Appearance-Based Gaze Block Estimation via CNN Classification

Xuemei Wu
School of Information Science and Engineering
Shandong University
Jinan, China
wuxue1991sdu@163.com

Jing Li
School of Mechanical and Electrical Engineering
Shandong Management University
Jinan, China
lijingjdsun@hotmail.com

Qiang Wu
School of Information Science and Engineering
Shandong University
Jinan, China
wuqiang@sdu.edu.cn

Jiande Sun
School of Information Science and Engineering
Shandong University
Jinan, China
jiandesun@hotmail.com

***Abstract*—Appearance-based gaze estimation methods have received increasing attention in the field of human-computer interaction (HCI). These methods tried to estimate the accurate gaze point via Convolutional Neural Network (CNN) model, but the estimated accuracy can't reach the requirement of gaze-based HCI when the regression model is used in the output layer of CNN. Given the popularity of button-touch-based interaction, we propose an appearance-based gaze block estimation method, which aims to estimate the gaze block, not the gaze point. In the proposed method, we relax the estimation from point to block, so that the gaze block can be estimated by CNN-based classification instead of the previous regression model. We divide the screen into square blocks to imitate the button-touch interface, and build an eye-image dataset, which contains the eye images labelled by their corresponding gaze blocks on the screen. We train the CNN model according to this dataset to estimate the gaze block by classifying the eye images. The experiments on 6- and 54-block classifications demonstrate that the proposed method has high accuracy in gaze block estimation without any calibration, and it is promising in button-touch-based interaction.**

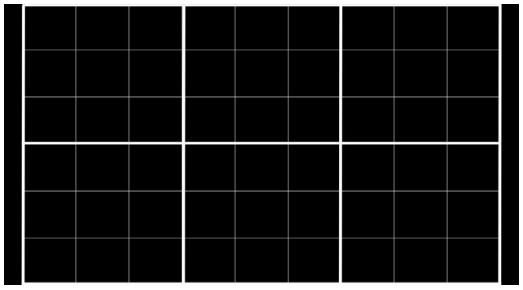
Keywords—Gaze estimation, Appearance-based, Gaze block, CNN, Button-Touch-Based Interaction

I. INTRODUCTION

Gaze serves as an important role in exploring outside world as eyes are one of the most vital organs of human. With the increasing development of non-contact human-computer interaction, gaze interaction technology is gradually becoming the highlight in this field. It will be great if gaze interaction can give one more "hand" to people, especially to the patients with severe paralysis but good vision.

At present, the gaze estimation methods can be divided into the model- and appearance-based methods. The model-based gaze estimation methods [1-3] usually require high resolution cameras to locate the pupil center accurately. In addition, one or multiple external infrared lights are used to get the corneal refection points, and the 3D geometric eye model is built to estimation gaze direction [4-5]. These methods usually have high estimation accuracy, but they need calibration and are easily affected by illumination.

Different from the model-based gaze estimation methods, the appearance-based methods estimate the gaze direction based on eye images or their features by mapping the relationship between gaze points and the related eye images. The intensity, histogram, and gradient of eye images are often used as the features, and the mapping is usually achieved by neural network [6], Gaussian Bayesian regression [7], and adaptive clustering [8]. Lu et al. [9-10] divided an eye image into several sub-regions manually and generated a 15-D intensity feature vector. They further adopted the Adaptive Linear Regression (ALR) to realize gaze point estimation. Zhang et al. [11] took the eye images as the input of CNN, and added the 3D head pose information into the last full connection layer. They changed the output of LeNet into a regression layer to predict the gaze point directly. But it can only achieve very low accuracy. Wang et al. [12] extracted the deep features of eye images using CNN, and used them as the input of random forest regression to predict the gaze point. Yusuke et al. [13] proposed a gaze sensing method using visual saliency maps and Gaussian process regression. Krafka et al. [14] established a large-scale dataset named GazeCapture consisting 1450 people for mobile equipment application. They built a robust eye tracker named iTracker by training both eyes and human faces. And in [15], they used a random forest based method to learn a regression function and estimate the gaze point. Almost all the appearance-based gaze estimation methods tried to adopt the regression models to estimate the accurate fixation points. However, it is really challenging to estimate the accurate gaze point based on eye images without calibration, so the estimation accuracies in these methods are very low. Different from the above methods, George et al. [16] proposed an eye gaze direction classification method to estimate the eye accessing cues (EAC) and infer the cognitive processes. In [16], the gaze direction was classified into 7 classes, i.e., Centre, UpRight, UpLeft, Right, Left, DownRight, and DownLeft.



