A Deep Learning Approach to Appearance-Based Gaze Estimation under Head Pose Variations

Hsin-Pei Sun
National Tsing Hua Univ.
Computer Science Dept.
Hsinchu, Taiwan
rex2246511@gmail.com

Cheng-Hsun Yang
National Tsing Hua Univ.
Computer Science Dept.
Hsinchu, Taiwan
xu3cj84wu0h9@gmail.com

Shang-Hong Lai
National Tsing Hua Univ.
Computer Science Dept.
Hsinchu, Taiwan
lai@cs.nthu.edu.tw

***Abstract*—In this paper, we propose a deep learning based gaze estimation algorithm that estimates the gaze direction from a single face image. The proposed gaze estimation algorithm is based on using multiple convolutional neural networks (CNN) to learn the regression networks for gaze estimation from the eye images. The proposed algorithm can provide accurate gaze estimation for users with different head poses, since it explicitly includes the head pose information into the proposed gaze estimation framework. The proposed algorithm can be widely used for appearance-based gaze estimation in practice. Our experimental results show that the proposed gaze estimation system improves the accuracy of appearance-based gaze estimation under head pose variations compared to the previous methods.**

***Keywords*—** Gaze estimation, deep learning, convolutional neural network

I. INTRODUCTION

There are more and more new applications based on human gaze estimation recently. For example, some systems based on eye tracking can detect what contents the website users are interested in and they are useful for website design [1], [2]. Besides, gaze locking can determine if the user is looking at the camera for subsequent processing [3]. Moreover, gaze control was touted as a crucial feature of future PC, such as Tobii eye tracking device. Therefore, the development of gaze estimation is important to many fields, including human-machine interaction, first-person vision, visual behavior analysis [4], e-learning [5], video communication [6], etc.

Nowadays, gaze direction could be estimated accurately by using model-based methods, such as Tobii eye tracking device. Model based methods estimate the gaze direction by using geometric eye features (e.g., the pupil position). Nakazawa et al. [7] used high-power infrared (IR) projector to obtain the corneal image reflected from the scene illuminated with structured light. However, this approach requires specialized and expensive hardware, which limits its application in practice. In contrast, appearance-based methods are more general since they can provide eye tracking without using active sensing technologies.

The gaze estimation methods based on appearance characteristics can be formulated as finding a mapping function from eye image features to gaze directions. And appearance-based methods usually require large amounts of training data. For example, TabletGaze [8] contains 100,000 images of people looking at a tablet screen, and MPIIGaze [9] contains 213,659 images collected during everyday laptop use over several months. Since it is difficult to collect real images with a wide range of head poses, some recent works tried to train on synthetic images. Wood et al. [10] used a generative 3D model of the human eye region to synthesize large amounts of eye images. However, networks learning from synthetic data may fail to generalize well to real images. Hence, Shrivastava et al. [11] proposed Simulated+Unsupervised learning to close the gap between synthetic and real images. Although synthetic images are not limited to synthesize head pose variations, it is hard to learn the correct features from eye images for gaze estimation under a large range of head poses.

To address the gaze estimation problem with variant head poses and different users, we use UT Multiview dataset since it contains a wide variety of eye images of numerous persons with different gazes at different head poses. To account for the large variations of head pose in the training data, we propose to learn multiple CNN models from the data with local head poses to estimate the gaze direction. The proposed gaze estimation system provides accurate gaze estimation with arbitrary head poses and non-specific users.

II. PROPOSED SYSTEM

Fig. 1 shows the whole training and estimation pipeline of the proposed system using multiple pose-based CNN models with a VGG-like architecture [12]. We first employ a state-of-the-art face detection and facial landmarks detection to locate eye landmark points from the calibrated monocular 4K camera. The next step is normalization of the input image obtained by cropping the eye image through landmark points, followed by image preprocessing. We then fit the 3D face model with several reasonable landmark points to estimate the 3D head pose. Multi-CNN models trained by eye images of adjacent poses and the estimated gaze direction is obtained

2327-0985/17 $31.00 © 2017 IEEE
DOI 10.1109/ACPR.2017.155

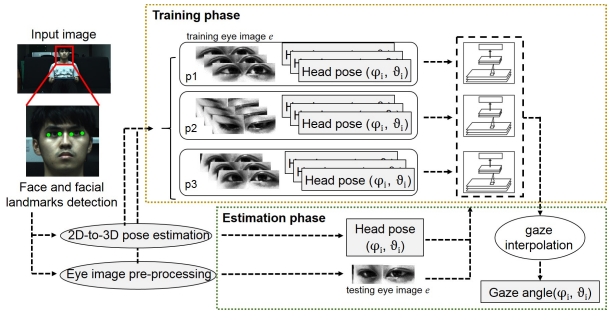

Fig. 1. A flowchart of the proposed system including training and estimation phase using multiple pose-based CNN models.

