Low-cost Gaze Detection with Real-time Ocular Movements Using Coordinate-Convolutional Neural Networks

Shubham Jain, Enda Fallon

Software Research Institute, Athlone Institute of Technology, Athlone, Ireland

Email: sjain@ait.ie; efallon@ait.ie

***Abstract* - Detection of ocular-movements unfolds various possibilities in computer vision but requires large datasets, expensive hardware and computational power. Prior research substantiates the belief that Convolutional Neural Network provides the highest recognition rate compared to traditional techniques, but they begin to overfit after achieving a certain accuracy due to the coordinate-transform-problem. Different image conditions like variation-in-viewpoint or illumination can be pragmatic for image processing and require on-device calibration. This paper proposes a framework that works with low-computational-complexity in varied environmental conditions to provide efficient gaze estimations that points out screen coordinates in real-time. We use a depth-wise convolution, an expansion and a projection layer along-with coordinate-channels to improve classification. The model is experimented against different environmental conditions, multiple subjects, image augmentation and different data sizes in real-time to estimate the coordinate classes using eye-movements on a standard web camera, yielding better accuracy and preventing overfitting of model with fewer hardware requirements.**

Keywords - Gaze detection, Convolutional Neural Networks, Deep Learning

I. INTRODUCTION

Gaze detection system have been proposed as a means of identifying visual cues in human computer interaction systems. Even after extensive research on desktop-based gazing applications, the state-of-the-art accuracies only predict limited classes. With the advancement of eye tracking techniques in deep learning, gaze detection has achieved extensive application prospect in medical, education, targeted advertisement marketing, manufacturing interfaces and activity recognition [1]. Activity recognition has been the main area of research after successful implementation of eye movement tracking using deep learning techniques. Gaze detection aids activity recognition by analysis of ocular movements to obtain information that relates to users point of attention. Such information can further aid marketing techniques, creation and presentation of content for individual users based on their activity statistics and point of attention, providing relevant advertisements with information gathered from activity recognition, collecting cues from eyes to detect interest level of students in education field, preventing production failures by providing gaze detection-based validation of inputs on manufacturing interfaces.

Convolutional Neural networks (CNNs) gained much attention in the area of eye movement tracking as it avoids complex preprocessing and image can be used directly as an input to the tensor, without manual feature extraction. CNNs have been significantly achieving state-of-the-art accuracies in image classification by adopting the use of SoftMax classifier which is efficient in learning ability as learning features are vital to the representation of images and more conducive to classification, and their usage on multi-dimensional inputs. Gaze detection in various applications require detection of (x, y) coordinates which may seem like a task Convolutional Neural Network would perform, but it surprisingly fails due to its inability to transform spatial representation between coordinates in Cartesian space and one-hot pixel space [2]. This results in overfitting of model after 86% accuracy in most image classification implementations, despite of providing it with an accurate labelled dataset. Reference [2] proposed the use of CoordConv which makes use of additional input channels with information related to coordinates which is implemented for gaze detection in this paper.

Gaze detection systems frequently becomes costly due to their requirement of high-quality cameras and complicated setups. The 2018 statistics for Mobile vs Desktop usage by reference [3] proved that end users prefer the use of Mobile devices compared to desktop due to ease of accessibility. Gaze analysis can be further utilized as input process for controlling handheld devices, enhanced safety features and efficient data analysis for traffic marketing [4]. Detecting ocular movements has always been challenging in unconstrained environments that has variation in illumination, background clutter, real-time efficiency against head movements and providing predictions with little or no calibration. Eliminating these challenges require high processing power which increases the cost. This paper implements a 3-layer Convolutional Neural Network architecture which is detailed in Section III that consists of an expansion layer, depth-wise layer and projection layer along with Coord input channels to provide low cost predictions with higher accuracy and less system complexities that works with minimal calibration on handheld devices. We examined how our model works in different conditions such as environmental factors, validation using different subjects and experimented with new techniques such as augmentation and training with more subjects to improve model accuracy. The experiments and results are illustrated in Section IV.

