ESTIMATION OF GAZE REGION USING TWO DIMENSIONAL PROBABILISTIC MAPS CONSTRUCTED USING CONVOLUTIONAL NEURAL NETWORKS

Sumit Jha and Carlos Busso

Multimodal Signal Processing (MSP) Laboratory, Department of Electrical Computer Engineering
The University of Texas at Dallas, Richardson TX 75080, USA
Sumit.Jha@utdallas.edu, busso@utdallas.edu

ABSTRACT

Predicting the gaze of a user can have important applications in *human computer interactions* (HCI). They find applications in areas such as social interaction, driver distraction, human robot interaction and education. Appearance based models for gaze estimation have significantly improved due to recent advances in *convolutional neural network* (CNN). This paper proposes a method to predict the gaze of a user with deep models purely based on CNNs. A key novelty of the proposed model is that it produces a probabilistic map describing the gaze distribution (as opposed to predicting a single gaze direction). This approach is achieved by converting the regression problem into a classification problem, predicting the probability at the output instead of a single direction. The framework relies in a sequence of downsampling followed by upsampling to obtain the probabilistic gaze map. We observe that our proposed approach works better than a regression model in terms of prediction accuracy. The average mean squared error between the predicted gaze and the true gaze is observed to be 6.89° in a model trained and tested on the MSP-Gaze database, without any calibration or adaptation to the target user.

Index Terms— Convolutional neural networks, gaze estimation, regression by classification

1. INTRODUCTION

Gaze is an important communicative cue in studying human interaction, playing a key role in understanding what he/she is attending to [1], who he/she is interacting with [2] or what he/she is interested in [3]. Estimation of gaze can be helpful in analyzing visual attention [4], creating smart interfaces [5], and predicting the engagement level of an interlocutor during a conversation [6]. While commercial gaze detection systems have achieved very accurate performance in a controlled environment, there are still open challenges when the intended system does not use any user calibration, or invasive equipments. Our study explores a novel deep learning architecture that not only estimates the gaze direction without user calibration, but also provides a gaze map describing the probability that the user is looking at different locations.

With the advancement in the development of deep learning algorithms, especially *convolutional neural networks* (CNNs), various algorithms have been developed to predict gaze location from the face image [7–9]. Most implementations to predict gaze rely on regression techniques, which predicts the horizontal and vertical position of the visual target. This paper proposes an alternative approach where the location is directly predicted in a 2D space (i.e., regression by classification). We use CNNs without fully connected layers such that each layer represents an image describing gaze direction with different resolutions. First, the model follows a sequence of max-pooling which reduces the resolution of the predicted image, capturing the key discriminative features needed to predict the gaze direction. Then, the model relies on a sequence of upsampling to obtain the final output image, which, in this study, is set to 48 × 24 pixels. This image provides a probabilistic map for the gaze.

We evaluate the proposed approach with the MSP-Gaze corpus [10], selecting the gaze direction in the gaze map with the highest probability. With this approach, we obtain a model that predicts the gaze with an average error of 6.89° in a user independent setting without any calibration or adaptation applied to the target user. This result is not only significantly better than the results achieved by a conventional regression model based on CNNs, but also provides extra information by predicting confidence regions obtained with the probability map.

2. RELATED WORKS

Algorithms for gaze estimation from images have seen major advancements in recent years. Early studies on predicting gaze estimation used intrusive equipments such as chin rests and head mounted devices, achieving very high gaze location estimation [11,12]. Some methods used active IR illumination [13–15]. While earlier studies were more focused on model-based approaches [13, 16, 17], recent efforts have focused on appearance based approaches [7–10], leveraging recent advancement in computer vision. A comprehensive review of eye tracking techniques can be found in Hansen and Ji [18] and Rahayfeh and Faezipour [19]. This section reviews studies on appearance based models.

Appearance-based methods directly use the eye image to predict the gaze. Some of these methods are purely based on machine learning, predicting the gaze location from the image without estimating landmarks on the eyes [5, 9, 20]. Li and Busso [5] proposed an appearance based algorithm based on *principal component analysis* (PCA). An image of both eyes was projected into the first 30 principal components. They trained linear regression models to learn the horizontal and vertical gaze pixels. They improved this model for user independent gaze estimation with similarity measures [21] and tensor based formulations [10].

