Appearance-based Gaze Estimation using Attention and Difference Mechanism

Murthy L R D
I3D Lab, CPDM,
Indian Institute of Science, Bangalore
lrdmurthy@iisc.ac.in

Pradipta Biswas
I3D Lab, CPDM,
Indian Institute of Science, Bangalore
pradipta@iisc.ac.in

Abstract

Appearance-based gaze estimation problem received wide attention over the past few years. Even though model-based approaches existed earlier, availability of large datasets and novel deep learning techniques made appearance-based methods achieve superior accuracy than model-based approaches. In this paper, we proposed two novel techniques to improve gaze estimation accuracy. Our first approach, I2D-Net uses a difference layer to eliminate any common features from left and right eyes of a participant that are not pertinent to gaze estimation task. Our second approach, AGE-Net adapted the idea of attention-mechanism and assigns weights to the features extracted from eye images. I2D-Net performed on par with the existing state-of-the-art approaches while AGE-Net reported state-of-the-art accuracy of 4.09° and 7.44° error on MPIIGaze and RT-Gene datasets respectively. We performed ablation studies to understand the effectiveness of the proposed approaches followed by analysis of gaze error distribution with respect to various factors of MPIIGaze dataset.

1. Introduction

The ability to estimate where a person is looking at opens up a plethora of opportunities. A few examples include understanding human vision like visual scan path analysis [15], reading analysis [7] and screening for dyslexia [27]. This technology also enable us to develop novel applications in human computer interaction across various domains like automotive [29], aviation [26] and accessibility [30, 8]. Even though various techniques like electrooculography existed earlier [6], use of imaging technology and advanced computer vision techniques enabled us to estimate gaze in non-intrusive manner. Commercial eye gaze trackers relies on infra-red based imaging to obtain eye-images to circumvent the effects of ambient illumination. These commercial gaze trackers claim to provide gaze accuracy of $<1.9°$error across 95% of population under real-world usage conditions

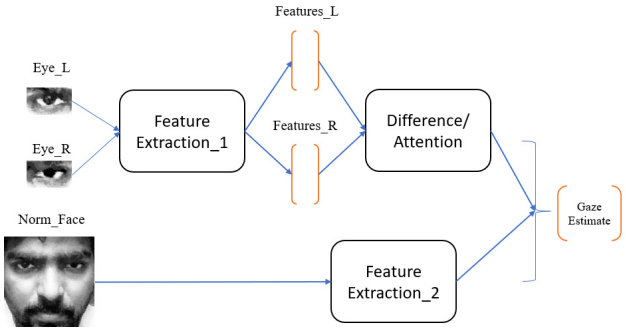

Figure 1. Illustration of the proposed approaches. Feature manipulation is performed on the features extracted from left and right eye images.

[2] and applications have been deployed using such gaze tracking systems to control cursor movement on a Windows PC [3].

Recent gaze estimation research focused on utilizing commodity hardware like webcam or the front-facing cameras available in ubiquitous mobile phones and tablet PC devices. Since this approach do not require additional hardware like infra-red illuminators, advancements in this direction allow eye gaze tracking to reach wider user groups.

Gaze estimation literature can be classified into feature-based, model-based, appearance-based approaches [19]. Numerous model-based and appearance-based methods were proposed for 3D and 2D gaze estimation problem and recent investigations [40] indicate that appearance-based methods obtained better gaze estimation accuracy than model-based approaches. This is made possible due to the availability of large datasets and novel deep learning techniques. In this paper, we proposed two feature manipulation approaches, I2D-Net and AGE-Net to improve gaze estimation accuracy. The first approach, I2D-Net relies on the intuition that omitting any person-dependent features which are extracted from eye images will help the deep neural network models to generalize well over unseen users. This approach assumes that these person-dependent

appearance-related features are present in both left and right eye images and hence obtaining a difference of left and right eye features should allow us to omit such person-specific information and retain only those features which are pertinent to person-independent gaze estimation.

Our second feature manipulation approach, AGE-Net is inspired from attention-mechanism, originally proposed for neural machine translation task. Attention-mechanism enables the model to search for a set of positions in a source sentence where the most relevant information is concentrated [4]. We adapt this idea of soft-selection of features for performing gaze estimation instead of using all features extracted from eye images. We believe that this is particularly relevant in cases where the input images contain variations in terms of appearance, illumination and head pose. Our proposed method have two branches, a feature extraction branch and an *Attention-branch* which produce feature vector and weights-vector respectively, of same dimension. The weight-vector assigns weights to each feature of the feature vector indicating their relevance for a given image.

We performed experiments using these two approaches on MPIIGaze [42] and RT-Gene [16] datasets. We demonstrated that the proposed feature manipulation techniques achieved state-of-the-art results under standard evaluation protocols. We analyzed the effectiveness of the proposed approaches using ablation studies. We also compared the performance of AGE-Net with I2D-Net under factors like illumination, head pose variations.

We summarize the contributions of this paper below:

- We proposed I2D-Net using difference layer and AGE-Net adapting attention-mechanism for gaze estimation.
- I2D-Net achieved on par performance with existing state-of-the-art methods while using fewer parameters ($\sim$87M).
- The proposed AGE-Net achieved a state-of-the-art performance of 4.09° and 7.4° error on MPIIGaze and RT-Gene datasets using $\sim$105M parameters.

