IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 22, 2020, accepted April 6, 2020, date of publication April 9, 2020, date of current version April 23, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2986815

Deep Fusion for 3D Gaze Estimation From Natural Face Images Using Multi-Stream CNNs

ABID ALI AND YONG-GUK KIM
Department of Computer Engineering, Sejong University, Seoul 05006, South Korea

Corresponding author: Yong-Guk Kim (ykim@sejong.ac.kr)

This work was supported by the Institute for Information and Communications Technology Promotion (IITP) funded by the Korea Government (MSIT), through personalized advertisement platform based on viewers' attention and emotion using deep-learning, under Grant2017-0-00731.

ABSTRACT Over the last few decades, eye gaze estimation techniques have been thoroughly investigated by many researchers. However, predicting a 3D gaze from a 2D natural image remains challenging because it has to deal with several issues such as diverse head positions, face shape transformation, illumination variations, and subject individuality. Many previous studies employ convolutional neural networks (CNNs) for this task, and yet the accuracy needs improvement for its practical use. In this paper, we propose a 3D gaze estimation framework based on the data science perspective: First, a novel neural network architecture is designed to exploit every possible visual attribute such as the states of both eyes and the head position, including several augmentations; secondly, the data fusion method is utilized by incorporating multiple gaze datasets. Extensive experiments were carried out using two standard eye gaze datasets, including comparative analysis. The experimental results suggest that our method outperforms state-of-the-art with 2.8 degrees for MPIIGaze and 3.05 degrees for EYEDIAP dataset, respectively, indicating that it has a potential for real applications.

INDEX TERMS Gaze estimation, data fusion, convolutional neural networks, MPIIGaze, EYEDIAP.

I. INTRODUCTION

Eye movement and gaze estimation are important in terms of visual and cognitive processing [1]. Specifically, eye movements have been widely studied for human visual attention [2], [3], emotion analysis [3] and for behavioral disorder identification [2], [4]. Gaze estimation has been studied thoroughly in the computer vision area because it has a wide range of applications in human-computer interaction [5], psychology [1], [6], [7], disability studies [8], navigation and detecting driver's behavior [9], surgical robots [10] and marketing research [11]–[15].

Given that the previous models and features-based methods for gaze prediction have certain limitations depending on the illumination condition, camera calibration method, and individual head-pose variations. Computer vision researchers have been explored the appearance-based methods to estimate the human gaze in an uncontrolled environment typically using a convolutional neural network (CNNs), due to recent availability of large gaze datasets. Even though deep learning approaches have achieved a remarkable success in estimating human gaze within a natural environment, the current approaches achieve about 3.6 degrees, which are still far away from applying it to real-time applications.

For the appearance-based gaze estimation method, the state of the art techniques typically utilize a full-face image as input [16]. Krafka *et al.* initially proposed a weight sharing mechanism where they used an Alex-net like architecture to estimate a 2D gaze from still images [17]. Given that a face has two eyes, it seems reasonable to use dual eye channels to estimate a gaze [18]. Since eye gaze behavior is not static, the head movement is responsible for a gaze to locate a target of interest. Lian *et al.* proposed a feature fusion technique for gaze direction and point estimation utilizing eye patches from multiple cameras [19]. Their goal was to use features from MPIIGaze for gaze direction prediction combined with ShanghaiTechGaze for gaze point estimation by weight sharing technique. In the present study, we show that the two eye patches, along with the head position, are essential to estimate a 3D gaze accurately, even in uncalibrated and uncertain environments. Our extensive experimental results suggest that the proposed network is a light and yet high-performing

The associate editor coordinating the review of this manuscript and approving it for publication was Gianluigi Ciocca.

gaze estimation method. Contributions of this research are summarized as below.

- Since the task of gaze estimation is highly dependent on eye movements and head translation and rotation, a small variation in the movements of eyes and head, makes large differences in gaze angle. Therefore, employing gaze estimation in a real-time application requires the most accurate system. In this paper, we presented a multi-stream shallow CNN with a dual spatial layer mechanism that was based method that combines features from both eye patches and a head position for 3D gaze estimation. Thus, our network architecture is light, fast, and highly accurate, and it outperformed the state-of-the-art methods (Section III-B).
- Data fusion is an effective technique to improve the system accuracy. For deep learning-based computer vision tasks, it usually needs a considerable amount of data in order to achieve the best results. However, it is quite a time-consuming and challenging task to collect such an enormous amount of data, especially for eye tracking and gaze estimation tasks since there is a limited number of datasets available. To solve this problem, a data fusion technique is designed by training our network using two publicly available datasets, MPIIGaze [20] and EYEDIAP [21] and testing one of them in turn (see Section IV-F). To the best of our knowledge, this is the first report that employs such a data fusion technique for 3D gaze estimation.
- To analyze the effects of a dual spatial layer mechanism efficiency, a comparative analysis between a single spatial layer and a dual spatial layer mechanism is preformed. It is found that the accuracy is much improved with a double spatial layer compared with a spatial layer [16] as described in Section IV-G).
- The resource-constrained devices, such as Raspberry-pi and mobile devices, have low computation power and it is difficult for a deep neural network to perform well. The proposed architecture is very light and fast, which makes it adaptable for resource-constrained devices (Section V).

The rest of the paper is organized as follows. Section II introduces the related work that is conducted about eye movement tracking. The proposed gaze estimation method using CNN is described in Section III. The experiments carried out are explained in Section IV. Further discussion is made in Section V and finally, we conclude our proposed method in Section VI.

II. RELATED WORK

In this section, we briefly review the previous literature on computer vision-based gaze estimation methods, which are typically categorized as feature-based, model-based, and appearance-based. Also, the relevant characteristics of the CNN-based architecture for regression tasks are be discussed in detail.

A. FEATURES-BASED GAZE ESTIMATION

Feature-based gaze estimation methods involve the usage of hand-crafted extracted local features, such as the pupil center, eye contours, and glints. Alternatively, other auxiliary information, such as the head position, is used to estimate the gaze direction. In earlier periods, an IR light source and a collection of mirrors and galvanometers were used to extract pupil-glints and head movement features for real-time eye tracking [22]. Huang *et al.* used six landmarks around a single eye as the feature with the head position in estimating a gaze [23]. Similarly, [24] proposed a Pupil-Center-Eye-Corner (PC-EC) method, that is later used to estimate the gaze direction on public displays [25]–[27] by combining the eye region landmarks model and the PC-EC features. Similarly, other eye-tracking methods utilized the local binary pattern (LBP) features [28], Gaussian Laplace [28], and the histogram of Gaussian features [21], [23]. However, these methods require different hand-crafted feature extraction techniques within a controlled calibrated environment, rather than the natural environment.