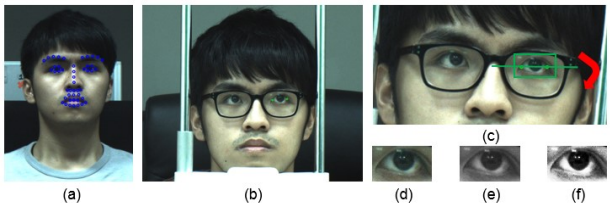

Fig. 2. (a) The detected face with 49 facial landmarks. (b) The two green points are the anchor points on the left eye. (c) The face image is rotated to make the line of the two anchor points parallel to the horizontal axis. (d) Normalize the eye patches to 60×36 images. The eye patch images after (e) gray-scale conversion and (f) histogram equalization.

by interpolating the gaze estimation results predicted from pose-based CNN models in the camera coordinate system.

A. Image Patch Normalization

Before training the multi-CNN models for gaze direction, we first preprocess the input data, estimate the 3D head poses, and extract local eye patches for the CNN models.

To apply supervised learning to appearance-based gaze estimation, we prepare eye images and associated head poses as input data and the corresponding gaze directions as the output label. First, we detect and track 49 facial landmarks by using IntraFace [13] library. An example of the detected facial landmarks is depicted in Fig. 2(a). The steps from Fig. 2(b)-(f) involve cropping the left eye region, rotating the eye image with the angle determined by the eye anchor points, and image intensity normalization.

We apply the above image preprocessing steps to extract normalized eye patches. Then, we use the extracted eye patches as input to the CNN models for gaze estimation.

B. Head Pose Estimation

To estimate the 3D head pose from a 2D face image, the 2D-to-3D pose estimation is employed here, which requires

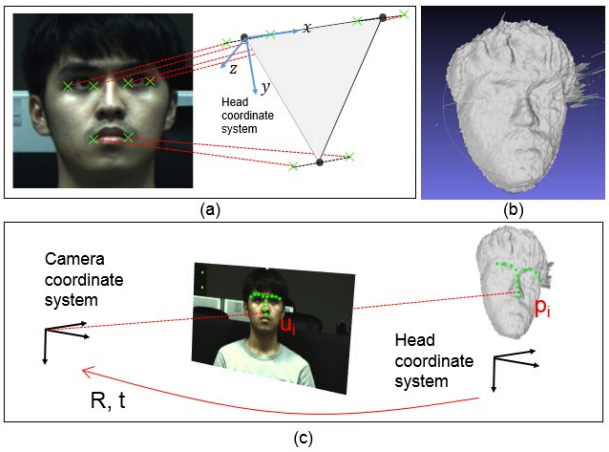

Fig. 3. (a) Definition of the head coordinate system from UT dataset. It is defined based on the triangle connecting three midpoints of the eyes and mouth. (b) The 3D face model built by Li et al. is used in this work. (c) The illustration of 2D-to-3D pose estimation.

the 2D-3D landmark point correspondences. To obtain the 3D facial landmark points, we used a 3D textured face model built by using the method proposed by Li et al. [14], as shown in Fig. 3. Then, we apply the EPnP algorithm [15] to estimate the 3D head pose from a set of correspondences between 3D points defined in the head coordinate system and their corresponding 2D image coordinates.

C. CNN Architecture

We use the CNN model to learn the regression function which is the mapping from input data (eye image as features and 2D head pose angles deciding the cluster of data for model) to gaze angles. VGG is used as a base net for our CNN architecture [12]. The characteristics of the VGG network architecture are the large number of weight layers and small 3×3 receptive fields throughout the whole net. Using small convolution kernels can effectively reduce the number of parameters. Also, incorporating more rectification layers makes the decision function more discriminative.

