layer. The goal of the domain adapter is to transfer the synthetic and real image features to a common domain. To achieve this goal, we use an adversarial training approach. The domain adapter in the proposed DNN is a real and synthetic image classifier. Given the features of an input image, it classifies whether the features are from a real image or a synthetic image. Similar to classifier training, the inputs to the domain adaptation network are real and synthetic images with annotations. The domain adaptation task in the proposed network is learned in two steps, and the loss function, \( L_{\text{domain}} \), is different as follows:

\[ L^{\text{step1\_domain}} = -\left( \sum_{j \in \text{real}} \log p_j + \sum_{j \in \text{synthetic}} \log (1 - p_j) \right) \]

\[ L^{\text{step2\_domain}} = -\left( 0.5 \log p_j + 0.5 \log(1 - p_j) \right) \]

where \( p_j \) is the probability that \( j \) is a real image, and obtained using the softmax function. In the first step, the gaze regressor and the domain adapter, as a usual classifier, are trained together. As a result, we have a well-trained gaze regressor and a real/synthetic classifier. In the second step, the parameters of the domain adapter are fixed, while the features and the gaze regressor are further refined. The trained real/synthetic classifier becomes confused by the loss function of the domain adapter in the second step. The classifier outputs the probability distribution \((0.5, 0.5)\) for all real and synthetic images due to the loss function, which means that the features are adapted until they are indistinguishable for these two domains.

Figure 2 shows \( t \)-distributed stochastic neighbour embeddings (t-SNE) 2D feature visualization. The red and blue colours denote real and synthetic images, respectively. Each digit indicates an annotation dimension. Using the CNN alone, the trained features show a clear gap between real images and synthetic images; however, through domain adaptation, they show no difference between the two domains. This result demonstrates that the real and synthetic data are mapped to a single domain through the proposed domain-adversarial neural network (DANN).

Datasets: In this study, we used three datasets, as shown in Figure 3. The MPIIGaze dataset [12] contains approximately 200K images. It was collected using a laptop during the daily lives of 15 participants. The data include a wide variety of appearance and lighting conditions because they were obtained from unconstrained environments. The Columbia Gaze dataset [26] was collected from 56 participants who were asked to look at a grid of dots on the wall. Their heads were held steady by a head-clamp device and captured in five different horizontal head orientations, from \([-30°, 30°]\). Although the dataset contains only 5880 images with annotations, the image resolution is extremely high, and the gender and race ratio of the data are evenly distributed. The dataset is composed of images of 32 males and 24 females, 21 of which are Asian, 19 are white, 8 are South Asian, 7 are black, and 4 are Hispanic or Latino.

In addition, we built a synthetic dataset (SynFace). We created 60 3D face models representing various ages, genders, and races using iClone7 software [27]. Then, we generated ground truth data by implementing a variety of head poses and gazes under various lighting directions using Unity software [28]. In the SynFace dataset, the horizontal head orientation (yaw angle) range is \([-75°, 70°]\) and the vertical head orientation (pitch angle) range is \([-45°, 45°]\). The gaze direction covers a range of \([-60°, 60°]\) in both phi and theta angles, which is not independent of the head pose. The dataset includes 120K images with corresponding head poses, gaze directions, and positions of 49 facial landmarks. The synthetic data supplement the real data that is difficult to obtain, so the result is not biased toward specific individuals or facial expressions.

Experiments: In this section, we demonstrate our gaze estimation algorithm. We verified the validity of the SynFace synthetic dataset and compared the results of the proposed method with those of state-of-the-art methods. To verify the validity of our SynFace synthetic dataset, we tested the model’s performance when synthetic data were added to real image data in specific proportions. The real data used for training consisted of approximately 110K images that included 10 subjects from the MPIIGaze dataset [12] and 37 subjects from the Columbia Gaze dataset [26]. The remaining subjects were used for validation. Table 2 shows the results of the validation set. As the quantity of added synthetic data increased, the accuracy also increased. The performance improved by approximately 28% when using both the real and SynFace datasets compared to real image data alone.
Table 1. Comparison of gaze estimation results of the proposed method with those of the state-of-the-art methods. Top: Mean errors of the state-of-the-art model-based methods and our methods on the Columbia Gaze dataset [26]. Bottom: Mean errors of the state-of-the-art appearance-based methods and our methods on the MPIIGaze dataset [12].

| Method | Dataset | Mean error [°] |
|--------|---------|----------------|
| Rigid points detection | Columbia Gaze | 12.30 |
| SDM tracker | Columbia Gaze | 10.20 |
| Iris detection | Columbia Gaze | 9.70 |
| 3D eye region model | Columbia Gaze | 9.44 |
| 3D eye free model | Columbia Gaze | 7.10 |
| Ours (real) | Columbia Gaze | 4.76 |
| Ours (+syn.) | Columbia Gaze | 4.24 |
| Ours (+syn. with DANN) | Columbia Gaze | 3.67 |
| GAN (synthetic data) | MPIIGaze | 11.20 |
| GAN (refined synthetic data) | MPIIGaze | 7.80 |
| MPIIGaze (CNN) | MPIIGaze | 6.10 |
| 3D eye free model | MPIIGaze | 4.80 |
| Ours (real) | MPIIGaze | 4.93 |
| Ours (+syn.) | MPIIGaze | 3.99 |
| Ours (+syn. with DANN) | MPIIGaze | 3.73 |
| GAN (CNN) | MPIIGaze | 3.73 |
| GAN (refined synthetic data) | MPIIGaze | 3.99 |

Table 2. Mean errors when adding synthetic images to real images. The labels indicate the number of synthetic images with respect to the number of real images used in training.

| Synthetic data ratio | Mean error [°] |
|---------------------|----------------|
| 0                   | 5.10           |
| 0.5                 | 4.78           |
| 1                   | 4.29           |
| 2                   | 3.67           |

We compared our method with the state-of-the-art methods of model-based and appearance-based approaches. The comparison results are shown in Table 1. As a result of the experiment, it was confirmed that the performance was improved by using the domain adaptation method rather than simply mixing and training real and synthetic data.