B. MODEL-BASED GAZE ESTIMATION

The model-based approach adopts geometric eye models for gaze estimation, and they are divided into shape-based and corneal-reflection-based methods, which depend on the requirement of external Infra-red light sources. Earlier work on eye-tracking involved the corneal-reflection methods, which are limited to only the stable head movement settings [29]–[32] and are improved to handle some head poses by imposing different light sources and cameras [20], [33], [34]. On the other hand, the shape-based methods [9], [35], [36] used the pupil center and the iris edges to estimate the gaze direction.

Although the model-based method achieves a high accuracy, which is around 1 degree, they require different cameras, light sources, and calibration systems. They cannot perform well in low light conditions and with low-quality images.

C. APPEARANCE-BASED GAZE ESTIMATION

Appearance-based methods aim to map directly gaze directions by taking raw images as the input. They typically utilize a single camera to capture the eye images [20], [31] and can predict the gaze with low-resolution images. There are several appearance-based gaze estimation methods such as adaptive linear regression ALR [37], artificial neural networks [17], [18], [38], linear interpolation [31], visual saliency mapping [39], and Gaussian process regression [30]. Previously, appearance-based method operated on a stationary head pose and required a specific training data for each person [30], [31], [37]. However, new methods have been focused on pose-independent gaze estimation either from RGB still images or using depth information from RGB-D images [16], [20], [21], [40], [41], but they still require user-specific training.

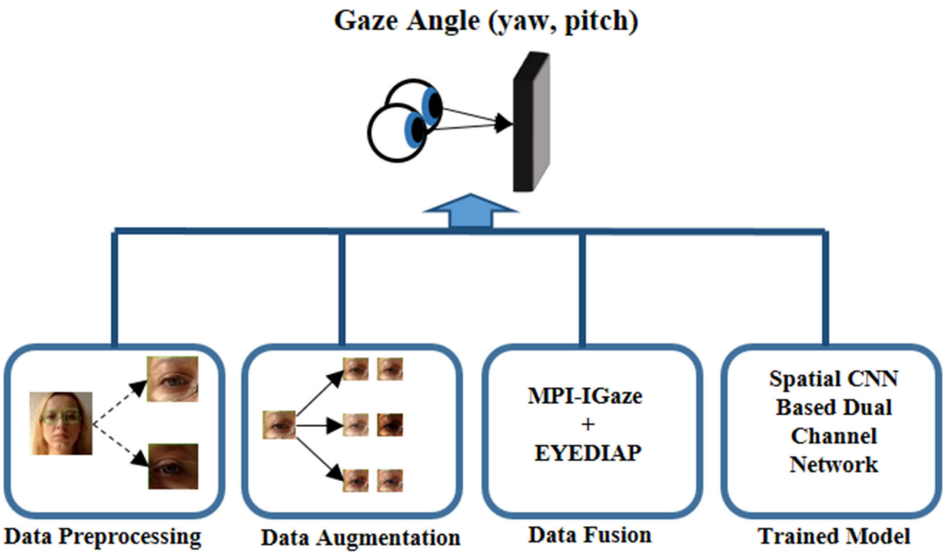

FIGURE 1. A schematic diagram of our proposed method where data preprocessing, data augmentation and dataset fusion are used for training the model.

D. CNN-BASED GAZE ESTIMATION

The CNN-networks have proved to be extremely effective not only for classification tasks [42] but also for regressions [43], which include gaze estimation [16], [17], [20], [44]. Several new methods have effectively employed deep learning and CNNs for 2D and 3D gaze estimation. Reference [17] proposed a 2D gaze estimation network for mobile devices, and later [45] included temporal information to improve the accuracy of the Itracker method by introducing a bi-directional LSTM network to the existing architecture. Zhang *et al.* used a full face image as input to a modified pre-trained Alex-Net to predict the gaze [16]. Park *et al.* introduced a stacked hourglass method for the eye region landmarks and gaze estimation [27]. Other CNN-based methods included multi-stream CNN architectures, such as an evaluation-guided asymmetric regression network [46], a recurrent CNN network that uses eye patches and facial landmarks as input [40], a deep 3D gaze estimation that uses a model ensemble technique [41], and a sequential neural network-based deep pictorial representation of a 3D gaze that uses a single eye the input [47]. Reference [48] proposed a differential network for gaze estimation by using reference samples from a specific person for the person-specific gaze estimation. A recent study, employed a professional eye tracker to train a camera far away from user for long distance gaze estimation involving CNNs for training [49]. Nonetheless, the accuracy of above methods is not satisfactory for real world application. The present study proposes an efficient multi-stream CNN based method that requires less computing resources and yet achieves a high accuracy.

III. MULTI-STREAM CNN NETWORK

In this section, a new approach is described on how to predict a 3D gaze angle using the combined features from both eye patches and the head position as shown in Fig. 1. The important step for 3D gaze estimation would be the data normalization before performing any regression tasks since the importance of data augmentation in term of system performance will be emphasized.

A. IMAGE NORMALIZATION

To overcome any appearance variation and to predict the gaze correctly regardless of the original camera parameters, Sugano *et al.* proposed a data normalization method for 3D appearance-based gaze estimation [39], which was further revised by [50]. This work used the revised version for data normalization. Given an input image $\boldsymbol{I}$, and a reference location $\boldsymbol{x}$, the goal is to calculate the conversion matrix $\boldsymbol{M}$ using (1). Using the rotation matrix $\boldsymbol{R}$, the x-axis of both head coordinates system and the camera are parallel.

$$M = SR \tag{1}$$

The scaling matrix S is defined so that the virtual camera looks at the reference point from a fixed distance d_s using (2).

$$S = diag(1, 1, d_x/||P||) \tag{2}$$

The images are normalized with the perspective warping using a transformation matrix in (3), where C_s is the projection matrix of the normalized camera, and C_n is the real camera matrix. The normalized images are cropped patches of size $W \times H$ centered at $\mathbf{p}$ with the head roll being removed.

$$W = C_s M C_{\mathrm{n}}^{-1} \tag{3}$$

The 3D ground truth gaze vector is also normalized using (4). After normalization, the gaze vector is further converted to spherical coordinates (horizontal and vertical