The details of our network design are shown in Fig. 4. The input to our network is a fixed-size 36×60 gray-scale eye image. The network consists of four modules. Previous three modules consist of two convolutional layers followed by one max pooling layer. The last module consists of two fully connected layers. For all of the convolutional layers, the kernel size is 3×3 and the convolutional stride is fixed to 1 pixel. In addition, all convolutional layers are equipped with the nonlinear rectification (ReLU [16]) function. The max pooling layer is performed over a 2×2 pixel window and the stride is fixed to 2 pixels. In the last module, the first fully connected layer has 200 channels and the following

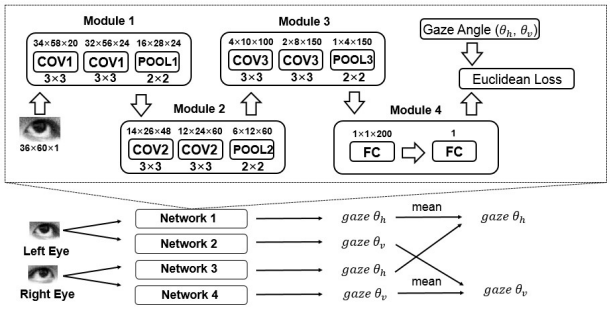

Fig. 4. The left and right eye regions are preprocessed and sent into the CNN network that is trained for estimating the horizontal and vertical angles for the gaze direction. For each network, VGG is the base net for our CNN architecture.

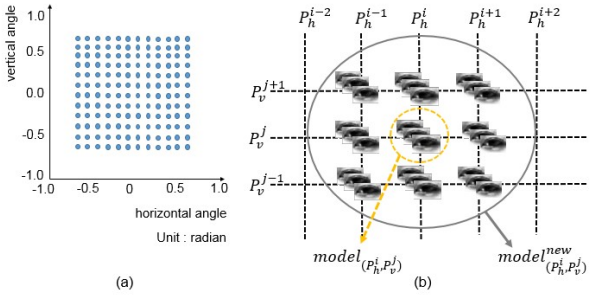

Fig. 5. (a) The distribution of head angles for UT dataset. The x-axis represents horizontal angle and the y-axis represents vertical angle. The unit of angle is radian in the diagram. (b) Instead of taking the eye image set with individual head pose, we take the eye image set with adjacent head poses to learn each of the multiple CNN models.

layer output one value for gaze angles which is calculated by summing up all activation values. For the loss function, we use the Euclidean loss that measures the distance between the predicted gaze angles and the ground truth.

To predict the horizontal and vertical gaze angles, we train two networks for the left and right eyes, respectively. Because both eyes share the same gaze angles, the mean of the estimated gaze angles from both eyes is taken as the final gaze estimation result.

We use UT Multiview dataset as the training dataset for our CNN networks. Fig. 5(a) shows the distribution of head poses in UT Multiview dataset. The head pose range in the dataset is larger than those of the state-of-art works. The range of horizontal angle is 66 degrees, from left 36 degrees to right 30 degrees, and the range of vertical angle is 66 degrees, from bottom 30 degrees to top 36 degrees. Each of the horizontal and vertical angle ranges is divided into 6-degree intervals. Fig. 5(b) illustrates how we cluster eye images based on the associated head poses to learn the gaze regression networks. There are H horizontal head pose P_h^i and V vertical head pose P_v^j, where i and j are indexes for horizontal and vertical head pose angles.

After training the multi-CNN models by using the eye image set with local head poses, the estimated 3D head pose is used to select the neighboring CNN models to predict the gaze angles with bilinear interpolation.

III. EXPERIMENTAL RESULTS

This section evaluates the gaze estimation accuracy of the proposed method on different datasets with comparison to the state-of-the-art methods. We randomly pick 30 gaze target samples from each person at different head poses from UT Multiview dataset for CNN training. Then, we conduct cross-dataset evaluation on two gaze datasets: MPIIGaze [9] and ours. For our own dataset, we collect data under the scenario of watching smart TV which can be controlled by gaze. We also compare the proposed method with baseline methods and different variants of our system.