978-1-5090-3649-3/17/$31.00 ©2017 IEEE

Fig. 1. The 6 and 54 gaze blocks used in our proposed method.

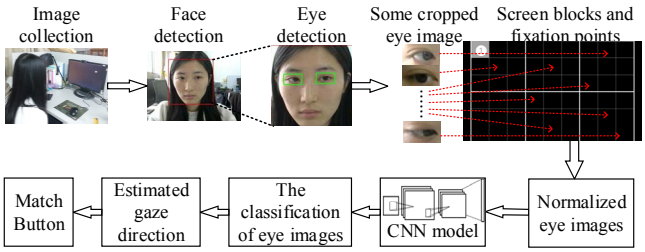


Fig. 2. Framework of the proposed gaze estimation method.

In this paper, the application of gaze interaction in button-touch-based interface is considered, where the user tends to touch the center of the button, but any touch on any position of the button can trigger the operation. Given this consideration, a gaze block estimation method is proposed. Different from most of the appearance-based gaze estimation methods, the proposed method aims to estimate the gaze block, i.e., touch-button in HCI, not a gaze point. The blocking of the screen is shown in Fig. 1, and there are 6- and 54-block levels respectively. The estimation on gaze block is implemented by the CNN-based classification, which reduces the difficulty of training the CNN with regression model to estimate the exact gaze points, and a new eye-image dataset is built for gaze block classification. The experiments verify the outperformance of the proposed method in gaze block estimation.

II. GAZE BLOCK ESTIMATION

Fig. 2 shows the framework of the proposed method in this paper. We collect the images that contain the subject's face, and detect out the eye images. Given the eye images, we train the CNN-based classification model to estimate the gaze block.

A. Eye-Image Dataset

Most exiting gaze estimation datasets are designed for regression task, i.e., for gaze point estimation, and are not suitable for screen-based interaction, so they are not suitable for our proposed method. Hence, we build our own eye-image dataset. Table I shows a comparison between several eye-image datasets used in the existing appearance-based gaze estimation methods and ours. The attributes of these dataset include the number of subjects (NS), the type of fixation points (TF), number of targets (NT), illumination conditions (IC) and the number of eye images (NI). The *cont.* in the "NT" column means continuous fixation targets. In our dataset, the fixation positions are quantized into the center of blocks, and we can call them quantized points (QP). Compared to our fixation point type, the fixation positions in most of the other datasets are not quantized, and they are considered as the discrete temporal points (DTP). The "3" in the "IC" column means there are 3 illumination conditions, i.e., natural, darker, and stronger illumination. Most of the above-mentioned datasets are captured with the sitting or standing pose, but the TabletGaze dataset is captured with more poses, such as lying and slouching.

In our eye image collecting, the web camera, Logic C270, and one normal 19-inch screen are used. The eye-image resolution is 640×480 and the aspect ratio of the screen is 16:9. We divide the computer screen into 6 big blocks with the size of 12.75×12.75 cm as shown in Fig.1 in thick lines. Each big block is further divided into 9 small blocks with the size of 4.25×4.25 cm so that we get 54 small blocks in Fig. 1 in thin lines. We set a white fixation point in the middle of each small block with the black screen background, ask the subject to gaze at it, and capture the subjects' eye images to build the dataset. The fixation point appears randomly and only one point is displayed on the screen at each time.

During the data collection, each subject is asked to sit about 60cm in front of the screen to ensure that the full face can be captured, and one centimeter on the screen is approximately equal to one degree in gaze direction in this distance. Totally 56 groups of eye videos are collected from 22 subjects aged in 20~30 range, and each video lasted about 6 minutes with 30fps. In order to prevent the fixation fatigue, there is a two-second rest between every two fixation points.

After the images acquisition, we detect the subject's full face in the original image by using Haar-like features [19]. Firstly, the centers of the two eyes are roughly detected, and then the human face is corrected by the angle between the line of the two eyes and the horizontal direction. Assuming this angle is marked as θ, then the face image can be corrected by rotating θ degree. Finally, the eyes can be located accurately in the face region after the face correction.