In the next section 2, we present a review of existing gaze estimation approaches. In section 3, we present the proposed approaches with details related to the network architectures. We present the experiments conducted on MPIIGaze and RT-Gene datasets along with results and ablation studies in section 4. We present analysis of both proposed approaches in section 5. Section 6 contains the discussion on the presented work followed by conclusion in section 7.

2. Related Work

2.1. Feature and Model-Based Approaches

Commercial gaze-trackers like Tobii X-series [2] are predominantly feature-based systems. An external IR light source using either bright or dark pupil method [1] is used to obtain eye features like corneal reflections [18] to perform gaze estimation.

On the other hand, model-based methods extract visual features from eye images like pupil center, iris contours and eye corners to fit a geometric 3D eye model to perform gaze estimation. Early model-based methods used infra-red illuminators and high-resolution cameras [20, 37] but recent approaches [33, 28, 5] overcome such requirements by extracting features from webcam images. They also empower their feature detectors with machine learning approaches to obtain robustness with respect to illumination variations.

2.2. Appearance-Based Approaches

In contrast to the earlier mentioned approaches, appearance-based methods attempts to directly map the images captured using commodity cameras to gaze direction vectors without any handcrafted features. These appearance-based methods are strongly supported by the creation of large datasets [17, 42, 16, 21] and advancements in deep learning techniques. We can classify the appearance-based approaches proposed so far into single channel and multi-channel methods.

One of the first attempts of appearance-based gaze estimation was GazeNet [42], a single channel approach where a single eye image is used as the input to an architecture based on 16-layer VGG CNN. Head pose information was concatenated to the first fully connected layer after convolutional layers. GazeNet reported a 5.4°mean angle error on the evaluation subset of MPIIGaze, termed as MPIIGaze+. This work was followed by Spatial-Weights CNN, another single-channel approach [41] where full face images were provided as input instead of eye crops. They used spatial weights technique to give significance for those regions of face which are pertinent for gaze estimation.

As an alternative to single-channel approaches, numerous multi-channel approaches were proposed. iTracker [23] was one of the first multi-channel architecture which uses left eye image, right eye image, face crop image and face grid information as inputs. Multi-Region Dilated-Net proposed by Chen and Shi [9] employed dilated convolutions in their proposed model and uses both eye images along with face image as inputs. This approach also reported the same result of 4.8° mean angle error as [41] did on MPIIGaze+ in cross-participant evaluation. Further, they extended their work by proposing GEDDNet [10], which uses gaze decomposition along with dilated convolutions and reported 4.5°error on MPIIGaze+. Most recent work based on multi-channel architecture is FAR*-Net [12] which proposed to utilize the asymmetry between two eyes of same person to obtain gaze estimates. In this work, they generated confidence scores for the gaze estimates obtained from two eye images to choose the more accurate prediction.

Wang et al [34] proposed to use bayesian framework for better generalization performance of gaze estimates than appearance-based approaches. They proposed to use an adversarial component along with CNN-based gaze estimator to learn generalizable gaze-responsive features.

Most of the above mentioned works used the features extracted from eyes and face images directly for the gaze vector regression. We observed that only Spatial-weights CNN [41] and Bayesian approach [34] proposed to employ feature extraction that focuses on obtaining features pertinent to gaze estimation. Due to the huge variance in terms of illumination and head pose variations in addition to the inherent person-specific variations, we believe that either person-specific or image-specific feature-manipulation can lead to obtain more accurate gaze estimates.

In the following section, we introduced our proposed feature manipulation approaches based on difference and attention mechanism.

3. Proposed Approach

3.1. I2D-Net

I-Gaze estimation using dilated and differential layer network (I2D-Net) primarily relies on two modules. Chen and Shi [9] showed that extracting features using dilated convolutions instead of regular convolutions improve gaze estimation accuracy. They argued that a series of maxpooling layers might not capture the finer details in eye images which are significant for gaze estimation. They also argued that the dilated convolutions preserve the resolution of feature maps while obtaining larger receptive fields on contrary to the use of maxpooling layers where larger receptive fields are obtained at the cost of feature map resolution.

The second module is the differential layer that obtains an absolute difference of the left and right eye features which are extracted using dilated convolutions. Zeiler and Fergus [38] demonstrated that shallow layers of CNNs capture low-level information such as edges and low-level contours while deeper-layers of CNNs attempts to learn higher level features like parts of the object with significant pose variation. In a specific example, they showed that the eyes and nose of the dog has been observed when the feature map of Layer 4 of AlexNet [24] was visualized. Based on this observation, we proposed the following approach.

We used shared-convolutional layers to extract features from the normalized right and left eye images. This feature extraction network is illustrated in figure 2c where we employ dilated convolutional layers. We changed the number of feature maps in each layer and dilation rate from our baseline [9] and added dilated convolutions to face channel as well (figure 2e).

These obtained feature maps might contain higher level features that encode information about various portions of the eye images like eyeball, sclera region or brow region. Such finer details vary from person to person and are present in both left and right eyes images. We argue that obtaining the absolute difference of these extracted features from left and right eye images removes common, redundant appearance-related information and hence retains only relevant features from both eyes. We posit that the resultant feature vector acts as a better feature transformation than the case where the features from both eyes are concatenated for subsequent fully connected layers. The entire network architecture of I2D-Net is illustrated in Fig 2a.

Reader may note that this proposed approach is fundamentally different from the Diff-NN [25]. We focused on improving person-independent gaze estimation task and we do not rely on any person-specific calibration samples as [25] did. Further, Diff-NN proposes to use images belonging to same eye (either left or right) and train the model to learn the gaze difference. We propose to obtain the difference between features extracted from left and right eye of same person to circumvent person-dependent features.