A. Data Collection

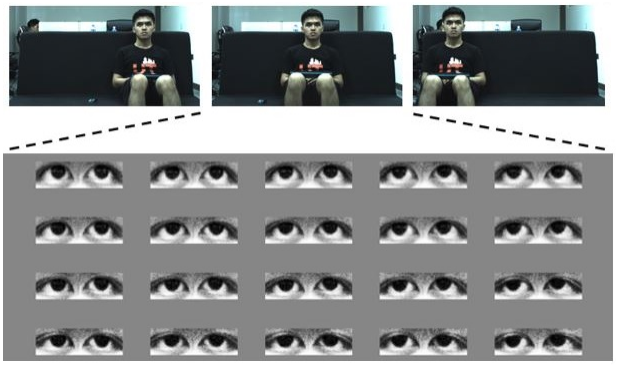

Fig. 6. Our data collection is based on the living room scenario. 76 participants were instructed to look at 20 visual targets with three different head poses.

Our data collection is based on the setting of watching a smart TV. In order to capture clear eye images at a distance, the system is equipped with a PointGrey Flea3 8.8-megapixel color camera with the image resolution at 4096×2160 pixels. The camera focal length is fixed to 9.7 mm. The camera is placed under a 42-inch BENQ TV (523.0 mm×929.7 mm).

A total of 60 (15 female and 45 male) people participated in our data collection and some of them had two rounds of data acquisition, i.e. with and without wearing glasses. During the video acquisition, participants were instructed to look at a visual target displayed on the monitor. Based on the scenario of watching TV, we design three viewing scenarios with different head poses: sit left position and look rightward, sit middle position and look to the front, sit right

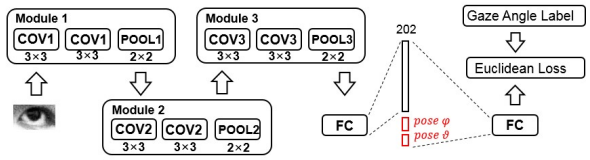

Fig. 7. A single network architecture is inspired by the multimodal CNN model.

position and look leftward (Figure 6). And the head pose range for the horizontal angle is nearly 33 degrees. In spite of three kinds of head pose, every head pose is different without using a chin rest on participants. There are 20 visual points arranged with a 4×5 pattern on the monitor. Every visual point is automatically displayed from top left to bottom right. The camera is triggered when the circle turns right light and finished until it turns yellow light. Therefore, there are 20 (gaze directions)×3 (head poses)×76 (subjects) eye images in our dataset.

B. Cross-Dataset Experiment

For experiments on our dataset, we compare the proposed multi-CNN method with some different methods, including different deep learning methods and the methods based on support vector regression (SVR). Schneider et al. [17] proposed different combinations between the regression methods and image features, and SVR with a concatenated vector of HOG [18] and LBP [19] features achieved the best performance for gaze estimation. On the selection of features as suggested in [17], we implement multi-level HOG [20] which concatenates different block permutations. Then, a concatenation of the HOG features and uniform LBP features is used as the input feature vector for the SVR model. Besides, we evaluate all the gaze estimation methods using the same facial landmark detection and head pose estimation.

We implement the single-CNN model and multi-SVR model with the same gaze interpolation scheme for comparison with our multi-CNN model. For the single CNN models, inspired by the multimodal CNN [22], we concatenate the head pose vector consisted of horizontal and vertical angles into blobs from second-last fully connective layer (Figure 7). To compare different architectures, we also use LeNet network similar to [9] to train a single CNN model. For multi-SVR models, we train each of the SVRs from the data of local head pose region and implement gaze interpolation on the results from SVR models of the neighboring head poses in the testing phase.

Figure 8 shows a comparison with different methods. The proposed multi-CNN method provides the highest accuracy and has the mean errors (7.75 $\pm$ 1.97 degrees), which is lower than (8.03 $\pm$ 1.88 degrees) by using multi-SVR models. Besides, the accuracy of VGG-based CNN model

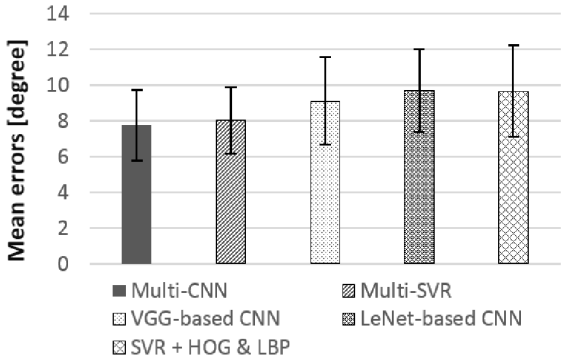

Fig. 8. Our method is compared with support vector regression (SVR) and the single CNN model. Also, we implement the multi-SVR method whose structure is similar to our proposed multi-CNN model for comparison.