II. RELATED WORK

Various conventional methods for eye-movement tracking have been proposed, a review of those methods is provided in reference [5]. In this section we limit the literature discussion to recent state-of-the-art methods for gaze-detection in real-time. Table I provides a brief overview of the deep learning methods discussed in this section and their dataset size.

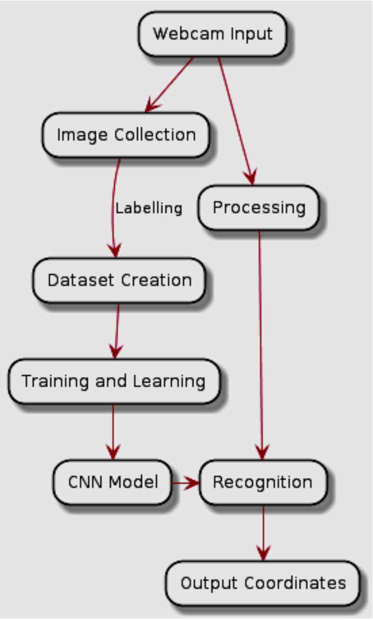

Figure 1. Flowchart of Research Activities.

Deep Learning has been useful to overcome challenges in image classification for eye movement tracking. The authors of [6] proposed the use of Convolutional Neural networks for gaze detection by combining the data obtained from facial posture area and eye region, deep features extracted from eye images are processed through multiple convolutions and pooling layers to generate predictions. They classify outputs on screen with their clustering algorithm which uses Random forest regression with minimized cross entropy loss. They also present another approach that works on detecting a gaze angle based on a regression model inspired by a LeNet-based CNN. The same authors in reference [7] represent a novel approach that utilized the entire facial region as input by allocating weights on separate regions allowing them to conceal or enhance information of facial regions. Their technique achieves higher accuracy in varied lighting conditions and robustness against head movements. Reference [8] on the other hand works on a 3D model for gaze tracking buy using two separate Convolutional neural networks with head pose and eye movement models, connected with a graze transform layer to provide state-of-the-art accuracy against free head movements.

TABLE I. DATASETS UTILIZED BY EXISITING METHODS

Citation	Image Resolution	Model	Size of Dataset
[1]	2x42x50	Two convolutional layers followed by a fully connected layer for each eye	1170 (40 Subjects with 7 classes)
[6]	60x36	Multiple convolution and pooling layers for facial pose and eye region. Based on LeNet architecture	MPII Gaze-213,659 (15 subjects)
[7]	448x448	Five convolutional layers, two fully connected layers. Linear regression layer combined with the last fully connected layer	MPII Gaze-213,659 (15 subjects)
[8]	62x62	AlexNet, two separate Convolutional neural networks with head pose and eye movement models, connected with a graze transform layer	24000 (200 Subjects)
[10]	2x60x36	Two Convolutional, two maxpooling and two fully connected layers for left and right eye.	MPII Gaze-213,659 (15 subjects)

Reference [1] describes a framework for gaze detection in real-time that is free from calibration even on unconstrained background conditions. Independent CNNs for each eye, trained separately using Stochastic Gradient Descent (SGD) algorithm [9] to classify gaze-directions from a set of 7 output labels using a dataset of 1170 images for training and validation. The framework achieved higher accuracy with less computational complexity and low-resolution cameras using pre-trained models. Finally, the authors of [10] experimented with two approaches, first where one network is utilized for both eyes, to determine gaze angle one of the eyes was flipped to measure the impact on accuracy and second was a modified version of reference [1] which makes use of dual eye channels to significantly increase accuracy. They also demonstrated the impact of data augmentation to overcome variation in distance in pre-trained models.

Existing literature mostly focuses on a three-step process: face recognition, eyes detection and classification of ocular movements. Almost all approaches follow either classification or regression outputs with a set of predefined labels in output class. The data required for training or calibration uses at least 1000 images from different subjects. The proposed method aims at gaining higher accuracy with less images in dataset on devices with low computation capabilities and unconstrained background conditions by predicting the x and y coordinates of gaze direction on the screen. The proposed method is discussed in later section.