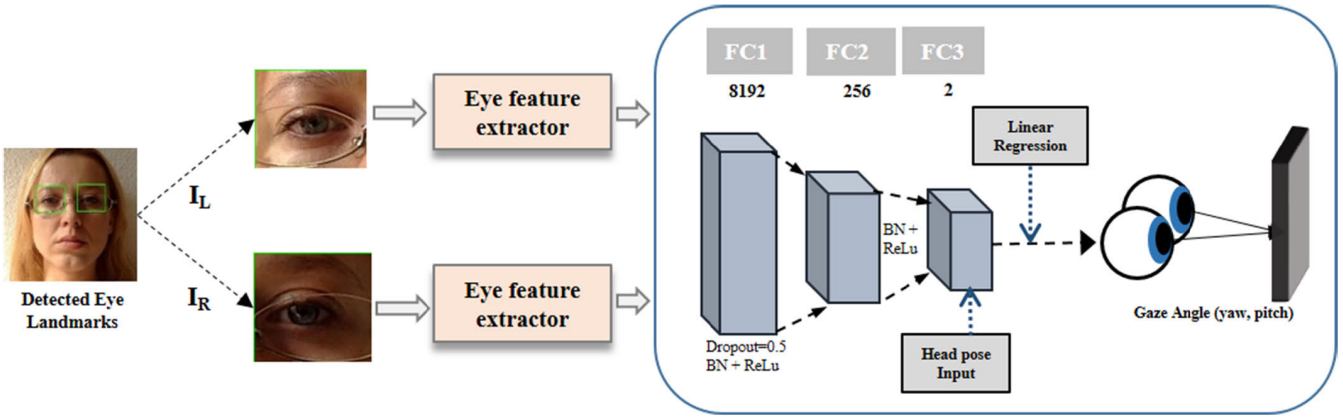

FIGURE 2. An overview of the proposed method. The eye landmarks were extracted from an image and used as the input for the eye feature extractor. A feature vector $V = U \odot W$ was extracted, which is passed to the FC layers for the gaze estimation. At FC3, the head pose is appended as the input to the network and the gaze angle (yaw and pitch) is computed with linear regression.

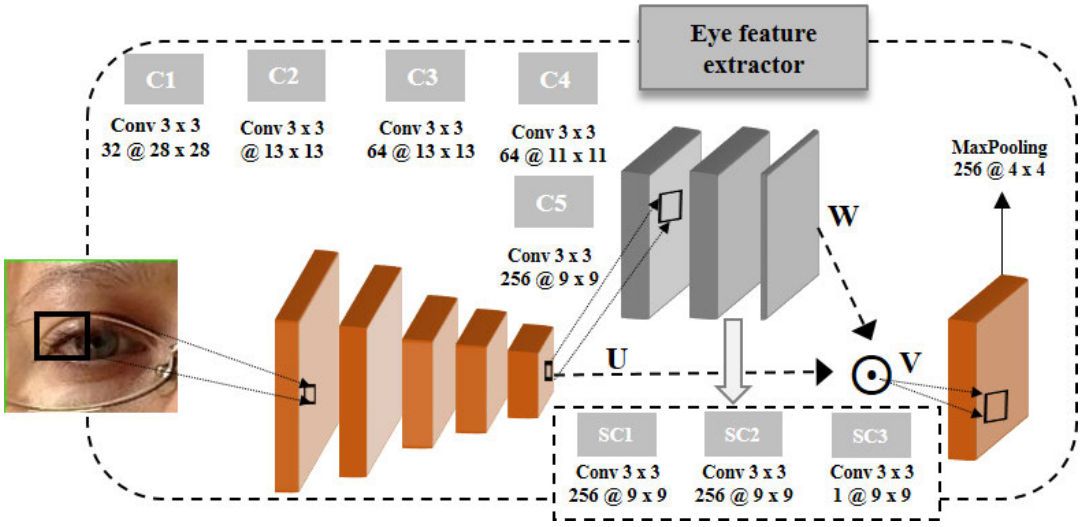

FIGURE 3. Baseline CNN architecture for eye feature extraction. C represents convolution layers, while SC represent spatial convolution layers.

gaze directions), assuming the unit length. All the data from both datasets are normalized similarly during both training and testing.

$$g_n = Rg \tag{4}$$

B. PROPOSED GAZE ESTIMATION NETWORK

The proposed shallow multi-stream CNN-based network have a spatial layer mechanism for 3D gaze estimation. The network consists of baseline CNN architecture for feature extraction, which is illustrated in Fig. 3, and it is used to extract the features from each eye separately, as shown in Fig. 2.

The network takes two eye images $\{I_L^{(i)}, I_R^{(i)}\}$ with a size of 60×60 and the head-pose angle $h^{(i)}$ as the input to learn the regression function f that predicts the 2D gaze angle $g^{(i)}$, where i is the index of each sample. Two previous studies are related to the present one. Lamely *et al.* proposed a small CNN based framework for 3D gaze estimation [18], and Zhang *et al.* implemented a spatial weight mechanism with a baseline that it could enhance some regions of the face for gaze estimation [16]. In this paper, two-stream CNNs are utilized for the two eye patches to process with a dual weight mechanism, which is slightly different from the above two methods. Our network has 5 convolutional layers for the feature extraction and a spatial mechanism, which consists of three convolutional layers with a filter size of 1×1 along with a rectified unit layer, and a final max-pooling layer is applied at the end of the baseline network, which is illustrated in Fig. 3.

$$V = W \odot U \tag{5}$$

The weighted activation maps were calculated using (5) where W and U represent the spatial weight matrix and the original activation tensor, respectively. It was found that using

the same spatial layers mechanism for each eye image can significantly improve the overall performance of the network as illustrated in Fig. 2.

C. 3D GAZE ESTIMATION NETWORK FLOW

An eye patch having input size of (60 × 60) was fed separately to a CNN, consisting of five convolutional layers, each followed by a Batch-Normalization (BN) layer and a rectified linear unit layer as illustrated in Fig. 3. The output from the 5^{th} convolutional layer (**C5**) was fed to the spatial layers. At this point, the last max-pooling layer reduced the dimensions of features received from the element-wise multiplication of the spatial layers and the original activation matrix with equation (5). Before concatenation of both eye features, a dropout layer ($p = 0.5$) was connected to a fully connected (FC) layer with a size of 512, which was followed by a BN and a rectified linear unit layer, and finally two more FC layers with sizes of 256 and 2, respectively. The head pose vector was appended to the final layer and then the output of the final layer were the gaze angles yaw and pitch.