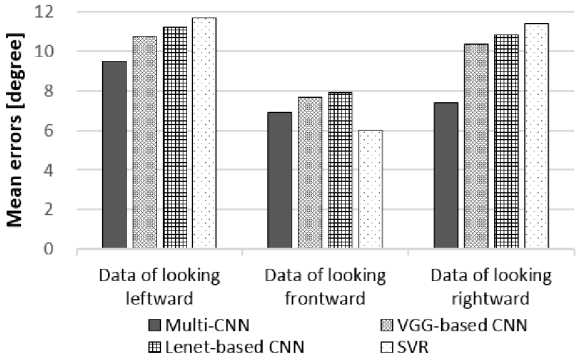

Fig. 9. Our proposed method is compared with the single CNN model and support vector regression (SVR) under three viewing scenarios.

is slightly better than the LeNet-based. And our proposed method shows a relative 14.7% improvement compared to the single CNN method, which indicates better utilization of head pose information.

In Figure 9, we analyze three mean errors from three clusters of our testing data with different main head poses, including looking leftward, looking to the front and looking rightward. For the case of looking frontward, the SVR method gives the best accuracy. Since CNN methods are sensitive to the quality of the images, eye images with uncommon eye shapes or low quality, such as squinting eyes, will cause higher errors of CNN methods. For the cases of looking leftward and looking rightward, the multi-CNN model effectively reduces the errors and achieves the highest accuracy. It indicates that the proposed method improves the accuracy of gaze estimation under head pose variations. Besides, it only costs about 0.043 seconds in average for

Method	Error
CNN with UT Multiview [9]	13.9
CNN with UnityEyes Synthetic Images [11]	11.2
kNN with UnityEyes [10]	9.9
multi-CNN with UT Multiview	**8.6**
CNN with UnityEyes Refined Images [11]	7.8

Table I. Comparison of the proposed method with previous methods for the gaze estimation experiment on the MPIIGaze dataset, measured in mean degree error.

gaze estimation with MATLAB implementation under a PC equipped with 3.4 GHz CPU, 16 GB memory, and GeForce GTX 750 Ti.

Table I gives a comparison among the state-of-the-art gaze estimation methods on the MPIIGaze dataset [9], which poses a practical task for gaze estimation since the dataset is collected in the wild. Compared to other methods that were also trained on synthetic images, the proposed method outperforms most of the state-of-the-art methods. Although [11] achieved 7.8 degrees in mean error, they used real images to refine synthetic images and the model was trained on more than 1.2M images. In contrast, the proposed multi-CNN system only used 13,500 images from UT Multiview dataset to train a single CNN model, but the trained models were tested on images from the MPIIGaze dataset. This experiment demonstrates the robustness of the proposed gaze estimation system through the cross-dataset experimental validation.

C. Cross-Subject Experiment on UT Multiview Dataset

We compare the proposed method on UT Multiview dataset with a few baseline methods reported in [21]. Based on their experiment setting, the synthesized images are used as training data and the acquired images are used as testing data. The cross-subject errors are evaluated by three-fold cross validation and synthesized training data of 33 different subjects. As shown in Fig. 10, the proposed method further improves the accuracy and achieves the lowest error compared to the previous methods.

IV. CONCLUSION

In this paper, we presented an appearance-based gaze estimation system that is robust against head pose variations. Our system is person-independent since we use UT Multiview dataset as the training data which includes a large number of participants. Through this dataset, multi-CNN models were trained and the gaze estimates from local head poses are fused to provide the final gaze estimation. Instead of training a single CNN model with or without head pose information, the proposed multi-CNN method provides