More recently, researchers have attempted to use deep learning methods to implement appearance based methods for gaze estimation. This approach requires large naturalistic databases with annotated ground truth gaze labels. Zhang et al. [22] presented the MPIIGaze database, where multiple users contributed to the data collection over several months via naturalistic laptop use. This database contains 213,659 faces of 15 users (six females and five with glasses). Mora et al. [23] presented the EYEDIAP database with 3D recording targets in a 3D environment. They included

This work was supported by Semiconductor Research Corporation (SRC) / Texas Analog Center of Excellence (TxACE), under task 2810.014.

static, and naturalistic gaze patterns with free head movements. Wood et al. [24] presented a large dataset of rendered eye images that can be used to pre-train networks for robust gaze estimation. Li and Busso [5] collected the MSP-Gaze corpus where 44 users were asked to look at various target locations on a monitor screen. Using these databases, various algorithms have been proposed for appearance based gaze estimation. Zhang et al. [7] proposed a CNN-based architecture for gaze prediction, which was compared with other machine learning algorithms (random forest, K-nearest neighbors, support vector regression, and iris edge detection). They tested the algorithm on the MPIIGaze and Eyediap corpora, showing that their CNN-based method performed better than alternative methods. Krafka et al. [8] proposed a deep model that trains in parallel (1) images from the left eye, (2) images from the right eye, and (3) images from the entire face. They used data collected with an iOS application. Zhang et al. [9] argued that using the entire face for gaze prediction is better than using the eye region alone. A mask was trained, which learned the importance of different facial regions for gaze detection.

Most CNN based architecture have used fully connected layer before the output layer. Then, the activations are combined to deduce a regression loss. The key difference in our method is that it formulates the regression problem as a classification problem. We design a network entirely based on convolutional layers, where the goal is to perform classification on discretized labels, creating a useful probabilistic visual map for the gaze. Our proposed method can be implemented on top of most of the proposed method for gaze estimation based on CNN.

3. PROPOSED MODEL

The gaze location is spatially related to the eye image. The target gaze location is determined by the location of the pupil in the image and the head orientation of the subject. To capture these relations, our proposed approach relies on CNN, which has achieved state-of-the-art performance in encoding spatial information from images. However, our approach is different from conventional gaze methods, which estimate the direction of the gaze using regression. Our method estimates a 2D visual map describing the probability of the gaze direction, formulating the problem as a classification problem. The model follows a sequence of max-pooling followed by a sequence of upsampling to obtain a probabilistic map describing the gaze. The approach applies convolution on the image multiple times to obtain a grid that describes the probability of the gaze in different discretized region. Li and Busso [10] suggested that using an image including both eyes provides information about the head rotation. Therefore, our approach also takes an image showing both eyes of the subject as input.

3.1. Proposed Model Architecture

The proposed method is inspired by the model presented by Jha and Busso [25], where a probabilistic model of the gaze was learned from the head pose of the user in a naturalistic driving environment. Notice that this framework is quite different from the proposed method, since it did not use any image of the eyes or the face, only the information about the location and orientation of the head (6D vector).

Figure 1 describes the architecture of our proposed model. Table 1 provides the details of the architecture. Our approach estimates a patch containing both eyes of the user (Fig. 1). The *region of interest* (ROI) is scaled to 100× 25 pixels before sending it as an input to the network. In the first half of the network, we successively downsample the image to retain only the discriminative information. Every convolutional layer has 16 filters of size 3×3. After two convolutional layers, we reduce the dimension of the image with a max-pooling layer. We repeat this process three times such that the output resolution is 12×3 pixels. After reaching this resolution, we successively upsample the image. In the first layer, we only upsample the y axis such that the resulting output is 12×6 pixels. This step is carried out to change the aspect ratio of the images, bringing it closer to the required aspect ratio imposed by the application, in our case, the resolution of the monitor (1680×1050 pixels, Sec. 3.2). The 12×6 images are then processed with convolutional layers and upsampling, increasing their resolution to get the final output of size 48×24. At the final layer, a single image of size 48×24 is obtained which is passed through softmax layer to obtain a probability distribution. While the architecture is flexible to obtain a higher resolution, we stop at this point to limit the depth and complexity of the model.

