Supplementary Material

1. Dataset Statistics

Following are the relevant data statistics:

1. **Age:** The curated DFW dataset has variation in subject's age. The age ranges from 18-63 years. Table 1 depicts the group-wise age distribution.
2. **Gender Distribution:** The data consists of 247 male and 91 female subjects.
3. **Daytime Distribution:** 55.2% of the data is collected in daylight and 44.8% during evening with different light sources. The daytime recording sessions were conducted during both sunny and cloudy weathers. The recording was performed in different places in the university resulting in variation due to different illumination sources. Additionally, few recordings have performed at night with very little light source. As face detector failed to detect faces in these videos, these recordings are discarded from the data.
4. **Duration:** The average length of a subject's recording session is 15-20 sec. The data was recorded over one and half month time span. **5) Face Size:** The average face size in the dataset is 179×179 (in pixel). To measure the face size, Dlib face detection library [7] is used.
5. **Other Attributes:** Specular reflection add more challenge to any gaze estimation data. In DGW data, 30.17% of the subjects have prescribed spectacles. The subjects having prescribed spectacles are requested to record data with and without spectacle (subject to the condition if they can). Thus, two different settings corresponding to these subjects were recorded. Additionally, there are variations in head pose due to sitting posture of the participants.

2. Validation of Automatic Data Annotation

Head Pose Variation. We analyse the variation in head pose values w.r.t. their zones by computing the density based clustering [3] on the head pose information [4] (i.e. yaw, pitch and roll). We observed that for zones 1-3, the head-pose is mainly centrally concentrated (forming 2, 3 and 2 clusters). For the remaining zones, there are more than 5 clusters each. This experiment indicates that it may be noisy, if the gaze data is based on head-pose only. Fig. 1 shows the clustering results of zones 2, 3, 5 and 7.

Comparison With Manual Annotation Process. For comparing the automatic annotation with manual annotation, expert and non-expert annotators are assigned. There were 3 annotators (2 experts and 1 non-expert[1]) who were assigned for this task. We asked the annotators to label 15 videos. Further, we calculate few statistics to judge the quality of labelling. 2 experts take approximately 10-15 min (on average) to annotate each video. The mean squared errors of automatic label with the 2 expert annotator's labels for 15 videos are 0.49 and 0.54 respectively. Similarly, the mean squared error in case of non-expert annotator is 0.78. The cohen's kappa between the expert annotators is 0.8. At microsecond level, there is a high probability of wrong labelling during annotation by human labellers.

Table 1. Age distribution in the DGW dataset.

Age Range	18-25	26-35	36-45	Over 45
Subjects (in %)	61.8	28.7	6.7	2.8

3. Illumination Robust Layer

As per [11], both Lambertian and Phong models can be formulated by the following equations:

$$L_{diffuse} = S_d E_d \, (n.l) \tag{1}$$

$$L_{diffuse} + L_{specular} = S_d E_d \, (n.l) + S_s E_s \, (v.r)^{\gamma} \tag{2}$$

In Lambertian Equation (Eq. 1), S_d is diffusion reflection coefficient; E_d denotes the diffuse lighting intensity; n corresponds to normal vector and l denotes normal vector along the direction of incoming light [11]. Similarly, in Phong Equation (Eq. 2), S_s is the specular reflection coefficient; E_s denotes the specular lighting intensity; v is the normal vector along observation direction and r is the normal vector along the reflected light. γ is a constant termed as 'shininess constant'.

4. Data Pre-processing

After labelling the curated data, few pre-processing methods are performed to remove noise from the training data.

1) Face Detection. Dlib face detector [8] is computed with low threshold value as there are large illumination variations in the dataset. As a result of the low threshold value, we observed that in few cases, the false face detection rate also increased.

2) Optical Flow-based Face Pruning. In order to deal with the false face detections, we compute dense optical flow [1] across the detected face frames. If two consecutive frames have high Forbenius norm of the optical flow magnitude (i.e. above an empirically decided threshold), we discarded the later one. An example of face pruning is shown in Fig. 2 in which the third frame is discarded due to its high Forbenius norm value. The comparison pairs are also marked in the figure. This removes the incorrectly detected faces in the training set.

[1]Expert refers to labeler with prior labelling experience.

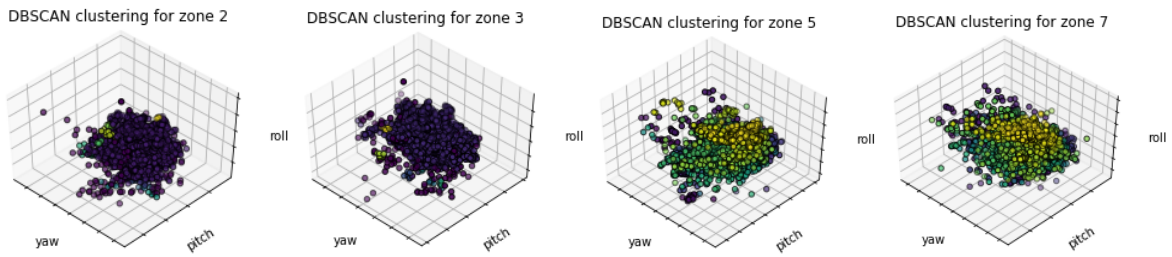

Figure 1. DBSCAN clustering of head pose along yaw, pitch and roll axis for zones 2, 3, 5 and 7 (left to right) respectively. Large number of clusters within a zone, implies that we cannot only rely on head pose information for labelling the data. (best viewed in colour)

