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

Appearance-based gaze estimation is believed to work well in real-world settings, but but existing datasets have been collected under controlled laboratory conditions. These conditions are characterised by limited variability of eye appearances. We introduce the first large-scale dataset MPIIGaze for appearance-based gaze estimation. Our dataset is one order of magnitude larger than existing datasets and significantly more variable with respect to illumination and appearance. The MPIIGaze dataset contains 213,659 images collected from 15 laptop users over several months (see Figure 1). Our recording software automatically asked participants to look at a random sequence of 20 on-screen positions every 10 minutes. No other instructions were given to them, in particular no constraints as to how and where to use their laptops. Therefore our MPIIGaze covers a realistic variability in appearance and illumination and represents a significant advance over existing datasets. The dataset and annotations are publicly available online. Compared with the other recent datasets, the Eyediap [2] dataset does not cover the range of gaze directions that can occur during laptop interactions and the UT Multiview [5] lacks diversity in variations of appearance. Details of the recording and feature of our MPIIGaze are described in the paper.

We also 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. Figure 2 provides an overview of our proposed method for in-the-wild appearance-based gaze estimation using multimodal convolutional neural networks (CNN) [4]. We first employ state-of-the-art face detection [3] and facial landmark detection [1] 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 [5] to crop and warp the head pose and eye images to the normalised training space. We use a multimodal CNN model to take advantage of both eye image and head pose information. The CNN is used to learn the mapping from the head poses and eye images to gaze directions in the camera coordinate system, which encodes head pose information into our CNN model by concatenating $\boldsymbol{h}$ with the output of the fully connected layer.

Current appearance-based gaze estimation methods are also not evaluated across different datasets, which bears the risk of significant dataset bias. So 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 We study two key tasks through extensive evaluations of appearance-based gaze estimation algorithms on three publicly available gaze estimation datasets: cross-dataset and within-dataset evaluation conditions. 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 across datasets. In contrast, if we can assume that training data is directly collected in the target daily-life environment, the rich training data can be fully utilised. We present both cross-dataset and within-dataset evaluations on three datasets: Eyediap [2], UT Multiview [5], and our own MPIIGaze. These evaluation allows us to identify key research challenges of gaze estimation in the wild. During the evaluation, we put more focus on person-independent scenario since such setting has the potential to bring appearance-based methods into settings that do not require any user-specific training.

We compare the performance of our CNN-based approach to several state-of-the-art image-based gaze estimation algorithms. In cross-dataset evaluation condition, we selected the UT Multiview dataset as the training dataset our CNN-based approach shows the best accuracy on both test 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 [6], $p < 0.05$) over the state-of-the-art Random Forests method. To analyse the limits of person-independent performance on the MPIIGaze dataset, we performed leave-one-person-out evaluation on the MPIIGaze dataset. 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. More details can be found in our paper.

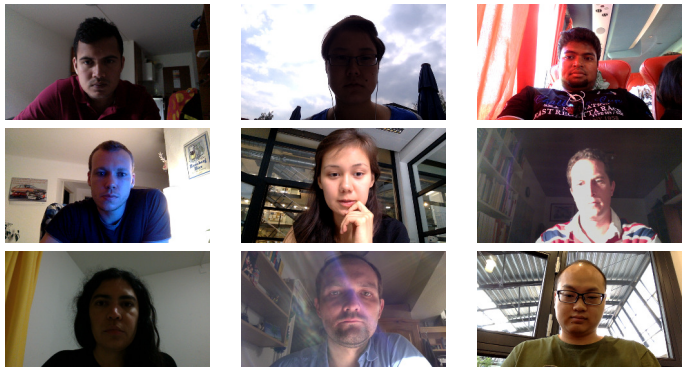

Figure 1: Sample images from our MPIIGaze dataset showing the considerable variability in terms of place and time of recording, directional light and shadows.

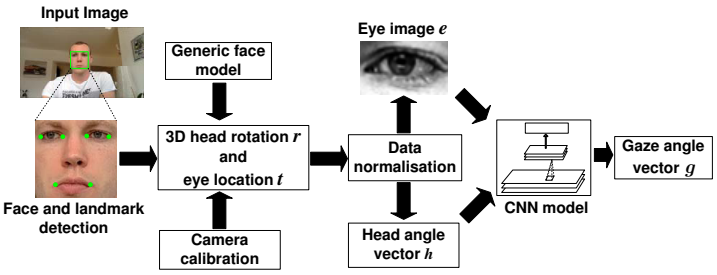

Figure 2: Overview of our method for in-the-wild appearance-based gaze estimation using multimodal convolutional neural networks.

[1] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Continuous conditional neural fields for structured regression. In *Proc. ECCV*, pages 593–608, 2014.

[2] 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 *Proc. ETRA*, pages 255–258, 2014.

[3] Jianguo Li and Yimin Zhang. Learning surf cascade for fast and accurate object detection. In *Proc. CVPR*, pages 3468–3475, 2013.

[4] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Ng. Multimodal deep learning. In *Proc. ICML*, pages 689–696, 2011.

[5] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *Proc. CVPR*, pages 1821–1828, 2014.

[6] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, pages 80–83, 1945.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.