3.2. AGE-Net

Attention-based Gaze Estimation Network (AGE-Net) proposes to adapt the attention-mechanism which was used in Natural Language Processing (NLP), Computer Vision [36] tasks, speech systems [13], recommender systems [32] and to predict the steering angle for self-driving cars [22]. In neural machine translation task, encoder generates the annotations for a given input sentence which in turn shall be used by decoder to generate the output sentence in a different language. Bahdanau et al proposed attention mechanism [4] to select specific words from input sentence those are significant for the translation task. They proposed to use weighted sequence of annotations instead of using them as is to generate output sequence.

Adapting this idea to gaze estimation task, we propose to assign weights to the features extracted from eye images. We propose to add an *Attention-branch* in parallel to feature-extraction branch to perform the intended feature manipulation. Both feature extraction branch and attention branch contains shared convolutional layers that takes eye images as the input. Feature extraction branch, indicated in figure 2c produces feature vectors from both left and right eye images while attention branch provides the necessary weight vectors for both left and right eye features. We used sigmoid activation function for the last layer of the attention-branch as illustrated in figure 2d to obtain weight-vectors with values in the range of (0-1). The feature vectors obtained from both eyes are multiplied with the corresponding weight vectors to obtain weighted features which will be passed further through the network for the regression task. The AGE-Net architecture is illustrated in Fig 2b.

Figures 2c, 2d and 2e indicate various parameter values

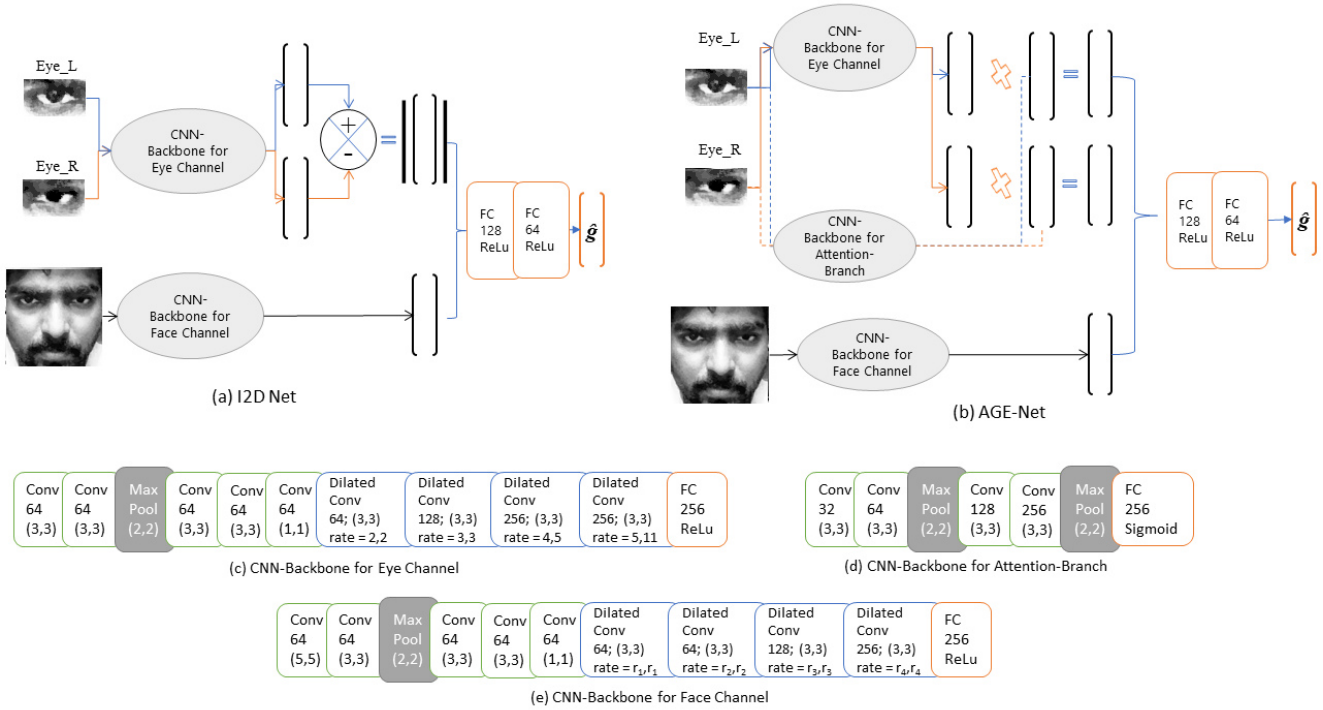

Figure 2. Network Architecture of proposed approaches. (2a) I2D-Net architecture (2b) AGE-Net architecture (2c) CNN-Backbone for Eye Channel (2d) CNN-Backbone for Attention-Branch (2e) CNN-Backbone for Face-Channel.

for each layer of the architecture like feature map size, kernel size and activation function for fully connected layers. We used ReLu activation function for all convolution and dilated convolution layers. The dilation rate parameters r_1, r_2, r_3 and r_4 for the face channel(figure 2e) assumes different values for AGE-Net and I2D-Net. We used 3, 5, 7 and 11 for r_1, r_2, r_3 and r_4 respectively for AGE-Net. In case of I2D-Net, these parameters take 2, 3, 5 and 11 values as the dilation rates. The normalized face images are first passed through the first six layers of VGG-Net [31] pre-trained on the ImageNet dataset [14] before feeding them to the CNN-backbone of face channel (figure 2e). Each layer in our proposed architectures is followed by batch normalization.