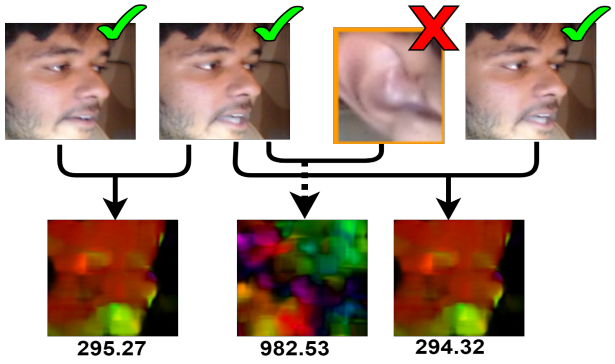

Figure 2. Overview of the optical flow based pruning method. The rejected frame is marked with a red cross.

5. Ablation Study

Eye Gaze Representation Learning. In order to analyze whether our network learnt a generalized face representation, we extracted the features from weights trained on DGW and fine tuned for the task of eye gaze estimation. We fine-tuned the network on the Columbia gaze [10] (CAVE) and TabletGaze [5] datasets. The results are shown in Table 3 and Table 2, respectively. In the case of CAVE, our model stabilizes the standard deviation. Similarly, for TabletGaze, the fine tuned network works well. These quantitative results indicate that our proposed network has learnt efficient representation from the DGW dataset.

Effect of Lip Movement. To check the effect of lips movement on the network, we performed the following experiment: images were cropped from the eye region up to nose tip point and the gaze zone prediction is re-trained. This resulted in drop of overall accuracy for the baseline network. Fig. 3 shows the overall pipeline of the aforementioned experiment. It is interesting to note that the performance difference is minimal due to the effect of loss of head pose information, when the eyes are used as input only.

Table 2. Results on Tablet Gaze with comparison to baselines [5]. Effectiveness of the learnt features from DGW dataset is demonstrated here. TG: TabletGaze, RP: Raw Pixels.

Methods TG	RP [5]	LoG [5]	LBP [5]	HoG [5]	mHoG [5]	Ours
k-NN	9.26	6.45	6.29	3.73	3.69	3.77
RF	7.2	4.76	4.99	3.29	3.17	
GPR	7.38	6.04	5.83	4.07	4.11	
SVR	-	-	-	-	4.07	

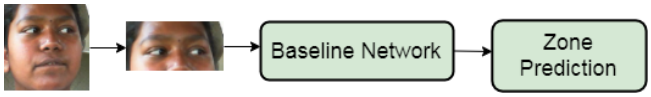

Figure 3. Lip movement effect analysis pipeline (Section **??**).

Jetson Nano Experiments. It is important to note that a more complex backbone network may achieve better performance. We chose Inception-V1 due to relatively better performance and to be able to evaluate the performance on a small platform such as the Nvidia Jetson Nano. These networks will be expected to run on close to real-time in a car based computer. On Nvidia Jetson Nano the best network runs at 10 Frames Per Second (FPS) with 34,532,961 parameters.

References

[1] L Alvarez, J Weickert, and J Sánchez. Reliable estimation of dense optical flow fields with large displacements. *International Journal of Computer Vision*, pages 41–56, 2000. 1

[2] N Dubey, S Ghosh, and A Dhall. Unsupervised learning of eye gaze representation from the web. In *International Joint Conference on Neural Network*, pages 1–7, 2019. 3

[3] M Ester, H Kriegel, J Sander, and X Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Knowledge Discovery and Data Mining*, volume 96, pages 226–231, 1996. 1

[4] Yin Guobing. Headpose. Technical report, 2016. 1

[5] Q Huang, A Veeraraghavan, and A Sabharwal. Tabletgaze: unconstrained appearance-based gaze estimation in mobile tablets. *Machine Vision and Applications*, pages 445–461, 2015. 2

[6] S Jyoti and A Dhall. Automatic eye gaze estimation using geometric & texture-based networks. In *IEEE International Conference on Pattern Recognition*, pages 2474–2479, 2018. 3

Table 3. Results on the CAVE dataset (of °0 yaw angle) using angular deviation, calculated as $mean\ error \pm std.\ deviation$ (in°).

Dataset	[6]		[9]		[2]		Ours	
	x	y	x	y	x	y	x	y
CAVE	1.67 ± 1.19	3.47 ± 3.99	2.65 ± 3.96	4.02 ± 5.82	1.67 ± 1.19	1.74 ± 1.57	2.17 ± 0.91	1.31 ± 0.73

[7] D E King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, pages 1755–1758, 2009. 1

[8] S Sharma, K Shanmugasundaram, and S K Ramasamy. Farec—cnn based efficient face recognition technique using dlib. In *IEEE International Conference on Advanced Communication Control and Computing Technologies*, 2016. 1

[9] E Skodras, V G Kanas, and N Fakotakis. On visual gaze tracking based on a single low cost camera. *Signal Processing: Image Communication*, 2015. 3

[10] B A Smith, Q Yin, S K Feiner, and S K Nayar. Gaze locking: passive eye contact detection for human-object interaction. In *ACM symposium on User interface software and technology*, pages 271–280, 2013. 2

[11] W Zhang, X Zhao, J Morvan, and L Chen. Improving shadow suppression for illumination robust face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):611–624, 2019. 1