Extensive experiments were carried out to find out the best network architecture. It was found that adding a BN layer before an activation layer was beneficial. It helped to improve the accuracy as well as to increase the generalization ability of a regression-based architecture [51]. Experiments with and without using a BN layer indicated that a BN layer for the spatial weight mechanism decreased the performance, so a BN layer for spatial weights was not used. However, it was found that there were no improvements with the BN layers during the training of the architecture, but the generalization ability was highly improved during validation and testing as shown in Table 1.

TABLE 1. Model generalization was very much improved by introducing a BN layer before the activation layer for both the CNN and the fully connected layers except the last layer. The model performed well on the test data when the BN layers were added. The best results are shown in bold.

Dataset	BN layer	Train acc. (degrees)	Test acc.(degrees)
MPIIGaze	without BN	1.7	7.8
	With BN	**1.85**	**2.65**
EYEDIAP	Without BN	2.8	12.9
	With BN	**2.68**	**3.60**

D. IMAGE DATA AUGMENTATION

To improve the robustness of our method, the training data were augmented in three different ways. Firstly, to cover the different illumination conditions, the gamma correction technique was adopted. A gamma value of 0.5 and 1.25 were used to cover both the darker and the brighter illumination conditions, respectively. Secondly, to make the system robust against camera blur conditions, the OpenCV Gaussian blur technique was applied with a kernel size of 7 × 7 and 3 × 3. Finally, different noises were applied to the original eye patches using Gaussian and salt pepper noise techniques as shown in Fig. 4. The size of our dataset was increased by 3 times using the present data augmentation method as shown in Table 2. Results were compared before and after applying augmentation to the whole data (MPIIGaze and EYEDIAP) drawn in Fig. 5.

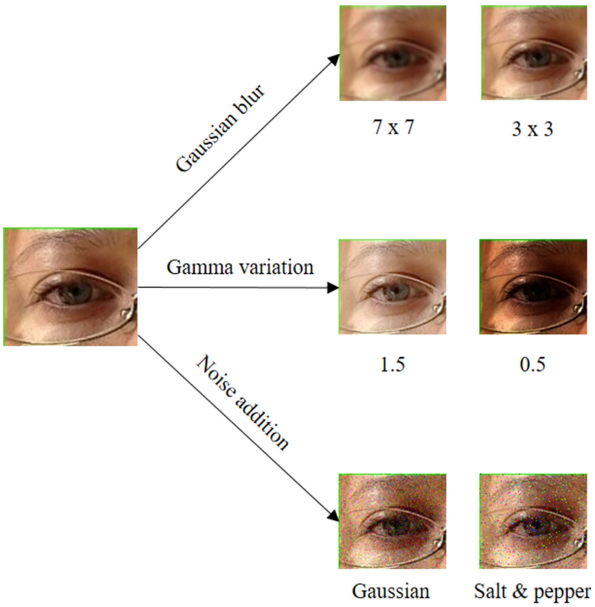

FIGURE 4. Sample images of how data augmentation is processed, which is best viewed in color.

TABLE 2. Datasets augmentation.

Dataset	Original (images)	After augmentation (images)
MPIIGaze	213,659	640,977
EYEDIAP	64,000	192,000

IV. EXPERIMENTS

Performance of this network was evaluated using two datasets. First, two eye gaze datasets used in this paper for training and evaluation are described. Secondly, data preparation and experimental details are explained in detail. Finally, detail information is provided on framework evaluation on both datasets, also single and multi-stream CNNs are compared followed by time complexity analysis in the end.

A. DATASETS

For both experiment and evaluation, two well-known datasets were used, such as the MPIIGaze [52] and the EYEDIAP datasets [21] as shown in Fig. 6. The former contained 213,659 images collected from 15 participants over several months. This dataset covered a wide range of head positions and illumination conditions. In each session, each subject was asked to look at 20 random positions. Each session was

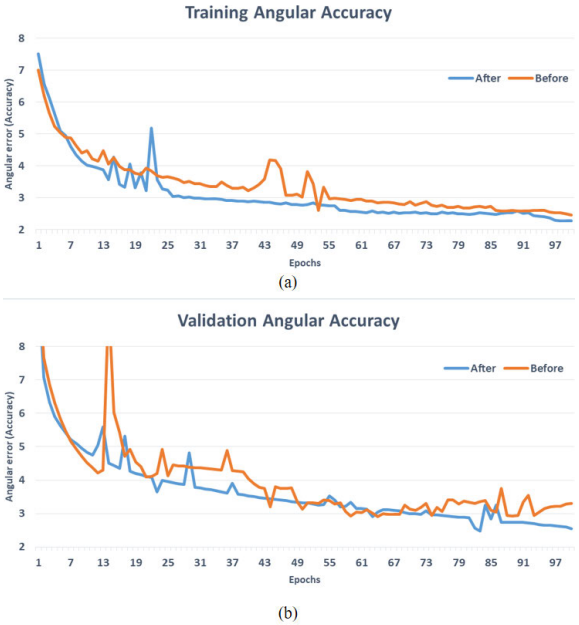

FIGURE 5. Training angular accuracy (error) (a) and validation angular accuracy (error) (b) of the proposed model for eye gaze estimation are shown before and after data augmentation.

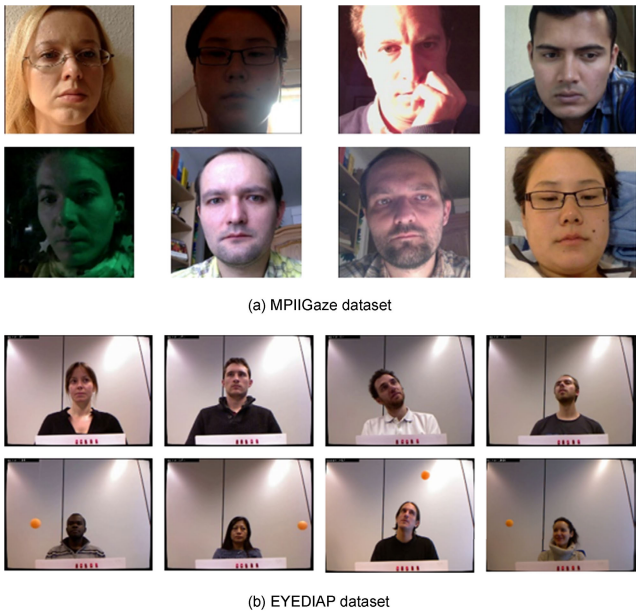

FIGURE 6. The sample images from (a) the MPIIGaze dataset and (b) the EYEDIAP dataset. The images in (a) are cropped images by removing the black background for visualization purpose. The first row in (b) are images with the stationary or movable head poses while gazing the target in a continuous fashion. The second row in (b) shows the sample images while gazing the floating target moving in 3D trajectory.

recorded during the daily routine of every subject without giving any specific instructions about how to record the sessions. The dataset contained diverse head poses, the illumination conditions, and the natural environment scenarios. So that each image had a full-face, head feature, and the 3D gaze target locations for each subject.