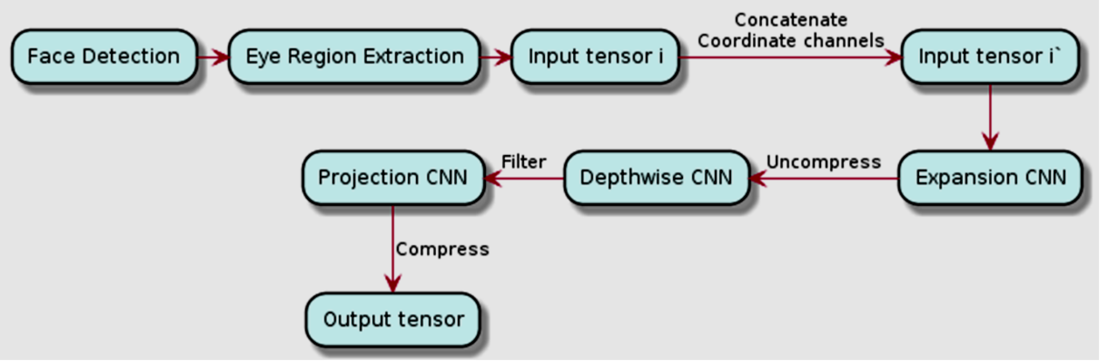

Figure 2. Framework for Gaze Detection.

III. RESEARCH METHODOLOGY

The flowchart of research activities is shown in Fig. 1 which contains various processes such as data gathering from webcam, labelling captured images, training and learning, model creation to recognize and detect gaze and plotting output coordinates.

The activities of database creation are manual and done using a web interface. User enables webcam and capture at least 50 samples that contains cropped eye region images with their respective x, y gaze coordinate points on screen. To obtain ocular region we have utilized Voila-Jones [11] method to detect frontal face regions as it is efficient and robust against head-rotations and then eye cascades to detect and crop eyes region. We then transform the obtained RGB image into a grayscale image and resize it to 30x60 matrix.

The eye region samples and coordinate values are then used in a classification framework for feature extraction and learning. We used a Convolutional Neural network based on MobileNetV2 [12] with 3 Convolution Layers followed by a fully connected layer, a global pooling layer and SoftMax classifier which would be discussed in this section for gaze detection. The proposed framework shown in Fig. 2. is a novel approach inspired by MobileNetV2 architecture that yields better results in gaze detection with the implementation of CoordConv. The first convolution layer is an expansion layer which is a 1x1 convolution layer responsible for expansion of number of channels in the tensor based on the expansion factor, we have used 6 as the default value for our expansion factor as recommended by the authors of [12]. The expansion layer increases the number of channels by creating new tensors based on the expansion factor. If the expansion factor is e and number of current channels is c, and new channels is n then total number of channels with new tensors can be calculate as:

$$n= e*c \qquad (1)$$

The second convolution layer, depth wise performs convolutions on individual channels to create an output filter equal to the number of channels that is followed by the third convolution layer, pointwise convolution layer which is responsible for dimensionality reduction from a tensor of n number of channels.

We feed the 30x60 image to the first convolutional layer with 24 filters of 7x7 dimension followed by batch normalization and activation. We wanted to accelerate the training of our network which was possible by limiting the covariate shift and normalizing the activations of each layer as presented by authors of [13]. They proposed that restricting the activations to be mean 0 and unit variance in each layer can allow the network to learn different parameters that can transform mean and variance to any value that network requires. So, we used batch normalization which allowed our network to increase training speed due to the stable distribution of inputs.