In the next section, we present the experiments conducted using both proposed approaches.

4. Experiments - I2D-Net & AGE-Net

4.1. Datasets

MPIIGaze

We conducted experiments on MPIIGaze, which was collected in real-world conditions with illumination and head poses variations. The dataset was collected with 15 people from diverse ethnic backgrounds and includes appearance-variations like wearing spectacles. We normalized face and eye images from the evaluation subset of MPIIGaze [42] using landmark annotations [41]. The evaluation subset of the dataset contained 45000 samples in total with 3000 samples from each person.

We utilized the method mentioned in [39] for normalizing the images and ground truth gaze labels. The normalization process, in summary, cancels out roll component of head pose in the captured image and positions the image at a desired distance d_v from the virtual camera with a focal length of f_v. In addition to the image normalization, we also transformed ground truth gaze label from camera co-ordinate system to normalized space with angular representation. We used d_v as 600 for eye image normalization and 1000 for face images. We selected the resolution of normalized eye images to be 36x60 while the normalized face images are of 120x120. Further, we selected f_v for both eye and face normalization to be 960. We applied histogram equalization on the resultant images to obtain the normalized eye and face images which shall be used for the subsequent stages.

RT-Gene

RT-Gene [16] dataset contains 122,531 images of 15 participants using wearable eyetracking glasses. Unlike MPIIGaze dataset where participants are sitting near to their computers, they are located at 0.5 to 2.9 meters from

Model	MPIIGaze	RT-Gene
iTracker (AlexNet) [23]	5.6°	-
MeNet [35]	4.9°	-
Spatial-Weights CNN [41]	4.8°	10.0°
Dilated-Net [9]	4.8°	-
RT-GENE (1 model) [16]	4.8°	-
RT-GENE (4 model) [16]	4.3°	8.6°
FAR* Net [12]	4.3°	8.4°
Bayesian Approach [34]	4.3°	-
I2D-Net **(Proposed)**	**4.3°**	**8.44°**
AGE-Net **(Proposed)**	**4.09°**	**7.44°**

Table 1. Comparison with existing Gaze estimation models.

the camera during this dataset creation. This dataset also has higher variation in head pose and gaze angles. Since the images captured in RT-Gene dataset contains eyetracking glasses along with the person, they used semantic inpainting to paint the area covered by eyetracking glasses with skin texture. Hence, the authors provided both original and in-painted version of the images after normalizing them. The resolution of the normalized eye and face images is 36x60 and 224x224 respectively. We did not do any further processing of these images apart from resizing of face images to 120x120. We observed noises in the in-painted set as [12] reported and hence used only the original dataset for experiments. We used grayscale images for all our experiments on both datasets.

4.2. Training & Results

We performed leave-one-out cross validation on MPIIGaze as mentioned in other works [41, 12, 9] using both proposed models. We leave one participant out for testing and considered other 14 participants for training. We implemented proposed models using Tensorflow and Keras. We have used 15% of the training data for validation split. Since most of the gaze labels are less than 1, we scaled them up by 100 and we used mean square error as the loss function. We have trained each model for 30 epochs with a batch size of 32 and we used Adam optimizer.

We conducted experiments on RT-Gene dataset as per the evaluation protocol provided by the dataset. We divided the original dataset into 3 folds and we performed a 3-fold cross validation. We followed similar training procedure as we did for MPIIGaze dataset, but we trained the model for 50 epochs due to the higher number of training samples.

We presented experimental results of the proposed approaches on both MPIIGaze and RT-Gene datasets and compared them against the various gaze estimation approaches which either use face or multi-channel approach in Table 1. The proposed I2D-Net achieved 4.3 $\pm$ 0.97 and 8.44 $\pm$ 1.08 degree mean angle error on MPIIGaze and RT-Gene datasets respectively. I2D-Net is on par with existing state-of-the-art methods like FAR* Net [12] and bayesian adversarial learning method [34] on both datasets. Since we used dilated convolutions in feature extraction phase, we consider Dilated-Net [9] as our baseline and hence the proposed differential feature transformation in I2D-Net achieved 10% improvement over the baseline. Diff-NN [25] reported 4.59°average error after applying their adaptation method with 9 reference samples and 4.64 with their default parameter setting. Hence the proposed I2D-Net with a difference layer of features extracted reported superior performance than the Diff-NN which proposes to use two same eye images of a person to learn the gaze difference.

Architecture	Mean Angle Error
No Attention	4.54°
Eye+Attention	4.64°
AGE-Net w/o Dilated Conv	4.24°

Table 2. Ablation study results of AGE-Net on MPIIGaze

The other proposed approach, AGE-Net achieved a state of the art performance of 4.09 $\pm$ 0.9 degree and a 7.44 $\pm$ 1.59 degree mean angle error on MPIIGaze and RT-Gene datasets respectively. We infer that the proposed attention-branch which assigns weights to the extracted features improved the overall mean angle error by 14.8% over the baseline. AGE-Net also achieved around 5% and 11.5% improvement on MPIIGaze and RT-Gene datasets respectively over the existing state of the art.