The latter was another large scale dataset for gaze estimation research. Sixteen participators were recruited for this project. Each session contained three different scenarios, which included a discrete screen target, continuous screen targets, and a 3D floating target. There were two head positions: one was a static position, and the other was a mobile head-position case. Three different cameras, which included a kinetic camera, a VGA camera (640 × 480), and an HD camera (1920 × 1080), were used for the eye data recording. For each participant, the videos were recorded in three different visual scenarios included a *Discrete screen target* (DS), a circle was drawn every 1.1 seconds uniformly, a *Continuous screen target* (CS), a circle moved along a random trajectory every 2 seconds, and a 3D *Floating target* (FT)), a ball that was 4cm in diameter attached to a stick with a thread that moved within a 3D region between the camera and the participant. To make the dataset robust against different head poses, the participants were instructed to record two videos (stationary (S) and mobile (moving head-position)) for each visual scenario. In this research, four videos from a VGA camera of each participant used for experimentation. The sample images from both datasets are shown in Fig. 6, which shows variations of both datasets in terms of the data collection, light intensity, head poses, and the camera angles used, respectively.

B. DATA PREPARATION

From the MPIIGaze dataset, the left and right eye patches were extracted from the full face dataset using perspective warping technique. As both eyes of a human looked at an object in a synchronized manner, the same ground truth vector for both the left and the right eye patches were used. The dataset was divided into training and evaluation with a ratio of 95% and 5%, respectively. Since the Rodrigues transformation was recommended to map a vector to an angle for both the head-pose and the gaze targets, the same method was followed in all experiments.

In terms of the EYEDIAP dataset, four videos were chosen, which included continuous screen targets and floating 3D targets videos, for both the stationary and the movable head-poses from each participant. The mid-point of both eye targets was chosen as the ground truth vector, and the eye patches were made by using the perspective warping technique used in III-A. We kept the same ratio of training and validation for the EYEDIAP dataset as well. Since the head positions were given for both datasets, they were used in these experiments.

C. EXPERIMENTAL DETAILS

The proposed framework was trained on a Linux system that has a NVIDIA GTX GForce 1070 GPU with python 3.6 and pytorch 1.0.1. The model was trained from scratch for 100 epochs with a batch size of 256. The weights of all the layers of the proposed network were initialized using the Kaiming He initialization [53]. Weight sharing was not used, because it decreased the performance. An Adam

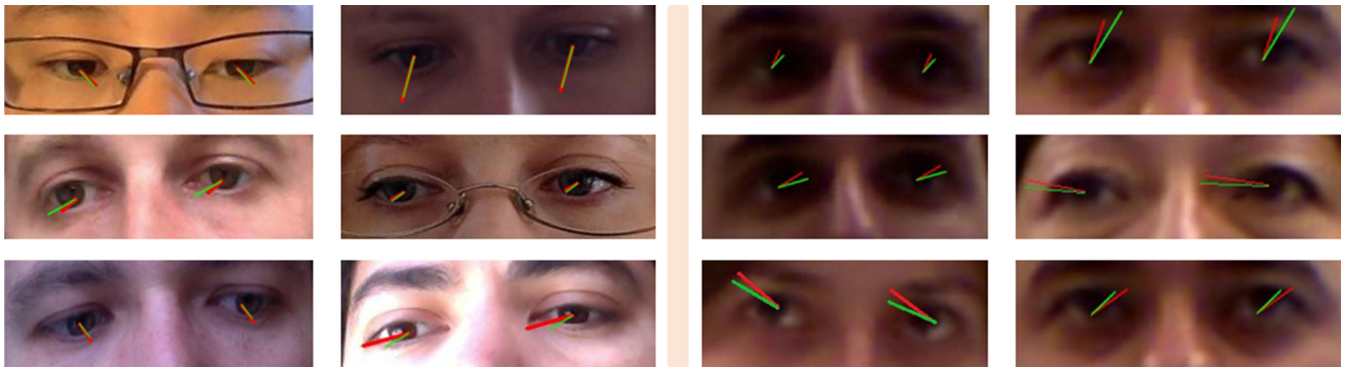

FIGURE 7. Sample 3D estimated gaze angle (green) and ground truth annotations (red) using the proposed method for the MPIIGaze dataset (left) and the EYEDIAP dataset (right), respectively.

optimizer [54] was used as an estimator with a learning rate of 0.01, a momentum 0.9, and weight decay.

D. EVALUATION METRICS

For gaze estimation the loss function was calculated by estimating the Euclidean distance between the predicted and ground truth gaze angle as shown in (6).

$$\mathcal{L}_{ed} = \sum_{i=1}^{N} \left\| \hat{g}_i - gi \right\|^2 \tag{6}$$

where N is the total number of images, $\hat{g}_i$ is the predicted regression angle of ith image, g_i is the actual ground truth of the ith image. $\mathcal{L}_{ed}$ is the averaged loss between the actual and predicted angle.

E. EVALUATION

The performance of this framework was evaluated using two standard eye gaze datasets, such as the MPIIGaze and the EYEDIAP datasets. A leave-one-out validation approach was used for MPIIGaze dataset. Since both eyes were used as input, two eye patches were extracted from the full-face image. The result was obtained by averaging all 15 participants. Given that the best state-of-the-art accuracy was 3.64 degrees [18], there was an improvement of 0.84 degrees since ours was 2.8 degrees. In addition, when data fusion technique described in Section IV-F was applied to this dataset, there was an additional improvement as shown in Table 3.

For the EYEDIAP dataset, the screen target sessions were used, as discussed in Section IV-A. The eye images were cropped using the same method as the MPIIGaze dataset. With a similar configuration to the MPIIGaze dataset, the EYEDIAP dataset was divided into 5-folds by splitting the 14 participants randomly into 5 groups. The initial accuracy was 3.77 degrees and yet it was further improved to 3.05 degrees by introducing the data fusion technique as described in Section IV-F, compared to the previous state-of-the-art, which was 3.23 degrees [40] on the EYEDIAP dataset (see Table 4).

TABLE 3. Comparison of the results with the state-of-the-art methods on the MPIIGaze dataset.