After removing the closed-eye images, 181440 eye images are obtained. The average number of eye images in each small block is about 3360, and the number of eye images from each subject varies from 3240 to 12960. The eye images of each subject are acquired at least twice, and the interval between every two times is at most 20 days and at least half a day. Several eye images are shown in Fig. 2. The resolution of the eye images is fixed to 40×72 pixels. We randomly selected the 151200 training images and 30240 testing images to train the CNN model. The left and right eyes are used separately.

B. Gaze Estimation via CNN Classification

In practical button-touch-based application, any touch on the range of the touch-button can trigger the touch operations, but users usually tend to touch the center of the touch-button. Given this consideration, we take the eye images corresponding to the same gaze block as the touch on the same touch-button. Therefore, we can estimate the gaze block by classifying the eye images into different classes via CNN. That is to say, CNN is used to build the relationship between the eye images and the gaze blocks on the screen.

TABLE I. COMPARION BETWEEN VARIOUS DATASETS USED IN APPEARANCE-BASED GAZE ESTIMATION

Database	NS	TF	NT	IC	NI
UT Multiview [15]	50	DTP	160	1	64,000
Eye Chimera [16, 17]	40	DTP	7	1	1,172
Wang et al. [12]	6	DTP	41	3	107,681
MPII Gaze [11]	15	DTP	*cont.*	Daily life	213,659
GazeCapture [14]	14 50	DTP	13+ *cont.*	Daily life	videos
TabletGaze [18]	51	QP	35	Daily life	videos
Ours	22	QP	54	Daily life	181,440

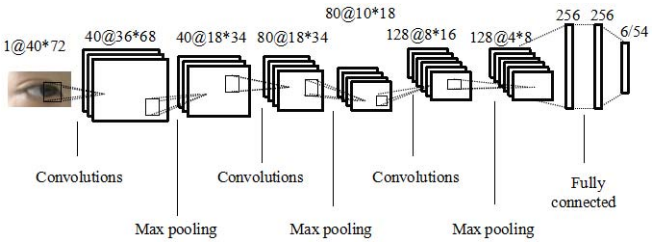


Fig. 3. Architecture of the CNN model used in the proposed method.

The CNN architecture used in the proposed method is shown in Fig. 3. We simplify the CNN model designed for ImageNet in [20], because the diversity of eye images in our dataset is not as much as that of the ImageNet images. The model we adopted consists three convolutional layers followed by the max-pooling layers, two fully connected layers and a soft-max layer for classification. Given the division of the screen, the network exports 6 classes at first and then 54 classes. The network inputs are the RGB eye images as the conversion from RGB images to gray images may lose some potentially useful information. For the three convolutional layers, the feature filter sizes is 5×5, 5×5, and 3×3 pixels, while the numbers of feature maps of the three convolutional layers is 40, 80, and 128 respectively. The number of hidden units for the two fully connected layers is 256. At the same time, we also set the corresponding excitation layers and Local Response Normalization (LRN) layers between the layers.

The feature map of convolutional layer is defined as:

$$x_j^l = relu(\sum_{i\in M} x_i^{l-1} * k_{ij}^l + b_j^l) \tag{1}$$

where x_j^l is jth feature map at lth layer, M is the number of feature maps at $l-1$ layer, "$*$"is the convolution operator, k_{ij}^l is the convolution kernel, and b_j^l is the bias of the jth feature map at the lth layer.

The feature map of sub-sampling layer is defined as:

$$x_j^l = f(\beta_j^l \max(x_j^{l-1}) + b_j^l) \tag{2}$$

We adopted the max-pooling operation, and the output images become one fourth through this operation.

In the last two fully connection layers, a 256-dimension deep feature is extracted and the deep feature is used for eye images classification by minimize the cross-entropy loss. The final class can be determined by the maximum probability $prob$, which is calculated by the softmax function. That means the sample belongs to the category which corresponds to the maximum probility.

$$class = \arg\max_{label}(prob) \tag{3}$$

III. EXPERIMENTS AND EVALUATIONS

A. Gaze Block Estimation Performance based on both our data set and MPIIGaze dataset

In the experiments, we first randomly selected the training and the testing samples out of the our dataset in a ratio of 5:1, i.e., the number of training samples is 151200, and there is no overlap between the training and testing samples. The average