In addition to that, the proposed approaches I2D-Net ($\sim$87M) and AGE-Net ($\sim$105M) utilizes less number of parameters than their counterparts like FAR* Net [12] ($\sim$848M), Spatial Weights CNN [41] ($\sim$196M), RT-GENE [16] ($\sim$122M) and GEDD Net [10] ($\sim$107M) and hence these approaches result in lower memory footprint.

4.3. Ablation Study

We investigated the significance of various modules that form the proposed architectures by performing ablation studies using MPIIGaze dataset.

AGE-Net Table 2 summarizes ablation study results on AGE-Net. First, we experimented with the CNN-backbones for face and eye channels alone as illustrated in figure 2c and figure 2e without the proposed attention branch and observed a 4.54°error. We have changed the number of features and the dilation rates used in each layer from the baseline [9] and formed these CNN backbones. Yet, the performance is on par with GEDDNet [10] which utilizes both dilated convolutions and gaze decomposition technique. Next, we experimented with CNN backbone for eye channel along with attention branch without using face information and observed a 4.64°error.

We then experimented by investigating the importance

Model	Mean Angle Error
Gaze-Net	5.4°
AR-Net	5.65°
ARE-Net	5.02°
AGE-Net (Eyes Only)	**4.64°**
I2D-Net (Eyes Only)	5.08°

Table 3. Comparing eye image-based methods on MPIIGaze

of dilated convolutions in the proposed AGE-Net architecture. We redesigned the architecture to achieve 90% of the input size as the receptive field without using dilated convolutions. First, all dilated convolution layers in both face and eye channels are replaced with regular convolutional layers. We added two more maxpooling layers in eye channel, one each after seventh and eighth convolutional layers. Furhter, we added three more maxpooling layers in face channel, one each after fourth, sixth and eighth convolutional layers. We obtained a 4.24°error using this approach which indicated that dilated convolutions indeed help the proposed network to achieve better performance. In summary, presence of attention branch to the CNN-backbones improves gaze estimation accuracy by 10% and presence of dilated convolutions improves the accuracy by 3.5%.

I2D-Net We investigated the significance of difference layer and face channel for I2D-Net. We first experimented with the depicted CNN-backbone in figure 2a including both eye channel and face channel with out difference layer. This model recorded a 4.6°error, a slight drop from the CNN-backbone used for AGE-Net, possibly due to the different dilation rates. This indicates that the presence of difference layer improves the performance by 6.5%.

Excluding face channel from the proposed I2D-Net resulted in 5.08°error. We observed that omitting face information from both proposed architectures resulted in significant drop in performance. Yet, AGE-Net without face channel obtained better gaze error when compared to other eye-only methods. As reported in Table 3, eye image-based methods like Gaze-Net [42], AR-Net and ARE-Net [11] reported 5.4, 5.65 and 5.02 degree error respectively against 4.64 and 5.08 degrees achieved by AGE-Net and I2D-net respectively without face information.

5. Analysis

In this section, we analyzed the proposed methods with respect to individual participant, illumination level, horizontal difference of mean illumination, gaze point and head pose in the MPIIGaze since factors like illumination variations are not reported for RT-Gene. We also present comparative analysis between AGE-Net and I2D-Net to understand the areas where AGE-Net performed better than I2D-Net.

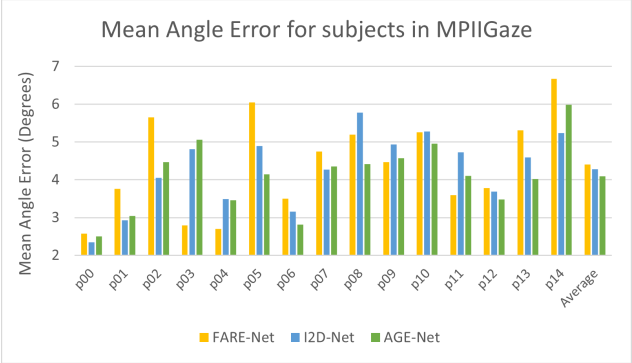

Figure 3. Participant-wise accuracy of MPIIGaze

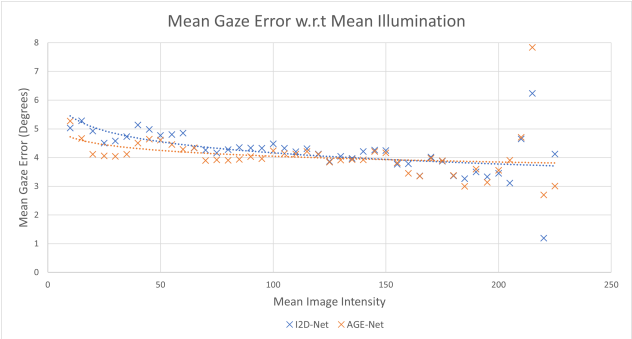

Figure 4. Gaze error distribution w.r.t. Illumination

5.1. Participant-wise Analysis on MPIIGaze

We analyzed the performance of various gaze estimation methods on each participant of the MPIIGaze dataset. We observed that both proposed approaches perform better than the baseline except for p03, p07 and p09. Further in figure 3, we compare our proposed methods with FARE Net [12]. Out of 15 participants, AGE-Net performed better for 11 participants while I2D-Net performed better for 9 participants than FARE-Net. We undertook paired t-tests which revealed that both proposed approaches, I2D-Net ($t[14] = 2.17$, $p = 0.047$, Cohen's $d = 0.5$) and AGE-net ($t[14] = 2.84$, $p = 0.01$, Cohen's $d = 0.7$) performed statistically significantly better than the baseline [9]. Further, it is observed that AGE-Net without face channel performed statistically significantly better ($t[14]=2.72$, $p=0.016$, Cohen's $d = 0.36$) than another eye-image based method, ARE-Net [12].