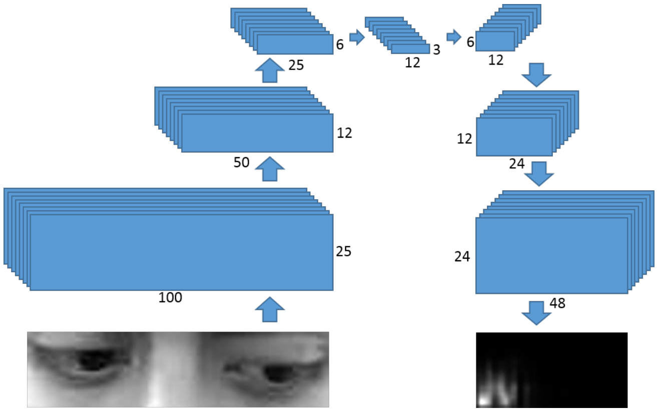


Fig. 1. Block diagram of our proposed framework. The architecture relies on max-pooling layers followed by upsampling layers to obtain the probabilistic gaze map with CNNs.

Table 1. Proposed deep learning architecture relying of max-pooling and upsampling layers with CNNs.

Layer	Spec	Activation	Dropout	Output Dimension
Conv2D	**16, 3x3**	**ReLU**	**0**	**16 x 100 x 25**
Conv2D	**16, 3x3**	**ReLU**	**0.25**	**16 x 100 x 25**
MaxPool2D	**2x2**	**-**	**-**	**16 x 50 x 12**
Conv2D	**16, 3x3**	**ReLU**	**0**	**16 x 50 x 12**
Conv2D	**16, 3x3**	**ReLU**	**0.25**	**16 x 50 x 12**
MaxPool2D	**2x2**	**-**	**-**	**16 x 25 x 6**
Conv2D	**16, 3x3**	**ReLU**	**0**	**16 x 25 x 6**
Conv2D	**16, 3x3**	**ReLU**	**0.25**	**16 x 25 x 6**
MaxPool2D	**2x2**	**-**	**-**	**16 x 12 x 3**
Conv2D	**16, 3x3**	**ReLU**	**0**	**16 x 12 x 3**
Conv2D	**16, 3x3**	**ReLU**	**0.25**	**16 x 12 x 3**
UpSampling2D	**1x2**	**-**	**-**	**16 x 12 x 6**
Conv2D	**16, 3x3**	**ReLU**	**0**	**16 x 12 x 6**
Conv2D	**16, 3x3**	**ReLU**	**0.25**	**16 x 12 x 6**
UpSampling2D	**2x2**	**-**	**-**	**16 x 24 x 12**
Conv2D	**16, 3x3**	**ReLU**	**0**	**16 x 24 x 12**
Conv2D	**16, 3x3**	**ReLU**	**0.25**	**16 x 24 x 12**
UpSampling2D	**2x2**	**-**	**-**	**16 x 48 x 24**
Conv2D	**1, 3x3**	**Softmax**	**-**	**1 x 48 x 24**

Some spatial component of the data is lost when we pool the image reducing the resolution to 12×3. However, the gaze output lies on a low dimensional space. Therefore, it is important to reduce the resolution, removing non-discriminant information from the image. Pooling also helps in increasing the receptive field of the network. While methods such as dilated convolution [26] helps increasing the receptive field without pooling, these methods are sensitive to the initialization of the parameters.

The loss functions generally used for classification problems such as the cross entropy loss are not cost sensitive. For our formulation, this is a problem since the cost of making an error of 1 pixel in the estimation would be the same as making an error of 10 or more pixels. For our purpose, we expect lower cost for an estimation that is spatially closer to the ground truth value. For this purpose, instead of using the ground truth labels as one hot encoded outputs, we use a Gaussian distribution around the true label. The distribution adds a weighted reward for choosing any value close to the target gaze. We also apply the Gaussian mask to the final layer of the network, so the output of the model matches the mask of the ground truth gaze. Then, we use cross entropy loss as the loss function. While the Gaussian mask is used on the estimation of the loss function computation, the final gaze map produced by our model corresponds to the prediction before applying the mask, which reduces the area of the probabilistic gaze map.