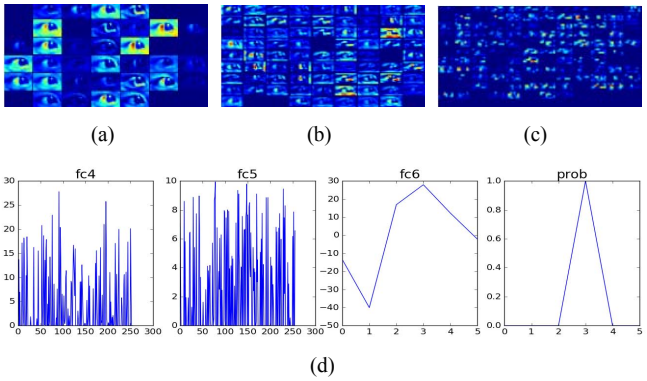


Fig. 4. Output of different layers in CNN. (a) output of the 1st pooling layer ($40\times18\times34$), (b) output of the 2nd pooling layer ($80\times10\times18$), (c) output of the 3rd pooling layer ($128\times4\times8$). (d) output of the fully connected layers.

training time with 20000 iterations based on Caffe is about 17.69 minutes. For left eyes, the proposed method can reach the average accuracy with 92.7% for the 6-class classification and 82.1% for the 54-class classification. And for the right eyes, the proposed method can reach 92.5% and 78.3% respectively.

The output deep features of an input eye image are shown in Fig. 4. Fig. 4(a) and 4(b) show the outputs of the first two pooling layers. We can see the contours and corners of pupils and eyes through the output of the first two pooling layers. But the output of the final pooling layer only shows the active units in Fig. 4(c). Fig. 4(d) presents the output of fully connected layers. The final probability distribution curve shows that the eye image belongs to the class with the highest probability. The confusion matrix for 6-class and 54-class classifications are shown in Fig. 5. Diagonal lines indicate the probability of correct classification. It can be concluded that the proposed method has very high estimation accuracy on every block.

We also test our gaze block gaze classification method on the MPIIGaze dataset in [11]. The dataset offered eye images and the corresponding gaze direction in three-dimensional space. In order to adapt our experiment, we transformed the raw gaze direction into two-dimensional angular coordinates and then quantized the gaze coordinates into our gaze blocks on the screen. We random select 66000 eye images from the MPIIGaze dataset and the results of classification accuracy for both our dataset as well as the MPIIGaze dataset are shown in TABLE II. The classification results show that although the mapping method from the exact gaze direction to the gaze block can be used for classification task, the performance is much worse. And it is necessary to build the specific dataset for the button-based gaze interaction.

B. *Cross-Subject Performance*

We investigated the cross-subject estimation performance of the proposed method. We adopted both our dataset and the MPIIGaze dataset in this section. We randomly selected 4 subjects from the two dataset respectively. The eye images from each subject are used as the testing set by turns and the eye images from the other subjects are used as the training set. The accuracies of the 6-class and 54-class classification are shown in TABLE III. We use n1-n4 to represent the selected 4 subjects from our dataset and s1-s4 to represent the selected 4 subjects from the MPIIGaze dataset. It demonstrates that although there are significant differences between subjects, our dataset can achieve higher accuracy for 6-class classification. And the performance are much worse in the case of 54-class classification. It can be concluded that the gaze estimation can't achieve high accuracy in the case of cross-subject due to the diversity of human eyes and the general characteristics extracted through CNN can only be used to represent the common parts between individuals. The performance can be upgrade when our dataset contain more individuals.