Methods	3D degrees error
Krafka K et al. 2016 [17]	5.6
Zhang X et al. 2017 [16]	5.4
Zhang X et al. 2017 [16]	4.8
Fischer T et al. 2018 [41]	4.3
Zhou X et al. 2019 [45]	4.64
Lemely J et al. 2019 [18]	3.64
Ours	**2.8**
Ours (data fusion)	**2.60**

TABLE 4. Comparison of the results with the state-of-the-art methods on the EYEDIAP dataset.

Methods	3D degrees error
Krafka K et al. 2016 [17]	8.3
Zhang X et al. 2017 [16]	6.0
Zhou X et al. 2019 [45]	6.02
Lian D et al. 2019 [55]	4.8
Ours	**3.77**
Ours (data fusion)	**3.05**

Our dual spatial weight mechanism-based multi-stream CNN network was compared with the previous state-of-the-art methods, such as a single face method, a face with spatial weight mechanism architecture [16], a recurrent based method that used both eye patches and the facial landmark [40], a deep ensemble network that used eye patches separately along with the head-position [41], and a multi-region method that employed both eyes, the face, and the face grid as the input [17]. Note that our method achieved the best result compared to all the previous methods by using just the eye patches and the head position, which is illustrated in Fig. 7.

F. DATA FUSION

Our experiment was further extended by involving data fusion of both datasets. First, both datasets were combined [1] and then

[1]The original MPIIGaze dataset was utilized instead of augmented data to make a fair comparison for data fusion

tested the trained model on the EYEDIAP dataset. Similar to previous experiments, the data were divided into similar fashion. During training, complete data were divided further into 90% and 10% for training and validation, respectively. The model converged and tested on the new data from the EYEDIAP dataset, it was noted that the angular error decreased further to 3.05 degrees, as described in Table 5.

Similarly, to analyze the results on the MPIIGaze dataset, another experiment was conducted keeping the same ratio of training and validation sets. Final results were produced by taking the mean value of the k-fold cross-validation as shown in Table 5.

TABLE 5. Model performance comparison with and without data fusion.

Dataset	Without data fusion	With data fusion
MPIIGaze	2.8	**2.60**
EYEDIAP	3.77	**3.05**

G. COMPARATIVE ANALYSIS

The experiments were conducted to compare a single and a dual spatial layer mechanism and the effect of the dual spatial layer on the model accuracy. In Fig. 8, it was observed that the models overall accuracy increased by introducing two spatial layers. We conducted the experiments using a single eye patch with a spatial layer and a head pose and evaluated the model on the MPIIGaze and the EYEDIAP datasets. Single spatial layer (single eye) results were compared with top models from [48] in Table 6. The results from the EYEDIAP dataset were worse than the MPIIGaze dataset due to the low resolution, and the high head pose variations of the EYEDIAP images.

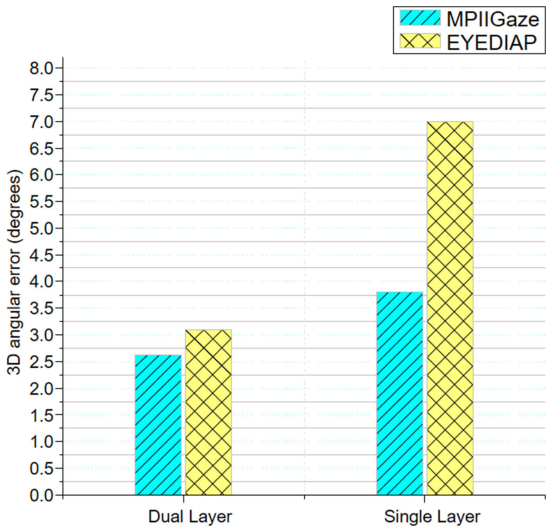

FIGURE 8. Comparative analysis between single spatial layer and dual spatial layer in our framework.

H. TIME COMPLEXITY ANALYSIS

The detail about processing speed for 3D gaze estimation is given in Table 7. These results were obtained by processing 6500 images and computing the average run-time. In Table 7, a comparison is given between a single and a batch of 256 images. The CPU is an Intel(R) Core(M) i7-4770 with eight kernels and 3.40GHz per kernel. The GPU is an Nvidia GForce GTX 1070. The program was written in Python and Pytorch. Note that the Pytorch process data is parallel to the GPU, which gave the best results for a batch of 256 images.

TABLE 6. Angular error (degree) using a single eye on MPIIGaze and EYEDIAP datasets. L, R, Avg denote the left, right eyes and the average of them, respectively.

	MPIIGaze			EYEDIAP		
-	L	R	Avg	L	R	Avg
Diff-NN-Ad [47]	4.61	4.56	4.59	**2.99**	**3.01**	**3.00**
Diff-VGG [47]	4.73	4.61	4.67	3.37	3.35	3.36
Ours (single eye)	**3.61**	**3.67**	**3.64**	6.96	7.66	7.31

TABLE 7. Processing speed (ms) of the proposed framework for 3D gaze estimation.

	No. of images	CPU	GPU
Run-time	1	12	6.64
	256	1190	29.4

V. DISCUSSION

Estimating gaze direction accurately from an image acquired with a mobile camera typically under the unstable illumination condition is not an easy task, given that the traditional way of estimating a human gaze was to gear up a massive eye movement setup, which was inconvenient and expensive. Of course, recent development of deep neural network makes it possible to estimate a reasonably accurate gaze direction from natural images. What the present study is trying to prove from our experimental results is that it is useful to employ a data science perspective in dealing with eye gaze datasets from data augmentation to data fusion.

The other important issue would be the image resolution of the gaze dataset. For instance, it seems that accuracy discrepancy between MPIIGaze and EYEDIAP (see Table 3 and 4) comes from the different image resolution within each eye patch, as observable from Fig 7. This suggests that the accuracy of gaze estimation by deep neural network could be further improved if there is any dataset that has more pixels within each eye patch.

VI. CONCLUSION

Detecting a human gaze during the interaction between people can play an essential role in social survival because one can understand what the other person intends during a conversation. Modern computer vision techniques with deep neural networks provide a new way to estimate a human gaze direction without gearing up such equipment.

In this work, we proposed a new 3D gaze estimation method from a natural face image taken with a desk-top computer that used a dual-channel convolutional neural network. The extensive evaluation was conducted with two standard gaze datasets. Our system has a spatial weight that is based

on a shallow network that outperformed all previous 3D gaze estimation methods. By using our method, we achieved an accuracy of 2.60 for MPIIGaze and 3.05 degrees for EYEDIAP, respectively. The improvement was 28% for the former and 4% for the latter over the state-of-the-art methods. Result suggests that our method is robust for any extreme head positions, gaze directions, and illumination conditions.