We also analyzed yaw and pitch errors for each participant. We observed absolute mean yaw error of 2.66 and 3.042 degrees and absolute mean pitch error of 2.56 and 2.98 degrees for AGE-Net and I2D-Net respectively.

5.2. Effect of Illumination

We analyzed the gaze errors obtained during cross-participant evaluation on MPIIGaze dataset with respect to

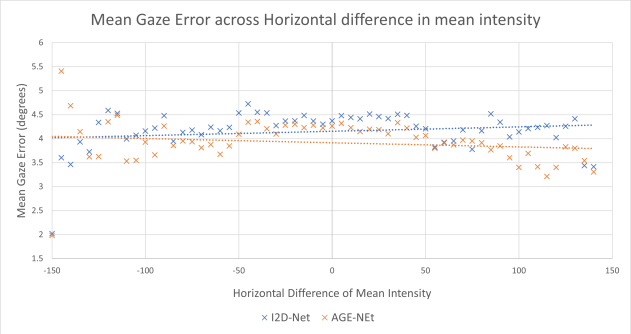

Figure 5. Gaze Error Distribution w.r.t. Mean intensity difference

Pitch (degree) \ Yaw (degree)	-43.4	-25.7	-8.1	9.6	27.2	44.9	Average
6.7	4.93	4.09	3.82	5.20	1.80		**3.97**
-1.9	3.72	4.19	4.07	4.12	5.10		**4.24**
-10.6	5.35	4.06	4.06	4.06	4.54		**4.42**
-19.2	6.07	4.10	4.04	4.05	5.26		**4.70**
-27.8	6.46	4.41	4.24	4.22	6.00		**5.07**
-36.5							
Average	**5.31**	**4.17**	**4.05**	**4.33**	**4.54**		

Figure 6. Gaze error analysis w.r.t Gaze region-AGE-Net

Pitch (degree) \ Yaw (degree)	-43.4	-25.7	-8.1	9.6	27.2	44.9	Average
6.7	5.31	5.05	4.82	6.39	2.63		**4.84**
-1.9	4.49	4.41	4.28	4.38	5.97		**4.71**
-10.6	5.08	4.32	4.30	4.24	5.13		**4.62**
-19.2	6.32	4.42	4.26	4.31	4.98		**4.86**
-27.8	4.62	4.50	4.50	4.34	5.75		**4.74**
-36.5							
Average	**5.16**	**4.54**	**4.43**	**4.73**	**4.89**		

Figure 7. Gaze error analysis w.r.t Gaze region - I2D-Net

the mean intensity of the face image. For this purpose, we grouped the mean intensity values into bins of width 5 and grouped the images accordingly. We obtained mean gaze errors corresponding to each bin and we plot the same for both I2D-Net and AGE-Net in figure 4. For illustration purpose, we omitted the mean gaze error for the first bin [0-5) since I2D-Net and AGE-Net reported 23.9 and 18.9 degree error respectively. We observed that for the mean intensity between 60 and 210, gaze errors of I2D-Net and AGE-Net lie with in the range of 3.96 ± 0.4 and 3.85 ± 0.33 degree respectively indicating the models' generalization performance across wide range of illumination variation. We also observed that the gaze error decreases with brighter illumination but increases when the mean intensity falls beyond the range of (5-200). This might be due to the less number of training samples in the mentioned range as reported in the dataset characteristics [42].

5.3. Effect of Horizontal Difference of Illumination

One of the challenging scenario that MPIIGaze dataset captured is the horizontal difference of illumination across the face image. We analyzed the performance of the proposed approaches under such scenarios. We used similar approach as described in sec 5.2 and we plotted the results in figure 5. We could not visually inspect the difference between the trend lines of both proposed approaches from figure 4, but in figure 5, it is evident that the AGE-Net reported lower mean gaze error than I2D-Net. We observed that AGE-Net reported lower gaze error than I2D-Net in the range of [-135, -50] and [80, 135]. The trend line of the mean gaze error across the horizontal intensity difference for both proposed approaches is close to a flatline, reiterating the models' generalization ability across the range of horizontal illumination difference.

5.4. Effect of Gaze Direction

The performance of a gaze estimation system may also vary based on where the person is gazing. We investigated the gaze error of our proposed models based on the ground truth gaze direction. We split the ground truth gaze pitch and yaw components into 5 bins each and hence we split the entire normalized gaze space into a grid of 25 cells. We have grouped the input images falling in these 25 cells and obtained mean gaze errors for each cell. In figure 6 and figure 7, we present the mean gaze errors by AGE-Net and I2D-Net respectively across all the 25 cells. We can observe a clear monotonic increase in the gaze error for AGE-Net in figure 6 as the pitch angle increases and such pattern is not observed in figure 7 for I2D-Net though. Both proposed methods reported similar trend of mean gaze error with respect to change in yaw angles and report best accuracy along the central column.