Each convolution layer is followed by batch normalization and activation except the projection layer. Since the output from the last layer is low dimensional, applying non-linearity damaged useful information as described by the authors of [12]. Rectified Linear Unit 6 (ReLu6) is used as an activation function to achieve non-linearity, the function of which can be denoted as:

$$f(x)= \min(\max(0, x), 6) \qquad (2)$$

where x is input and f(x) is the output that ranges from 0 to 6 in a ReLu unit. The authors of [12] compared the results of ReLu with ReLu6 and detected the robustness of ReLu6 for low precision computation on handheld devices. Our focus was to reduce computations which was possible by dimensionality reduction in the projection layer which reduced the size of a 7x7 output to 1x1. Another issue that can be resolved by the proposed architecture describe in Fig. 3 is the problem of overfitting in Convolutional Neural Networks caused by its inability to transform spatial representation between coordinates in Cartesian space and one-hot pixel space [2]. As seen in Fig. 2 and Fig. 3, we concatenated two input channels with our input tensor.

We used SoftMax function for multiclass classification and cross-entropy function for loss estimation. Cross-entropy the actual class output value with the predicted probability and assign a score between 0 and 1. The cross- entropy loss is reduced in training with the use of Stochastic Gradient Descent [14] algorithm.

Gaze detection systems should be invariant to ocular distance from image capturing device and varied illumination conditions, the architectural design was focused on accommodating the real-time distance of the subject on a low-cost image capturing device. We performed down-sampling to simulate information loss by relative distance. The authors of [10] proposed the use of random resizing to capture the range of desired distance rather than down sampling. On the other hand, our proposed architecture works well with down-sampling due to the use of expansion and projection layers.

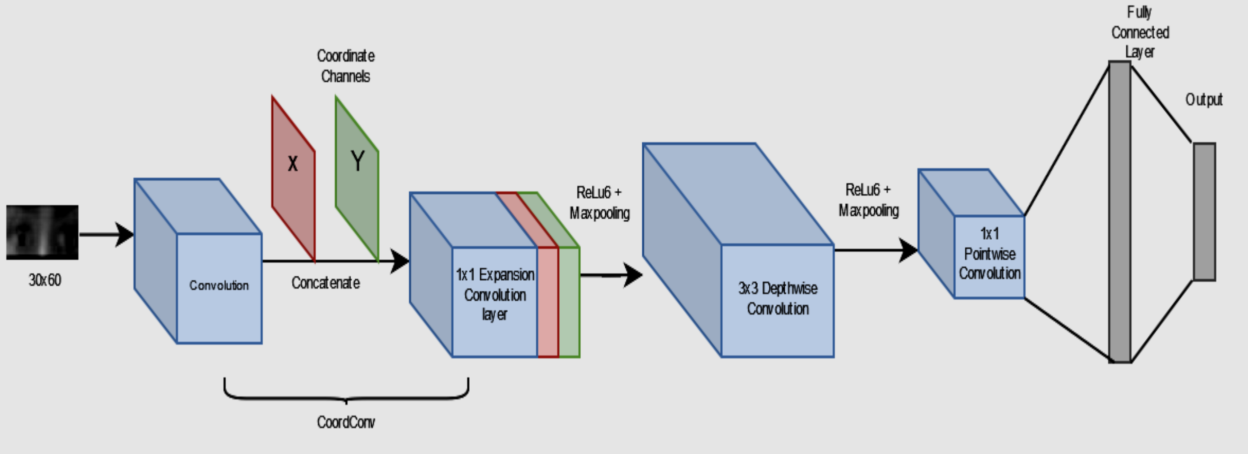

Figure 3. Architecture of proposed method for Gaze Detection.

IV. EXPERIMENTS AND RESULTS

In this section, we discuss and depict results of different experiments that had an impact on our model accuracy. We first evaluate the minimum number of labelled images required for training and then compare the results after applying augmentation to the training set. We compare the accuracy of our model against different environmental conditions that may have a higher impact on accuracy in the later sub-section. We also analyze the impact on accuracy when trained and validated with different number of subjects. Finally, after inputs from each experiment, we analyze our model and illustrate the results in a confusion matrix.

A. Dataset Evaluation

Experiments to determine how many labelled images are necessary for a better accuracy were conducted. We trained the model with different dataset sizes, applied image augmentation and compared the accuracy. The images were captured with one subject in same environmental conditions using a laptop webcam. The results are illustrated in Table II.