We use Keras [27] on top of Tensorflow [28] to train our model. We use a learning rate of 0.001, training the network for 150 epochs.

3.2. Database

We use the MSP-Gaze [5, 10] database to train our network. This database was collected by asking subjects to click on points appearing on a monitor at random locations. The database contains data collected from 44 subjects (21 males and 23 females). Subjects were chosen from multiple ethnic backgrounds representing the student population demographic at The University of Texas at Dallas (Caucasian, Asian, Indian and Hispanic). The data was collected in two different sessions across different days to capture session variability. The protocol includes natural gaze actions. It also included controlled recordings where the subjects were asked to avoid moving their head. The protocol also includes different distances between the subject and the screen. For our network, the inputs are 100×25 ROI images including both eyes, and the ground truth labels are the target pixel locations in a 1680×1050 resolution screen. In total, we have 324,771 gaze directions associated with the ROI images. The details of this corpus are provided in Li and Busso [10].

The evaluation considers user-independent partitions. We divide the corpus into seven partitions, where data from each subject is exclusively included in one of the partitions. We repeat the experiment seven times using one partition as the test set. The remainder six partitions are used to train the models. The reported results correspond to the average results obtained across the seven folds.

3.3. Baseline Model

We also train a basic regression model with six convolutional layers and two fully connected layers to compare the performance of the proposed approach. Table 2 explains the model architecture of the baseline method. The loss function corresponds to the *mean squared error* (MSE) between the predicted and true gaze. The model jointly predicts the horizontal and vertical directions of the gaze.

4. RESULTS

This section reports the experimental evaluation of our method. The evaluations considers comparisons with the baseline method (Sec. 4.1). It also evaluates the probabilistic map, creating confidence intervals containing the gaze directions (Sec 4.2).

Table 2. Architecture of the baseline regression model using CNNs.

Layer	Spec	Activation	Dropout	Output Dimension
Conv2D	**16, 3x3**	**ReLU**	**0**	**16 x 100 x 25**
Conv2D	**16, 3x3**	**ReLU**	**0.25**	**16 x 100 x 25**
MaxPool2D	**2x2**	**-**	**-**	**16 x 50 x 12**
Conv2D	**16, 3x3**	**ReLU**	**0**	**16 x 50 x 12**
Conv2D	**16, 3x3**	**ReLU**	**0**	**16 x 50 x 12**
Conv2D	**16, 3x3**	**ReLU**	**0.25**	**16 x 50 x 12**
MaxPool2D	**2x2**	**-**	**-**	**16 x 25 x 6**
Conv2D	**16, 3x3**	**ReLU**	**0**	**16 x 25 x 6**
Conv2D	**16, 3x3**	**ReLU**	**0**	**16 x 25 x 6**
Conv2D	**16, 3x3**	**ReLU**	**0.5**	**16 x 25 x 6**
MaxPool2D	**2x2**	**-**	**-**	**16 x 12 x 3**
Dense	**512**	**ReLU**	**0.5**	**1 x 512**
Dense	**128**	**ReLU**	**0.5**	**1 x 128**
Dense	**2**	**-**	**-**	**1 x 2**

Table 3. Angular error between the true and predicted gaze on the screen using the proposed approach and the baseline method.

	Proposed approach error [°]	**Regression model error [°]**
Mean	**6.89**	**10.28**
Median	**6.11**	**9.69**
95 percentile	**14.20**	**19.45**