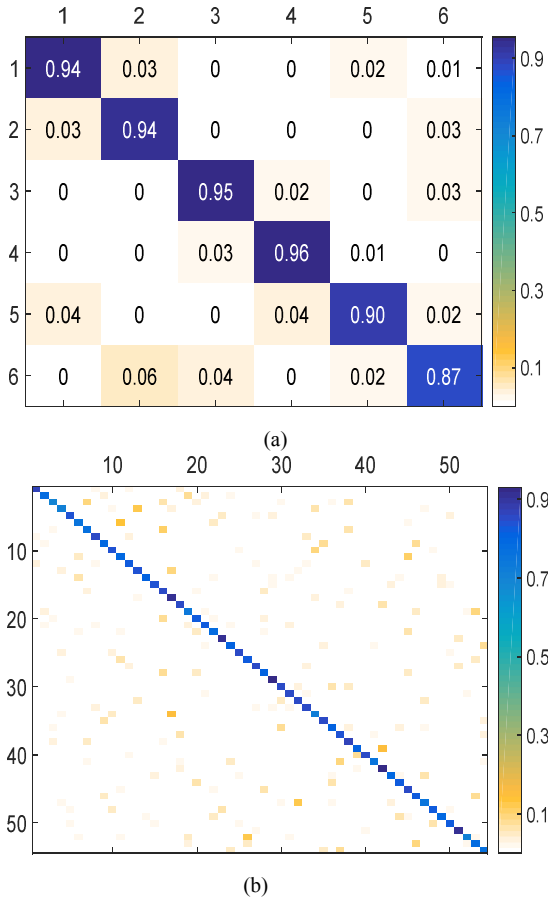


Fig. 5. (a) The confusion matrix for 6-class classification. (b) The confusion matrix for 54-class classification.

C. *Comparison with other methods*

George et al. [16] proposed a screen-based eye gaze direction classification method using CNN. They trained the network independently for left and right eyes and combine the two scores to get the final class labels. Zhang et al. [11] proposed appearance-based gaze estimation method in the wild. They estimated the gaze direction via using a CNN model and trained a regression model in the output layer. In this experiment, we carry out their methods on our database and make a comparison with our proposed method. The comparison results in Table IV show that our proposed method has 4% accuracy improvement for the 6-class classification and 27.3% accuracy improvement for the 54-class classification considering only left eyes. And the accuracy improvement can be 3.8% and 23.5% respectively considering only right eyes. We also adopted the regression method in [11] based on our dataset and mapped the predicted gaze coordinates to the screen blocks. The results are much worse than our proposed method as the best performance of mean error is reported as 10.5 degree in [11], which is much larger than our largest error, i.e., 6.375 degree.

TABLE II. COMARISON BETWEEN OUR DATASET AND THE MPIIGAZE DATASET

Datasets		**Estimation Accuracy (%)**	
		6-class	***54-class***
Our Dataset	Left eyes	92.7	82.1
	Right eyes	92.5	78.3
MPIIGaze in [11]		75.6	39.2

TABLE III. INVESTIGATION ON CROSS-SUBJECT GAZE BLOCK ESTIMATION

Dataset	**Test Subject**	**Estimation Accuracy (%)**	
		6-class	***54-class***
Ours	n1	90.1	42.4
	n2	86.9	20.2
	n3	78.0	12.2
	n4	77.9	17.7
MPIIGaze in [11]	s1	69.3	15.2
	s2	52.2	14.8
	s3	51.5	16.0
	s4	48.5	13.0

TABLE IV. COMARISON BETWEEN THE PROPOSED METHOD AND THE METHODS IN [11] AND [16]

Methods		**Estimation Accuracy (%)**	
		6-class	***54-class***
Method in [11]		23.8	
Method in [16]		88.7	54.8
Our method	Left eyes	92.7	82.1
	Right eyes	92.5	78.3

D. *Analysis*

We can observe from the above experiments that: 1) The CNN-based classification is helpful to estimate the gaze block accurately. Given the size of block, 6-class classification amounts to 6.375 degree in the gaze point estimation accuracy when the gaze point is at the center of the gaze block. And 54-class classification means 2.125 degree. 2) To improve the accuracy of cross-subject gaze estimation could be a challenge due to various physiological properties of human eyes. But the accuracy can be improved greatly, when the training set cover the data of the subject referring to Fig. 5. Therefore,

incremental learning might be considered when the method is used in practice.

IV. CONCLUSION

In this paper, we present a block-based gaze estimation method, which is designed for the gaze-based button-touch interaction. We imitate the button-touch interface via dividing the screen into square blocks, estimate the gaze block based on eye images by using CNN, and verify the performance of the proposed method on blocks with two different sizes. The experiment shows that our proposed method has high classification accuracy for the two chosen accuracy ranges. For the future work, we will continue to enrich our dataset with more subjects and more situations. Better CNN model considering more information and incremental learning will be studied to provide more support to the practical implementation in consumer electronics.

ACKNOWLEDGMENT

This work is supported by Key Research and Development Foundation of Shandong Province (2014GGX101009, 2016GGX101009), Natural Science Foundation of Shandong Province (ZR2014FM012), and Scientific Research and Development Foundation of Shandong Provincial Education Department (J15LN60). We acknowledge the support of NVIDIA Corporation with the donation of the TITAN X GPU used for this research. The contact author is Jiande Sun.