This paper proposes a method that achieves low computational complexity that results in accuracy trade-off, so we classified the output coordinates into eight equal quadrants Q to predict better results in limited regions rather than detecting the exact coordinate.

$$(x_i, y_i) \Rightarrow \{x \in Q_e{:}0< x< \max\} \text{ and } \{y \in Q_e{:}0< y< \max\}$$

where x, y are coordinates and Qe is the element of set Q, Q= {Top-Left (TL), Top-Middle-Left (TML), Top-Middle-Right (TMR), Top-Right (TR), Bottom-Left (BL), Bottom-Middle-Left (BML), Bottom-Middle-Right (BMR), Bottom-Right (TR)}.

TABLE II. DATASETS OF EXISITING METHODS

Dataset samples (Per class)	**Accuracy (%)**
2000	96.67
1000	96.12
500	95.82
250	93.87
100	92.37

The eight classes describe the region on the screen where gaze was detected and plotted by tracking ocular movements.

B. Image Augmentation to increase dataset

We also used data augmentation in our training sample to provide large amount of data for training. We performed scaling to provide predictions invariant to distance of subject from camera. We added salt and pepper noise (adding black and white dots on images) and blurring to images so that the model can learn from blurry input and perform classification and finally we applied various lightning conditions to our images to simulate the real-world scenarios. The accuracy after image augmentation with each technique were compared in different conditions and are illustrated in Table III. We used 100 samples for each class using one subject.

C. Model accuracy in different conditions

Environment factors such as bad illumination, background clutter, variation in distance, bad camera quality have a greater impact on results. We trained out model with 100 samples for each class after augmentation in different environment conditions.

TABLE III. COMPARISON OF ACCURACY AFTER IMAGE AUGMENTATION

Image Augmentation	**Accuracy before (%)**	**Accuracy after (%)**	**Test Condition**
Scaling	89.42	91.67	Distance variation
Salt and pepper noise	90.57	91.53	Variation in image quality
Blurring	90.57	91.45	Variation in image quality
Shading	87.77	90.87	Variation in illumination
Resizing	87.63	90.66	Distance, Illumination variation

Table IV illustrates the results in different conditions. The results generated were similar or with little difference which suggested that our architecture was robust against various environmental changes and required no calibration.

TABLE IV. AFFECTS OF ENVIRONMENTAL CONDITIONS ON ACCURACY

Environmental Factor	**Average Accuracy (%)**
Illumination- Invariant, Distance- Accurate, Head Movement- Minimum	92.625
Illumination- Changing, Distance- Normal, Head Movement- Minimum	92.325
Illumination- Invariant, Distance- Far, Head Movement- Minimum	92.575
Illumination- Changing, Distance- Normal, Head Movement- Yes	92.525

D. The effect of different subjects on the model

TABLE V. AFFECTS OF MULTIPLE SUBJECTS ON ACCURACY

Subjects (Training)	**Subjects (Validation)**	**Average Accuracy (%)**
1	5	91.21
3	5	92.32
5	5	92.67
10	5	92.63
10	10	92.73
12	10	92.67

The model is trained by capturing a subject's ocular regions and the x, y coordinates on the screen at the same time. Since different subjects have ocular sizes and shapes, the model was tested against multiple subjects which is illustrated in Table V. The data suggested that training the model with multiple subjects yielded better accuracies but there was little difference. The proposed architecture can estimate gaze coordinates even with a single subject as the images were augmented for better training and classification.

E. Results

The classification accuracy is illustrated using a confusion matrix in Fig. 4. The results were calculated after training 100 sample images per label for 50 epochs on a Dell Inspiron i5 5th gen 2.0 GHz laptop with 4GB ram using a low-resolution webcam. With larger dataset the algorithm yields even better accuracies for multiclassification of gaze detection on handheld devices. The working paradigm of the prototype is described in Fig. 5.