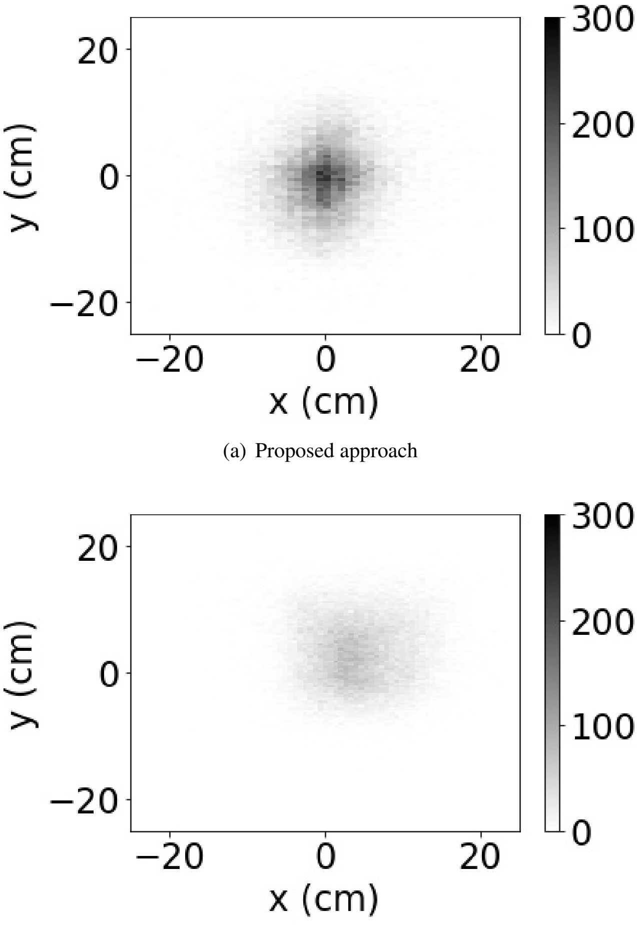
4.1. Comparisons with the Baseline

Our proposed model creates a probabilistic gaze map. We estimate the most likely point in the probabilistic visual map to directly compare the results between the regression model and our proposed approach. We estimate the mean, median and 95 percentile of the errors for each subject in the corpus. Then, we average the performance across subjects, reporting the results in Table 3. The error is measured in term of the angular error between the predicted and true gaze location (degree). The angular error was calculated by interpolating the size of the head to approximate the distance between the user and the screen, following the approach explained in Li and Busso [10]. Table 3 shows that our method performs significantly better than its regression counterpart (one-tail t-test over the errors across the 44 subjects, asserting significance at p-value = 0.001). The regression by classification formulation is effective for this task.

Figure 2 shows the vertical and horizontal errors for the predictions of our approach and the baseline method. For our approach, the errors are clustered around the center of the coordinate (i.e., low error). For the baseline model, the spread of the errors is higher. These results show the advantage of using our approach over a traditional regression method, even when both are implemented with CNNs.

4.2. Confidence Intervals: Accuracy versus Resolution

One of the advantages of using our proposed approach is that we can derive a probability density map instead of a single gaze direction as an output. With a probability map, we can create different confidence regions. If we include regions with low probabilities, the area of the confidence region will be big. This setting will result in high accuracy (i.e., the confidence region includes the target gaze direction), but low spatial resolution (i.e., the area of the confidence



(a) Proposed approach

(b) Baseline regression model

Fig. 2. Distribution of gaze estimation error.

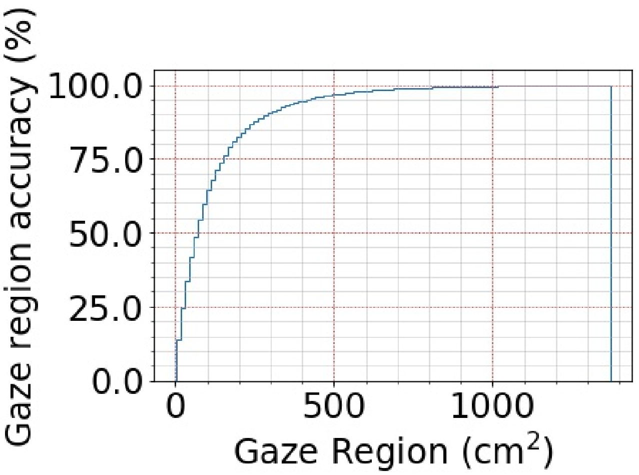
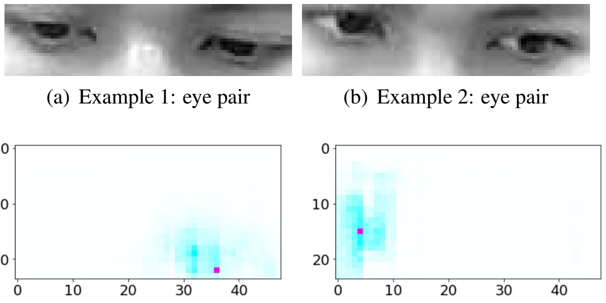


Fig. 3. Accuracy versus resolution of the predicted gaze region.