5.5. Effect of Head Pose

We investigated proposed models' performance with respect to head pose of the participant which is a significant variable in the dataset. Similar to the analysis based on gaze location presented in 5.4, we have split the entire yaw and pitch range of head orientation into 50 bins each and hence the normalized head pose space is split into a total of 2500 cells. We have clustered the images corresponding to these cells and we considered the cells with at least three images. We obtained mean gaze errors from both AGE-Net and I2D-

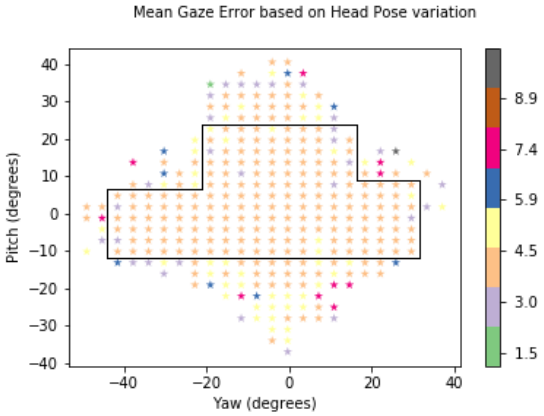

Figure 8. Gaze Error w.r.t Head Pose-AGE-Net

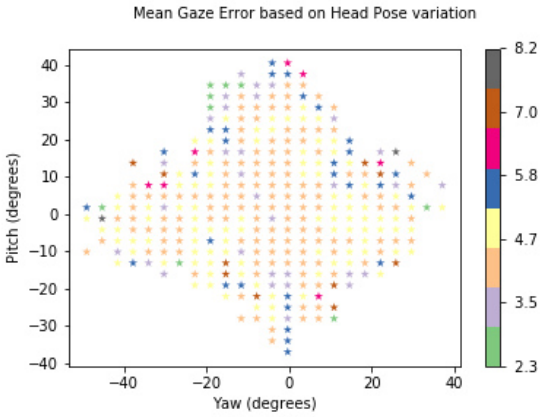

Figure 9. Gaze Error w.r.t Head Pose-I2D-Net

Net for each cell and plotted in figure 8 and figure 9 respectively. In comparison, AGE-Net have more uniform gaze error distribution with respect to head pose than I2D-Net. Further, clusters with high mean gaze error can be located only on the boundaries of the head pose distribution indicating that extreme head poses results in high gaze prediction errors. For AGE-Net, this means that a clear region of head pose variations, indicated with black lines in figure 8, is identified which shall result in an almost uniform distribution of gaze error. Such regions are helpful in defining the range of operation when this model is deployed for the real-time usage.

6. Discussion

Our first feature manipulation approach, I2D-Net reported on-par performance with existing state-of-the-art methods like FAR* Net on both MPIIGaze and RT-Gene datasets while the other feature manipulation approach, AGE-Net reported a state-of-the-art performance on both datasets. In addition to that, AGE-Net also found to have better generalization performance over I2D-Net based on the analysis presented in section 5. Both methods reported a flat trend of gaze errors across the range of horizontal difference of illumination and a linear trend across mean illumination range. Further, the distribution of gaze error with respect to head pose using AGE-Net provided a clear boundary with uniform error which is useful in defining range of operation when used for real-time interaction. AGE-Net also reported a clear trend of increasing error as the pitch component of gaze direction increases. This might be due to the occlusion of eyeball region by eyelids when pitch component of gaze vector is higher. Since the majority of existing laptops and tablet computers have cameras placed above the screen, this limitation need to be overcome to obtain uniform gaze error across the viewing range.

We experimented AGE-Net architecture without dilated convolutions with ∼27M parameters and observed an error of 4.24 compared to AGE-Net's 4.09 and ∼105M parameters. This approach can be taken further to investigate if we can improve gaze error with less number of parameters. Such smaller memory footprint models can be useful for achieving less latency and less stringent requirements of high end GPUs for real-time gaze estimation. A real-time demonstration of the current AGE-Net system can be seen at **https://youtu.be/2pyX6O2xTFw**. On the other hand, we did not use attention-branch for face channel in our proposed architecture which is another avenue for exploration.

In section 5, we compared and analyzed the performance of the proposed approaches with respect to various parameters in MPIIGaze dataset. We performed cluster-wise analysis on the gaze errors rather than individual-image wise analysis. Even though we obtained a macro-level trend of gaze errors with respect to each parameter, further investigation is required to understand the factors behind images which reported high error. Such failure mode analysis is imperative to understand the influence of the inherent visual-optical axis offset and model's shortcomings in the observed gaze error to build robust gaze estimation models.

7. Conclusion

In this paper, we proposed feature manipulation techniques based on differential feature vector and attention-mechanism for appearance-based gaze estimation task. The proposed I2D-Net reported on-par performance and AGE-Net reported superior performance when compared with existing state-of-the-art methods on both MPIIGaze and RT-Gene datasets. Our approaches also shown to be robust to various factors like illumination, head-pose and horizontal difference of mean intensity. Further, we have demonstrated the significance of the proposed techniques using ablation studies. Finally, we discussed the implications and prospective extensions to our proposed approaches to further improve the gaze estimation accuracy.

References

[1] Dark and bright pupil tracking. https://www.tobiipro.com/learn-and-support/learn/eye-tracking-essentials/what-is-dark-and-bright-pupil-tracking/. Accessed: 2021-03-27.

[2] Tobii dynavox pceye mini. http://tdvox.web-downloads.s3.amazonaws.com/PCEye/documents/TobiiDynavox_PCEyeMini_UserManual_v1-2_en-US_WEB.pdf. Accessed: 2021-03-08.