REFERENCES

[1] S. P. Liversedge and J. M. Findlay, "Saccadic eye movements and cognition," *Trends Cognit. Sci.*, vol. 4, no. 1, pp. 6–14, Jan. 2000. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364661399014187

[2] Q. Guillon, N. Hadjikhani, S. Baduel, and B. Rogé, "Visual social attention in autism spectrum disorder: Insights from eye tracking studies," *Neurosci. Biobehavioral Rev.*, vol. 42, pp. 279–297, May 2014. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0149763414000682

[3] S. Fan, Z. Shen, M. Jiang, B. L. Koenig, J. Xu, M. S. Kankanhalli, and Q. Zhao, "Emotional attention: A study of image sentiment and visual attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7521–7531.

[4] H.-O. Karnath and W. Huber, "Abnormal eye movement behaviour during text reading in neglect syndrome: A case study," *Neuropsychologia*, vol. 30, no. 6, pp. 593–598, Jun. 1992. [Online]. Available: http://www.sciencedirect.com/science/article/pii/002839329290062Q

[5] R. Santos, N. Santos, P. M. Jorge, and A. Abrantes, "Eye gaze as a human-computer interface," *Procedia Technol.*, vol. 17, pp. 376–383, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2212017314004836

[6] A. Kennedy, "Book Review: Eye Tracking: A Comprehensive Guide to Methods and Measures," *Quart. J. Experim. Psychol.*, vol. 69, no. 3, pp. 607–609, Mar. 2016. [Online]. Available: http://journals.sagepub.com/, doi: 10.1080/17470218.2015.1098709.

[7] S. P. Liversedge and J. M. Findlay, "Saccadic eye movements and cognition," *Trends Cognit. Sci.*, vol. 4, no. 1, pp. 6–14, 2000.

[8] M. A. Eid, N. Giakoumidis, and A. El Saddik, "A novel eye-gaze-controlled wheelchair system for navigating unknown environments: Case study with a person with ALS," *IEEE Access*, vol. 4, pp. 558–573, 2016.

[9] T. Ishikawa, S. Baker, I. Matthews, and T. Kanade, "Passive driver gaze tracking with active appearance models," Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep. CMU-RI-TR-04-08, Feb. 2004.

[10] P. Li, X. Hou, X. Duan, H. Yip, G. Song, and Y. Liu, "Appearance-based gaze estimator for natural interaction control of surgical robots," *IEEE Access*, vol. 7, pp. 25095–25110, 2019.

[11] R. Pieters and L. Warlop, "Visual attention during brand choice: The impact of time pressure and task motivation," *Int. J. Res. Marketing*, vol. 16, no. 1, pp. 1–16, Feb. 1999. http://www.sciencedirect.com/science/article/pii /S0167811698000226

[12] M. Wedel and R. Pieters, "A review of eye-tracking research in marketing," in *Review of Marketing Research*. Abingdon, U.K.: Routledge, 2017, pp. 123–147.

[13] R. Pieters, E. Rosbergen, and M. Wedel, "Visual attention to repeated print advertising: A test of scanpath theory," *J. Marketing Res.*, vol. 36, no. 4, pp. 424–438, Nov. 1999.

[14] R. Pieters and M. Wedel, "Attention capture and transfer in advertising: Brand, pictorial, and text-size effects," *J. Marketing*, vol. 68, no. 2, pp. 36–50, Apr. 2004. [Online]. Available: http://www.jstor.org/stable/30161988

[15] M. Wedel and R. Pieters, "Eye tracking for visual marketing," *Found. Trends Marketing*, vol. 1, no. 4, pp. 231–320, Aug. 2008. [Online]. Available: https://www.nowpublishers.com/article/Details/MKT-011

[16] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "It's written all over your face: Full-face appearance-based gaze estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*. Honolulu, HI, USA: IEEE, Jul. 2017, pp. 2299–2308. [Online]. Available: http://ieeexplore.ieee.org/document/8015018/

[17] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 2176–2184. [Online]. Available: http://ieeexplore.ieee.org/document/7780608/

[18] J. Lemley, A. Kar, A. Drimbarean, and P. Corcoran, "Convolutional neural network implementation for eye-gaze estimation on low-quality consumer imaging systems," *IEEE Trans. Consum. Electron.*, vol. 65, no. 2, pp. 179–187, May 2019.

[19] D. Lian, L. Hu, W. Luo, Y. Xu, L. Duan, J. Yu, and S. Gao, "Multi-view multitask gaze estimation with deep convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 10, pp. 3010–3023, Oct. 2019.

[20] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-Based Gaze Estimation in the Wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4511–4520. [Online]. Available: http://arxiv.org/abs/1504.02863

[21] 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 *Proc. Symp. Eye Tracking Res. Appl. (ETRA)*. Safety Harbor, Florida: ACM Press, 2014, pp. 255–258. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2578153.2578190

[22] J. Merchant, R. Morrissette, and J. L. Porterfield, "Remote measurement of eye direction allowing subject motion over one cubic foot of space," *IEEE Trans. Biomed. Eng.*, vol. BME-21, no. 4, pp. 309–317, Jul. 1974. [Online]. Available: http://ieeexplore.ieee.org/document/4120787/

[23] M. X. Huang, T. C. K. Kwok, G. Ngai, H. V. Leong, and S. C. F. Chan, "Building a self-learning eye gaze model from user interaction data," in *Proc. ACM Int. Conf. Multimedia (MM)*, Orlando, FL, USA: ACM Press, 2014, pp. 1017–1020. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2647868.2655031

[24] J. J. Bengoechea, J. J. Cerrolaza, A. Villanueva, and R. Cabeza, "Evaluation of accurate eye corner detection methods for gaze estimation," *J. Eye Movement Res.*, vol. 7, no. 3, pp. 1–8, 2014.