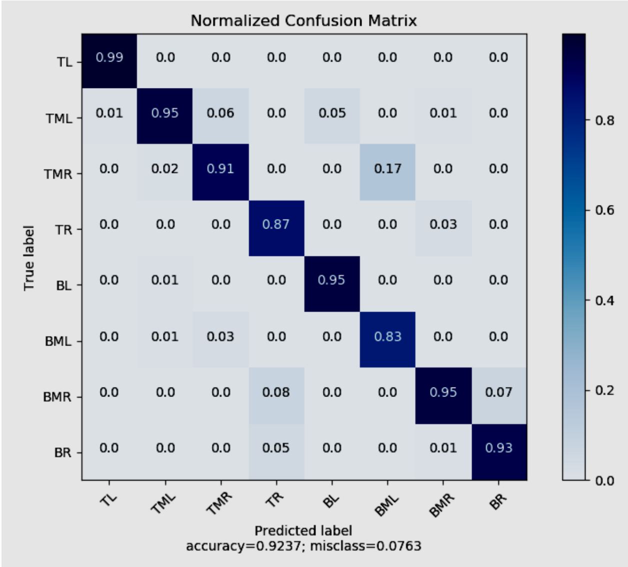

Figure 4. Confusion Matrix for classification of 8 classes.

The framework without the implementation of CoordConv [2] appeared to be overfitting after 40 epochs which was improved after concatenating coordinate channels with the CNN. Features of the architecture are listed below:

- The model is trained with low computational complexity.
- The model works in different illumination conditions.
- The model required little or no calibration for eye movement tracking.
- The model is invariant to distance of a subject from the image capturing device.
- The model requires less image from the subject as image augmentation is applied to increase data size.
- The model predicts eight class labels and plots x, y coordinates at the point of attention on screen even on devices with less computational power.

- The computational time per frame is 67 ms, which is much less compared to the results of the state-of-the-art framework by [15].

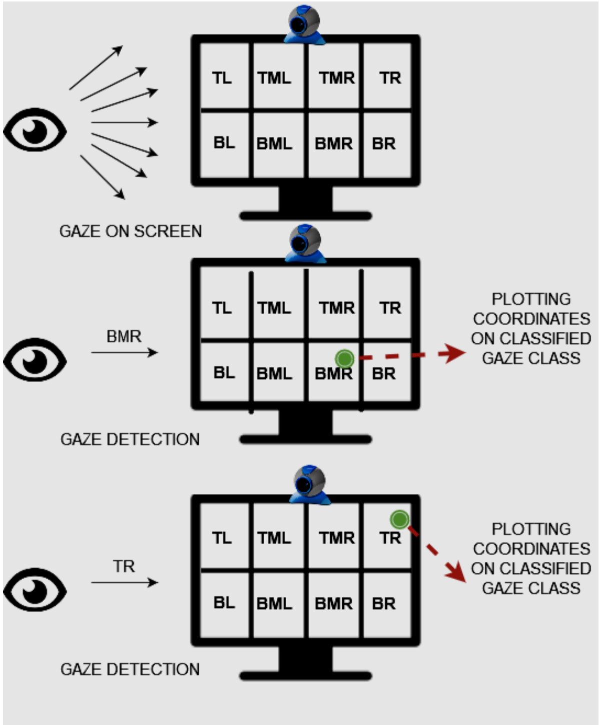
TL
TML
TMR
TR
BL
BML
BMR
BR
GAZE ON SCREEN
BMR
PLOTTING COORDINATES ON CLASSIFIED GAZE CLASS
GAZE DETECTION
TR
PLOTTING COORDINATES ON CLASSIFIED GAZE CLASS
GAZE DETECTION

Figure 5. Prototype paradigm for Gaze Detection.