(a) Example 1: eye pair

(b) Example 2: eye pair

(c) Example 1: predicted gaze

(d) Example 2: predicted gaze

Fig. 4. Probabilistic gaze maps created by our proposed approach for two examples.

region is big). If we only include regions with high probability values, the accuracy will decrease (i.e. the region won't include the target point), but the spatial resolution will be better (i.e., smaller area).

Figure 3 analyzes the tradeoff between accuracy and spatial resolution. This figure is created by estimating the area of the smallest confidence region such that the region includes the target accuracy (e.g., 75% accuracy). For example, to include 95% of the data point, we need an area of $400cm^2$ (20cm×20cm). For a 75% confidence region, the area of the region is approximately 150 cm^2 (≈12cm×12cm). While the results are encouraging, there is room for improvement.

Figure 4 shows two examples for the probability distribution of the predicted gaze maps. The dot is the true prediction label, while the cyan region is the predicted distribution. We observe that the probabilistic maps are reasonably accurate. While for some examples the true label may not exactly overlap with the dense area of the predicted region, the outputs are still reasonably close to the actual gaze direction to be used in most practical applications.

5. CONCLUSIONS AND FUTURE WORK

This paper proposed a novel approach to predict the gaze of a user. Instead of separately performing regression, we formulate the problem as a regression by classification task. Our proposed architecture performs classification of gaze direction in a 2D space. The approach directly obtains not just a gaze direction, but also a probability map describing the gaze distribution. This approach provides a new formulation for solving the gaze estimation problem by providing confidence regions for the gaze. The results show that the accuracy of the predictions are better than the accuracy of a regression model of comparable depth.

While we stop the upsampling at a resolution of 48 × 24 in the model, it is possible to construct deeper models to obtain results at a higher resolution. We will investigate variations of the proposed formulation that can lead to better performance (i.e., higher accuracy, better spatial resolution). Another drawback of the model is that we decrease the resolution of the image by pooling before upsampling. While this step helps the network to filter out information that is not useful for gaze estimation, this approach might lead to loss of spatial information. We plan to investigate models with lateral (residual) connections [29], such that information at higher resolutions can be retained. We used the MSP-Gaze database for our experiments, since it includes several aspects that are important for gaze detection without calibration or intrusive equipments (i.e., variability in terms of users, distance to screen, head movements, and sessions). For our future work, we also plan to investigate the use of the proposed model trained with other popular databases such as the MPIIGaze corpus [22]. We are also interested in exploring the accuracy of this approach in challenging environments for gaze detection such as in vehicles during naturalistic driving conditions [25].

6. REFERENCES

[1] G. Underwood, P. Chapman, N. Brocklehurst, J. Underwood, and D. Crundall, "Visual attention while driving: sequences of eye fixations made by experienced and novice drivers," *Ergonomics*, vol. 46, no. 6, pp. 629–646, May 2003.

[2] T. Ohno, N. Mukawa, and A. Yoshikawa, "FreeGaze: a gaze tracking system for everyday gaze interaction," in *Symposium on Eye tracking research & applications (ETRA 2002)*, New Orleans, LA, USA, March 2002, pp. 125–132.

[3] K. Rayner, C.M. Rotello, A.J. Stewart, J. Keir, and S.A. Duffy, "Integrating text and pictorial information: eye movements when looking at print advertisements," *Journal of Experimental Psychology: Applied*, vol. 7, no. 3, pp. 219–226, September 2001.

[4] S. Jha and C. Busso, "Probabilistic estimation of the driver's gaze from head orientation and position," in *IEEE International Conference on Intelligent Transportation (ITSC)*, Yokohama, Japan, October 2017, pp. 1630–1635.

[5] N. Li and C. Busso, "Evaluating the robustness of an appearance-based gaze estimation method for multimodal interfaces," in *International conference on multimodal interaction (ICMI 2013)*, Sydney, Australia, December 2013, pp. 91–98.

[6] Y. I. Nakano and R. Ishii, "Estimating user's engagement from eye-gaze behaviors in human-agent conversations," in *Estimating User's Engagement from Eye-gaze Behaviors in Human-agent Conversations (IUI 2010)*, Hong Kong, China, February 2010, pp. 139–148.