Conclusions: In this study, we propose a 3D gaze estimation method using a DNN and a domain adaptation method that reduces the feature gap between real and synthetic image domains. In addition, we build and release a large synthetic dataset called SynFace that represents various lighting, races, ages, genders, head poses, and gaze directions.

Compared to using simply mixed data, the proposed domain adaptation method results in a better accuracy. The experimental results demonstrate the superiority of our method by comparing it with the state-of-the-art approaches of both model-based and appearance-based methods.

Acknowledgments: This study was supported by a research grant from Hanbat National University (No. 202003290001), NRF grant funded by the Korean Government (MIST, No. 2019029049) and partially supported by a research grant from the National Research Council of Science Technology under the RD Program of the Ministry of Science, ICT and Future Planning, as a part of the research project, "Development of core machinery technologies for autonomous operation and manufacturing (NK230G)."

© 2021 The Authors. Electronics Letters published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Received: 10 February 2021 Accepted: 29 May 2021 doi: 10.1049/el.2021.0247

References

1 Holzman, P.S., et al.: Eye-tracking dysfunctions in schizophrenic patients and their relatives. Arch. Gen. Psychiatry 31(2), 143–151 (1974)
2 Rayner, K.: Eye movements in reading and information processing: 20 years of research. Psychol. Bull. 124, 372–422, 12 (1998)
3 Morimoto, C.H., Mimica, M.R.: Eye gaze tracking techniques for interactive applications. Comput. Vision Image Understanding 98(1), 4–24 (2005)
4 Zhang, X., et al.: An advertisement video analysis system based on eye-tracking. In: 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity), Chengdu, China, 19–21 Dec. 2015
5 Vicente, F., et al.: Driver gaze tracking and eyes off the road detection system. IEEE Trans. Intell. Transp. Syst. 16(4), 2014–2027 (2015)
6 Hansen, D.W., Ji, Q.: In the eye of the beholder: a survey of models for eyes and gaze. IEEE Trans. Pattern Anal. Mach. Intell. 32(3), 478–500 (2010)
7 Feng Lu, Y.S., et al.: A head pose-free approach for appearance-based gaze estimation. In: Proceeding of the British Machine Vision Conference (BMVC), BMVA Press, 29 Aug.–2 Sept. 2011
8 Lu, F., et al.: Head pose-free appearance-based gaze sensing via eye image synthesis. In: IEEE International Conference on Pattern Recognition (ICPR), Tsukuba, Japan, 11–15 Nov. 2012
9 Kratka, K., et al.: Eye tracking for everyone. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016
10 Mora, K.A.F., Odozeb, J.-M.: Gaze estimation from multimodal kinetic data. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Providence, RI, USA, 16–21 June 2012
11 Sugano, Y., et al.: Learning-by-synthesis for appearance-based 3d gaze estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014
12 Zhang, X., et al.: Mpiigaze: real-world dataset and deep appearance-based gaze estimation. IEEE Trans. Pattern Anal. Mach. Intell. 41(4), 162–175 (2019)
13 Chen, J., Ji, Q.: 3D gaze estimation with a single camera without IR illumination. In: IEEE International Conference on Pattern Recognition (ICPR), Tampa, FL, USA, 8–11 Dec. 2008
14 Xiong, X., et al.: Eye gaze tracking using an RGBD camera: A comparison with a RGB solution. Ubicomp 2014 – Adjunct Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, New York, NY, USA, 13–17 Sept. 2014
15 Wood, E., et al.: A 3d morphable eye region model for gaze estimation. In: European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 Oct. 2016
16 Wang, K., Ji, Q.: Real-time eye gaze tracking with 3d deformable eye-face model. In: IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 22–29 Oct. 2017
17 Shrivastava, A., et al.: Learning from simulated and unsupervised images through adversarial training. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017
18 Goodfellow, I.J., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (NIPS), MIT Press, Cambridge, MA, USA, (2014), pp. 2672–2680.
19 Zhang, X., et al.: It’s written all over your face: full-face appearance-based gaze estimation. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017
20 Park, S., et al.: Few-shot adaptive gaze estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea (South), 27 Oct.–2 Nov. 2019
21 Guo, Z., et al.: Domain adaptation gaze estimation by embedding with prediction consistency. In: Proceedings of the Asian Conference on Computer Vision, Kyoto, Japan, 30–4 Dec. 2020
22 Funes Mora, K.A., et al.: Eyedat: 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, New York, NY, USA, 26–28 Mar. 2014, pp. 255–258.
23 Ahn, B., et al.: Real-time head orientation from a monocular camera using deep neural network. In: Asian Conference on Computer Vision (ACCV), Singapore, Singapore, 1–5 Nov. 2014
24 Ahn, B., et al.: Real-time head pose estimation using multi-task deep neural network. Rob. Auton. Syst. 103, 1–12 (2018)
25 Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: International Conference on Machine Learning (ICML). Omnipress Haifa, Israel, June 21–24, 2010
26 Smith, B.A., et al.: Gaze locking: Passive eye contact detection for human-object interaction. In: ACM Symposium on User Interface Software and Technology (UIST), New York, NY, USA, 8–11 Oct. 2013
27 iClone7 software [Online]. Available: https://www.realusion.com/icleone/
28 Unity software [Online]. Available: https://unity.com/