V. CONCLUSION

This research provides a framework for real-time gaze detection on devices with low-computational power, varied illumination conditions and robustness again head movements using a modified Convolutional Neural Network that solves the coordinate transform problem required for plotting of coordinates for gaze estimation. The proposed architecture works on low resolution webcams with a smaller dataset without calibration even when trained with a single subject. This research can be used for UI control applications in handheld devices, human-computer interaction systems, manufacturing systems for validation by gaze, estimating subject attentions in educational sector and many more. The low computational complexity makes it most suitable for various eye-tracking and monitoring systems, smart devices that require little or no calibration or applications that need model training based on environmental conditions to predict better results as compared to pre-trained models. The future work of this research includes modifying the architecture to classify coordinates on a handheld device using the front camera of a mobile phone. We would also like to improve the accuracy of our model to predict more quadrants on the screen. Implementation of an algorithm to improve the quality of blurred or damages images in the training set can optimize the framework by improving the training set.

REFERENCES

[1] A. George and A. Routray, “Real-time eye gaze direction classificationusing convolutional neural network,” in2016 International Conference on Signal Processing and Communications (SPCOM), June 2016, pp.1–5.

[2] R. Liu, J. Lehman, P. Molino, F. P. Such, E. Frank,A. Sergeev, and J. Yosinski. An intriguing failing of convo-lutional neural networks and the coordconv solution.arXivpreprint arXiv:1807.03247, 2018.

[3] Similarweb.com, The State of Mobile Web – US 2015 ', 2014. [Online]. Available: https://www.similarweb.com/blog/report-the-state-of-mobile-web-us-2015. [Accessed: 03- Aug- 2019].

[4] V. Vaitukaitis and A. Bulling, “Eye gesture recognition on portabledevices,” inProceedings of the 2012 ACM Conference on UbiquitousComputing, ser. UbiComp ’12.New York, NY, USA: ACM, 2012,pp. 711–714. [Online]. Available: http://doi.acm.org/10.1145/2370216.2370370.

[5] D. W. Hansen and Q. Ji, “In the eye of the beholder: A survey ofmodels for eyes and gaze,”Pattern Analysis and Machine Intelligence,IEEE Transactions on, vol. 32, no. 3, pp. 478–500, 2010.

[6] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, “Appearance-based gazeestimation in the wild,” in2015 IEEE Conference on Computer Visionand Pattern Recognition. IEEE Computer Society, 2015.

[7] Zhang, Y. Sugano, M. Fritz, and A. Bulling, “It’s written all overyour face: Full-face appearance-based gaze estimation,” in2017 IEEEConference on Computer Vision and Pattern Recognition Workshops(CVPRW), July 2017, pp. 2299–2308.

[8] H. Deng and W. Zhu, “Monocular free-head 3d gaze tracking withdeep learning and geometry constraints,” in2017 IEEE InternationalConference on Computer Vision (ICCV), Oct 2017, pp. 3162–3171.

[9] L. Bottou, “Large-scale machine learning with stochastic gradient de-scent,” inProceedings of COMPSTAT’2010. Springer, 2010, pp. 177–186.

[10] J. Lemley, A. Kar, A. Drimbarean, P. Corcoran, “Efficient CNN Implementation for Eye-Gaze Estimation on Low-Power/Low-Quality Consumer Imaging Systems”, 2018.

[11] P. Viola and M. Jones, “Rapid object detection using a boosted cascadeof simple features,” inComputer Vision and Pattern Recognition, 2001.CVPR 2001. Proceedings of the 2001 IEEE Computer Society Confer-ence on, vol. 1. IEEE, 2001, pp. I–511.

[12] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks", arXiv:1801.04381v3 [cs], Apr. 2018.

[13] Sergey Ioffe , Christian Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, Proceedings of the 32nd International Conference on International Conference on Machine Learning, July 06-11, 2015.

[14] L. Bottou, “Large-scale machine learning with stochastic gradient descent,” in Proceedings of COMPSTAT’2010. Springer, 2010, pp. 177–186

[15] R. Vrˆanceanu, C. Florea, L. Florea, and C. Vertan, “Gaze direction estimation by component separation for recognition of eye accessing cues,” Machine Vision and Applications, vol. 26, no. 2-3, pp. 267–278, 20.