[25] Y. Zhang, A. Bulling, and H. Gellersen, "Pupil-canthi-ratio: A calibration-free method for tracking horizontal gaze direction," in *Proc. Int. Work. Conf. Adv. Vis. Interface (AVI)*. Como, Italy: ACM Press, 2014, pp. 129–132. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2598153.2598186

[26] Y. Zhang, A. Bulling, and H. Gellersen, "SideWays: A gaze interface for spontaneous interaction with situated displays," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst. (CHI)*. Paris, France: ACM Press, 2013, p. 851. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2470654.2470775

[27] S. Park, X. Zhang, A. Bulling, and O. Hilliges, "Learning to find eye region landmarks for remote gaze estimation in unconstrained settings," in *Proc. ACM Symp. Eye Tracking Res. Appl. (ETRA)*. Warsaw, Poland: ACM Press, 2018, pp. 1–10. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3204493.3204545

[28] Q. Huang, A. Veeraraghavan, and A. Sabharwal, "TabletGaze: Unconstrained appearance-based gaze estimation in mobile tablets," Aug. 2015, *arXiv:1508.01244*. [Online]. Available: http://arxiv.org/abs/1508.01244

[29] S. Baluja and D. Pomerleau, "Non-intrusive gaze tracking using artificial neural networks," in *Proc. NIPS*, 1993, pp. 753–760.

[30] O. Williams, A. Blake, and R. Cipolla, "Sparse and Semi-supervised Visual Mapping with the S^3GP," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2006, pp. 230–237.

[31] K. Tan, D. Kriegman, and N. Ahuja, "Appearance-based eye gaze estimation," in *Proc. 6th IEEE Workshop Appl. Comput. Vis. (WACV)*, Jan. 2002, pp. 191–195.

[32] C. H. Morimoto, A. Amir, and M. Flickner, "Detecting eye position and gaze from a single camera and 2 light sources," in *Proc. Object Recognit. Interact. Service Robots*, vol. 4, Aug. 2002, pp. 314–317.

[33] Z. Zhu and Q. Ji, "Eye gaze tracking under natural head movements," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 918–923.

[34] Z. Zhu, Q. Ji, and K. P. Bennett, "Nonlinear eye gaze mapping function estimation via support vector regression," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*. Hong Kong, China: IEEE, 2006, pp. 1132–1135. [Online]. Available: http://ieeexplore.ieee.org/document/1699089/

[35] D. W. Hansen and A. E. Pece, "Eye tracking in the wild," *Comput. Vis. Image Understand.*, vol. 98, no. 1, pp. 155–181, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S107731420400116X

[36] J. Chen and Q. Ji, "3D gaze estimation with a single camera without IR illumination," in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.

[37] F. Lu, Y. Sugano, T. Okabe, and Y. Sato, "Adaptive linear regression for appearance-based gaze estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 10, pp. 2033–2046, Oct. 2014.

[38] S. Park, A. Spurr, and O. Hilliges, "Deep pictorial gaze estimation," in *Proc. ECCV*, 2018, pp. 721–738.
[39] Y. Sugano, Y. Matsushita, and Y. Sato, "Appearance-based gaze estimation using visual saliency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 329–341, Feb. 2013.
[40] C. Palmero, J. Selva, M. A. Bagheri, and S. Escalera, "Recurrent CNN for 3D gaze estimation using appearance and shape cues," in *Proc. BMVC*, 2018, pp. 1–13.
[41] T. Fischer, H. J. Chang, and Y. Demiris, "Rt-gene: Real-time eye gaze estimation in natural environments," in *Proc. Comput. Vis. (ECCV)*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 339–357.
[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf
[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Sep. 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[44] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.* Springer, Sep. 2014, pp. 818–833.
[45] X. Zhou, J. Lin, J. Jiang, and S. Chen, "Learning a 3D gaze estimator with improved Itracker combined with bidirectional LSTM," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 850–855.
[46] Y. Cheng, F. Lu, and X. Zhang, "Appearance-based gaze estimation via evaluation-guided asymmetric regression," in *Proc. Comput. Vis. (ECCV)*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 105–121.
[47] S. Park, A. Spurr, and O. Hilliges, "Deep pictorial gaze estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 721–738.
[48] G. Liu, Y. Yu, K. A. Funes Mora, and J.-M. Odobez, "A differential approach for gaze estimation," 2019, *arXiv:1904.09459*. [Online]. Available: http://arxiv.org/abs/1904.09459
[49] W. Li, Q. Dong, H. Jia, S. Zhao, Y. Wang, L. Xie, Q. Pan, F. Duan, and T. Liu, "Training a camera to perform long-distance eye tracking by another eye-tracker," *IEEE Access*, vol. 7, pp. 155313–155324, 2019.
[50] X. Zhang, Y. Sugano, and A. Bulling, "Revisiting data normalization for appearance-based gaze estimation," in *Proc. ACM Symp. Eye Tracking Res. Appl. (ETRA)*, 2018, pp. 1–9.
[51] S. Lathuilière, P. Mesejo, X. Alameda-Pineda, and R. Horaud, "A comprehensive analysis of deep regression," Mar. 2018, *arXiv:1803.08450*. [Online]. Available: http://arxiv.org/abs/1803.08450
[52] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "MPIIGaze: Real-world dataset and deep appearance-based gaze estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 162–175, Jan. 2019.
[53] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.
[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[55] D. Lian, Z. Zhang, W. Luo, L. Hu, M. Wu, Z. Li, J. Yu, and S. Gao, "RGBD based gaze estimation via multi-task CNN," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jul. 2019, pp. 2488–2495.

ABID ALI received the bachelor's degree in electrical engineering from COMSATS University Islamabad (CUI) at Wah Campus, Wah Cantt, Pakistan, in 2016, and the master's degree from the Department of Computer Engineering, Sejong University, Seoul, South Korea. He worked as a Trainee Engineer with Pakistan Telecommunication Limited (PTCL), from 2017 to 2018. He is currently working as a Researcher of the Human–Computer Interaction Lab (HCI Lab), Sejong University. His major research domains are eye tracking, gaze estimation, feature extraction, computer vision, pattern recognition, and deep learning for object detection.

YONG-GUK KIM received the B.S. and M.S. degrees in electrical engineering from Korea University, Seoul, South Korea, in 1982 and 1984, respectively, and the Ph.D. degree in computational vision from the University of Cambridge, U.K., in 1997.

His work experience includes with LG Electronics and Korea Telecom (KT). From 1995 to 1996, he was a Research Fellow of the Helmholtz Robotics Institute, Utrecht, The Netherlands. From 1998 to 2001, he was a Research Associate with the Smith–Kettlewell Vision Institute, San Francisco, USA. He was the Director of a start-up incubator and the Dean of international affairs with Sejong University, Seoul. Since 2001, he has been with the Department of Computer Engineering, Sejong University, where he is currently a Full Professor. His research areas are facial expression recognition, driver drowsiness detection, fake emotion detection, human head pose estimation, and vision-based autonomous drone.