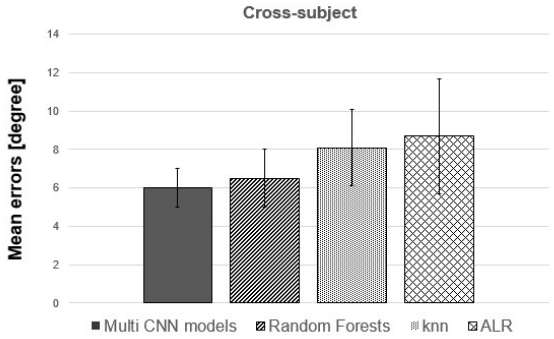

Fig. 10. The proposed method is evaluated by using cross-subject experiment on UT Multiview dataset. The accuracy of the proposed method is compared with those of the baseline methods reported in [21].

higher accuracy. Further, the proposed method demonstrates its practical value for use under head pose variations as we justified in cross-dataset experiments. In addition, the output is obtained by interpolating gaze estimates provided by the multi-CNN models corresponding to adjacent head poses. Thus, it further improves the accuracy of gaze estimation from face images with head pose variations.

REFERENCES

[1] J. Nielsen, K. Pernice. *Eyetracking Web Usability*, Berkeley, CA: New Riders Press, 2009.

[2] J. Nielsen, K. Pernice. *How to conduct eyetracking studies*, Nielsen Norman Group, 2009.

[3] 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.

[4] C. H. Morimoto and M. R. Mimica, "Eye gaze tracking techniques for interactive applications", *Comput. Vi. Image Understand., Special Issue on Eye Detection and Tracking*, vol. 98, no. 1, pp. 4-24, 2005.

[5] Asteriadis, S., Tzouveli, P., Karpouzis, K., & Kollias, S. (2009). Estimation of behavioral user state based on eye gaze and head pose—application in an e-learning environment. *Multimedia Tools and Applications, 41*(3), 469-493.

[6] J. P. Rae, W. Steptoe, and D. J. Roberts, "Some Implications of Eye Gaze Behavior and Perception for the Design of Immersive Telecommunication Systems", *2011 IEEE/ACM 15th Int. Symp. Distrib. Simul. Real Time Appl.*, pp. 108-114, 2011.

[7] A. Nakazawa and C. Nitschke, "Point of gaze estimation through corneal surface reflection in an active illumination environment", *in Proc. ECCV*, pp. 159-172, 2012

[8] Huang, Q., Veeraraghavan, A., & Sabharwal, A. (2015). TabletGaze: unconstrained appearance-based gaze estimation in mobile tablets. *arXiv preprint arXiv:1508.01244*.

[9] ZHANG, Xucong, et al. Appearance-based gaze estimation in the wild. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2015. p. 4511-4520.

[10] Wood, E., Baltruaitis, T., Morency, L. P., Robinson, P., & Bulling, A. (2016, March). Learning an appearance-based gaze estimator from one million synthesised images. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications* (pp. 131-138). ACM.

[11] Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., & Webb, R. (2016). Learning from simulated and unsupervised images through adversarial training. *arXiv preprint arXiv*:1612.07828.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", *CoRR*, abs/1409.1556, 2014.

[13] De la Torre, F., W.-S. Chu, X. Xiong, F. Vicente, X. Ding and J. Cohn, "Intraface", *11th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2015.

[14] C175. Y. Lee, J. Chen, C. W. Tseng and S.-H. Lai, "Accurate and robust face recognition from RGB-D images with a deep learning approach", *Proc. of British Machine Vision Conference*, 2016.

[15] 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.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks", *in NIPS*, pp. 1097-1105, 2012.

[17] T. Schneider, B. Schauerte, and R. Stiefelhagen, "Manifold alignment for person independent appearance-based gaze estimation", *in ICPR*, 2014.

[18] N. Dalal and W. Triggs, "Histograms of Oriented Gradients for Human Detection", *in CVPR*, 2004.

[19] D. C. He and L. Wang, "Texture Unit, Texture Spectrum And Texture Analysis", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 28, no. 4, pp. 509-512, 1990.

[20] F. Martinez, A. Carbone, and E. Pissaloux, "Gaze estimation using local features and non-linear regression", *in ICIP*, 2012.

[21] Y. Sugano, Y. Matsushita, and Y. Sato, "Learning-by-synthesis for appearance-based 3d gaze estimation", *in Proc. CVPR*, pages 1821-1828, 2014.

[22] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning", *in Proc. ICML*, pages 689-696, 2011.