REFERENCES

[1] E. D. Guestrin and M. Eizenman, "General theory of remote gaze estimation using the pupil center and corneal reflections," in *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 6, pp. 1124-1133, June 2006.

[2] C. H. Morimoto and M. R. Mimica, "Eye Gaze Tracking Techniques for Interactive Applications," *Computer Vision and Image Understanding*, vol. 98, no. 1, pp. 4-24, April 2005.

[3] Y. m. Cheung and Q. Peng, "Eye Gaze Tracking With a Web Camera in a Desktop Environment," in *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 4, pp. 419-430, Aug. 2015.

[4] C. Yang, J. Sun, J. Liu, X. Yang, D. Wang, and W. Liu, "A Gray Difference-Based Pre-Processing for Gaze Tracking," *Proceedings of IEEE International Conference on Signal Processing*, pp. 1293-1296, 2010.

[5] C. Niu, J. Sun, J. Li, and H. Yan, "A Calibration Simplified Method for Gaze Interaction Based on Using Experience," *Proceedings of 2015 IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1-5, 2015.

[6] S. Baluja and D. Pomerleau, "Non-Intrusive Gaze Tracking Using Artificial Neural Networks," *Technical Report CMU-CS-94-102, Carnegie Mellon University*, January, 1994.

[7] N. Ye, X. Tao, L. Dong and N. Ge, "Mouse Calibration Aided Real-Time Gaze Estimation Based on Boost Gaussian Bayesian Learning," *Proceedings of 2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, 2016, pp. 2797-2801.

[8] Y. Sugano, Y. Matsushita, Y. Sato and H. Koike, "Appearance-Based Gaze Estimation with Online Calibration from Mouse Operations," in *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 6, pp. 750-760, Dec. 2015.

[9] F. Lu, Y. Sugano, T. Okabe and Y. Sato, "Adaptive Linear Regression for Appearance-Based Gaze Estimation," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 10, pp. 2033-2046, Oct. 2014.

[10] F. Lu, Y. Sugano, T. Okabe and Y. Sato, "Inferring Human Gaze from Appearance via Adaptive Linear Regression," *Proceedings of 2011 IEEE International Conference on Computer Vision*, pp. 153-160, 2011.

[11] X. Zhang, Y. Sugano, M. Fritz and A. Bulling, "Appearance-Based Gaze Estimation in the Wild," *Proceedings of 2015 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015, pp. 4511-4520.

[12] Y. Wang, T. Shen, G. Yuan, J. Bian, and X. Fu, "Appearance-Based Gaze Estimation Using Deep Features and Random Forest Regression," *Knowledge-Based Systems*, vol. 110, pp. 293-301, 2016.

[13] Y. Sugano, Y. Matsushita and Y. Sato, "Appearance-Based Gaze Estimation Using Visual Saliency", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 2, pp. 329-341, Feb. 2013.

[14] K. Krafka, A. Khosla, P. Kellnhofe, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye Tracking for Everyone," *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 2176-2184.

[15] Y. Sugano, Y. Matsushita and Y. Sato, "Learning-by-Synthesis for Appearance-Based 3D Gaze Estimation," *Proceedings of 2014 IEEE International Conference on Computer Vision and Pattern Recognition*, Columbus, OH, 2014, pp. 1821-1828.

[16] A. George and A. Routray, "Real-Time Eye Gaze Direction Classification Using Convolutional Neural Network," *Proceedings of 2016 IEEE International Conference on Signal Processing and Communications (SPCOM)*, Bangalore, 2016, pp. 1-5.

[17] L. Florea, C. Florea, R. Vranceanu, and C. Vertan, "Can Your Eyes Tell Me How You Think? A Gaze Directed Estimation of the Mental Activity," *Proceedings of the British Machine Vision Conference*, pp. 60 1-11, 2013.

[18] Q. Huang, A. Veeraraghavan, and A. Sabharwal. "TabletGaze: A dataset and baseline algorithms for unconstrained appearance-based gaze estimation in mobile tablets," arXiv:1508.01244, 2015, pp. 2-8.

[19] R. Lienhart and J. Maydt, "An Extended Set of Haar-Like Features for Rapid Object Detection," *Proceedings of 2002 IEEE International Conference on Image Processing*, vol. 1, pp. 900 -903, 2002.

[20] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", in *Advances in Neural Information Processing Systems 25*, pp. 1106-1114, 2012.