[7] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, Boston, MA, USA, June 2015, pp. 4511–4520.

[8] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, Las Vegas, NV, USA, December 2016, pp. 2176–2184.

[9] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "It's written all over your face: Full-face appearance-based gaze estimation," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2017)*, Honolulu, HI, USA, July 2017, pp. 2299–2308.

[10] N. Li and C. Busso, "Calibration free, user independent gaze estimation with tensor analysis," *Image and Vision Computing*, vol. 74, pp. 10–20, June 2018.

[11] D. Scott and J.M. Findlay, *Visual Search, Eye Movements and Display Units*, IBM UK Hursley Human Factors Laboratory, 1991.

[12] V. Raudonis, R. Simutis, and G. Narvydas, "Discrete eye tracking for medical application," in *International Symposium on Applied Sciences in Biomedical and Communication Technologies*, Bratislava, Slovakia, November 2009, pp. 1–6.

[13] T. Ohno and N. Mukawa, "A free-head, simple calibration, gaze tracking system that enables gaze-based interaction," in *Symposium on Eye tracking research & applications (ETRA 2004)*, San Antonio, TX, USA, March 2004, pp. 115–122.

[14] B. Noureddin, P. D. Lawrence, and C. F. Man, "A non-contact device for tracking gaze in a human computer interface," *Computer Vision and Image Understanding*, vol. 98, no. 1, pp. 52–82, April 2005.

[15] D. Beymer and M. Flickner, "Eye gaze tracking using an active stereo head," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003)*, Madison, WI, USA, June 2003, vol. 2, pp. 451–458.

[16] S. Baluja and D. Pomerleau, "Non-intrusive gaze tracking using artificial neural networks," Tech. Rep. CMU-CS-94-102, Carnegie Mellon University, Pittsburgh, PA, USA, January 1994.

[17] D.W. Hansen and A.E.C. Pece, "Eye tracking in the wild," *Computer Vision and Image Understanding*, vol. 98, no. 1, pp. 155–181, April 2005.

[18] D. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 478–500, March 2010.

[19] A. Al-Rahayfeh and M. Faezipour, "Eye tracking and head movement detection: A state-of-art survey," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 1, no. 1, pp. 2100212 –2100212, 2013.

[20] Y. Sugano, Y. Matsushita, Y. Sato, and H. Koike, "An incremental learning method for unconstrained gaze estimation," in *Computer Vision ECCV 2008*, D. Forsyth, P. Torr, and A. Zisserman, Eds., vol. 5304 of *Lecture Notes in Computer Science*, pp. 656–667. Springer Berlin Heidelberg, Marseille, France, October 2008.

[21] N. Li and C. Busso, "User-independent gaze estimation by exploiting similarity measures in the eye pair appearance eigenspace," in *International conference on multimodal interaction (ICMI 2014)*, Istanbul, Turkey, November 2014, pp. 335–338.

[22] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "MPIIGaze: Real-world dataset and deep appearance-based gaze estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 162–175, January 2019.

[23] K.A.F. Mora, F. Monay, and J.-M.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 (ETRA 2014)*, Safety Harbor, FL, USA, March 2014, pp. 255–258.

[24] E. Wood, T. Baltrušaitis, X. Zhang, Y. Sugano, P. Robinson, and A. Bulling, "Rendering of eyes for eye-shape registration and gaze estimation," in *IEEE International Conference on Computer Vision (ICCV 2015)*, Santiago, Chile, December 2015, pp. 3756–3764.

[25] S. Jha and C. Busso, "Probabilistic estimation of the gaze region of the driver using dense classification," in *IEEE International Conference on Intelligent Transportation (ITSC 2018)*, Maui, HI, USA, November 2018, pp. 697–702.

[26] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *International Conference on Learning Representations (ICLR 2016)*, San Juan, Puerto Rico, May 2016, pp. 1–13.

[27] F. Chollet, "Keras: Deep learning library for Theano and TensorFlow," https://keras.io/, April 2017.

[28] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D.G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: A system for large-scale machine learning," in *Symposium on Operating Systems Design and Implementation (OSDI 2016)*, Savannah, GA, USA, November 2016, pp. 265–283.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, Las Vegas, NV, USA, June-July 2016, pp. 770–778.