[3] Windows control - tobii dynavox. https://www.tobiidynavox.com/software/windows-software/windows-control-software/. Accessed: 2021-03-19.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[5] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE, 2018.

[6] Rafael Barea, Luciano Boquete, Manuel Mazo, and Elena López. System for assisted mobility using eye movements based on electrooculography. *IEEE transactions on neural systems and rehabilitation engineering*, 10(4):209–218, 2002.

[7] David Beymer and Daniel M Russell. Webgazeanalyzer: a system for capturing and analyzing web reading behavior using eye gaze. In *CHI'05 extended abstracts on Human factors in computing systems*, pages 1913–1916, 2005.

[8] Maria Borgestig, Jan Sandqvist, Gunnar Ahlsten, Torbjörn Falkmer, and Helena Hemmingsson. Gaze-based assistive technology in daily activities in children with severe physical impairments–an intervention study. *Developmental Neurorehabilitation*, 20(3):129–141, 2017.

[9] Zhaokang Chen and Bertram E Shi. Appearance-based gaze estimation using dilated-convolutions. In *Asian Conference on Computer Vision*, pages 309–324. Springer, 2018.

[10] Zhaokang Chen and Bertram E Shi. Geddnet: A network for gaze estimation with dilation and decomposition. *arXiv preprint arXiv:2001.09284*, 2020.

[11] Yihua Cheng, Feng Lu, and Xucong Zhang. Appearance-based gaze estimation via evaluation-guided asymmetric regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 100–115, 2018.

[12] Yihua Cheng, Xucong Zhang, Feng Lu, and Yoichi Sato. Gaze estimation by exploring two-eye asymmetry. *IEEE Transactions on Image Processing*, 29:5259–5272, 2020.

[13] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. *arXiv preprint arXiv:1506.07503*, 2015.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[15] Sukru Eraslan, Yeliz Yesilada, and Simon Harper. Eye tracking scanpath analysis techniques on web pages: A survey, evaluation and comparison. *Journal of Eye Movement Research*, 9(1), 2016.

[16] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 334–352, 2018.

[17] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 255–258, 2014.

[18] Elias Daniel Guestrin and Moshe Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on biomedical engineering*, 53(6):1124–1133, 2006.

[19] Dan Witzner Hansen and Qiang Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):478–500, 2009.

[20] Takahiro Ishikawa. Passive driver gaze tracking with active appearance models. 2004.

[21] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6912–6921, 2019.

[22] Jinkyu Kim and John Canny. Interpretable learning for self-driving cars by visualizing causal attention. In *Proceedings of the IEEE international conference on computer vision*, pages 2942–2950, 2017.

[23] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2176–2184, 2016.

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[25] Gang Liu, Yu Yu, Kenneth Alberto Funes Mora, and Jean-Marc Odobez. A differential approach for gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[26] LRD Murthy, Abhishek Mukhopadhyay, Varshit Yellheti, Somnath Arjun, Peter Thomas, M Dilli Babu, Kamal Preet Singh Saluja, DV JeevithaShree, and Pradipta Biswas. Evaluating accuracy of eye gaze controlled interface in military aviation environment. In *2020 IEEE Aerospace Conference*, pages 1–12. IEEE, 2020.

[27] Mattias Nilsson Benfatto, Gustaf Öqvist Seimyr, Jan Ygge, Tony Pansell, Agneta Rydberg, and Christer Jacobson. Screening for dyslexia using eye tracking during reading. *PloS one*, 11(12):e0165508, 2016.

[28] Seonwook Park, Xucong Zhang, Andreas Bulling, and Otmar Hilliges. Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pages 1–10, 2018.

[29] Gowdham Prabhakar, Aparna Ramakrishnan, Modiksha Madan, LRD Murthy, Vinay Krishna Sharma, Sachin Deshmukh, and Pradipta Biswas. Interactive gaze and finger controlled hud for cars. *Journal on Multimodal User Interfaces*, 14(1):101–121, 2020.

[30] Vinay Krishna Sharma, LRD Murthy, KamalPreet Singh Saluja, Vimal Mollyn, Gourav Sharma, and Pradipta Biswas. Webcam controlled robotic arm for persons with ssmi. *Technology and Disability*, 32(3):1–19, 2020.

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[32] Yi Tay, Anh Tuan Luu, and Siu Cheung Hui. Multi-pointer co-attention networks for recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2309–2318, 2018.

[33] Roberto Valenti, Nicu Sebe, and Theo Gevers. Combining head pose and eye location information for gaze estimation. *IEEE Transactions on Image Processing*, 21(2):802–815, 2011.

[34] Kang Wang, Rui Zhao, Hui Su, and Qiang Ji. Generalizing eye tracking with bayesian adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11907–11916, 2019.

[35] Yunyang Xiong, Hyunwoo J Kim, and Vikas Singh. Mixed effects neural networks (menets) with applications to gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7743–7752, 2019.

[36] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[37] Hirotake Yamazoe, Akira Utsumi, Tomoko Yonezawa, and Shinji Abe. Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions. In *Proceedings of the 2008 symposium on Eye tracking research & applications*, pages 245–250, 2008.

[38] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[39] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Revisiting data normalization for appearance-based gaze estimation. In *Proc. International Symposium on Eye Tracking Research and Applications (ETRA)*, pages 12:1–12:9, 2018.

[40] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Evaluation of appearance-based methods and implications for gaze-based applications. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

[41] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 51–60, 2017.

[42] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):162–175, 2017.