FLAME: Facial Landmark Heatmap Activated Multimodal Gaze Estimation

Neelabh Sinha, Michal Balazia, Francois F Bremond

FLAME: Facial Landmark Heatmap Activated Multimodal Gaze Estimation

Neelabh Sinha[1,2], Michal Balazia[1,3], and François Bremond[1,3]

[1]INRIA Sophia Antipolis - Méditerranée, France
[2]Birla Institute of Technology and Science, Pilani, India
[3]Université Côte d'Azur, France
https://github.com/neelabhsinha/flame

Abstract

3D gaze estimation is about predicting the line of sight of a person in 3D space. Person-independent models for the same lack precision due to anatomical differences of subjects, whereas person-specific calibrated techniques add strict constraints on scalability. To overcome these issues, we propose a novel technique, Facial Landmark Heatmap Activated Multimodal Gaze Estimation (FLAME), as a way of combining eye anatomical information using eye landmark heatmaps to obtain precise gaze estimation without any person-specific calibration. Our evaluation demonstrates a competitive performance of about 10% improvement on benchmark datasets ColumbiaGaze and EYEDIAP. We also conduct an ablation study to validate our method.

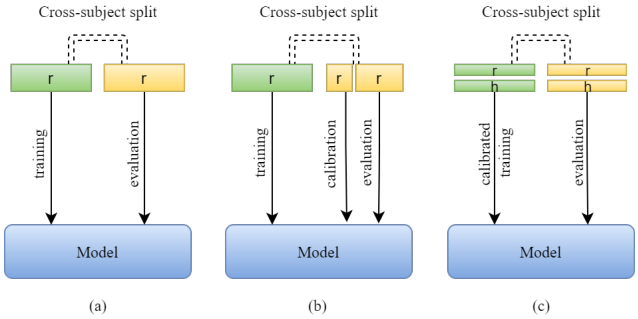


Figure 1: Different types of gaze estimation methods: (a) person-independent technique, (b) person-specific technique, (c) **FLAME**. Training subjects are green and test subjects are yellow. r stands for RGB image and h stands for eye landmark heatmap.

1. Introduction

Eye gaze is an important non-verbal cue of humans and is capable of describing human emotion and behaviour. It is used in large cross-domain applications such as study of human behavior [24], clinical diagnosis [27], human-computer and human-robot interaction [11, 19]. To have precision in such applications, it is very important to have an accurate estimation of eye gaze. 3D gaze estimation is about inferring the actual line-of-sight of a person in 3D space and is the main focus of this paper.

Owing to the importance of study of this problem, vast research has been conducted. With the evolution in the field of deep learning, this problem has also moved significantly towards using it to predict gaze [10, 30, 33, 35]. Many techniques focus on person-independent gaze estimation [33, 35] (Figure 1(a)) where the model is evaluated on un-encountered subjects. But gradually, it was proved that it is not possible to reduce error below a certain point due to heavy dependency of gaze on the anatomy of the eye of different subjects [12]. Few classical model-based techniques [12, 13] were able to incorporate these anatomical features with few calibration samples, but they are not robust to variation in the image under different settings. As a result, subject-specific personalized models [30, 22] (Figure 1(b)) got introduced which were able to further reduce errors using additional calibration steps on the test subjects to incorporate anatomical features.

However, although these person-specific models tend to give a low value of error, using them to design practical systems is difficult. This is because, in real-world scenarios, systems keep on encountering new subjects. So, for additional calibration, there have to be steps to take test samples with accurate ground truth and feed it to the model to calibrate before the actual task can be carried out. In different settings with varying surroundings and distance between subject and camera, it is challenging to procure test samples with accurate ground truth to calibrate the model. Further, for applications like clinical diagnosis of psychological dis-

orders, for example autism and schizophrenia, where gaze is a very important non-verbal cue, the affected people cannot follow precise instructions to be performed for acquiring calibration samples. This makes it practically impossible to rely on person-specific methods for many applications.

In order to address this issue, and also use the anatomical features while predicting the gaze, we propose the Facial Landmark Heatmap Activated Multimodal Gaze Estimation (FLAME), which focuses on precise gaze estimation without calibration by incorporating the anatomical information contained in eye landmarks along with RGB features (Figure 1(c)). Landmarks corresponding to the outline of the eye hold key information about its anatomical features, and as also shown by [29], depict strong correlation with gaze direction. However, since the extraction of eye landmarks using other algorithms are also prone to errors, to make the model robust to these, we take a heatmap based approach inspired by [9] for skeleton-based action recognition. Instead of using absolute coordinates of landmark points, we use a Gaussian probability distribution heatmap for each landmark. This allows us not to pose strong constraints on the obtained value of the landmark coordinates and rather to provide it in the form of a probability distribution, while also allowing the system to extract salient anatomical features from it.

The high-level idea is depicted in Figure 2, which follows a two-stream CNN based neural network, working on RGB and eye landmark heatmap modalities to predict gaze angles. These angles can easily be converted to directional vector using basic trigonometry. To the best of our knowledge, this is the first work in gaze estimation using the information from eye landmarks in the form of a heatmap along with the RGB image.

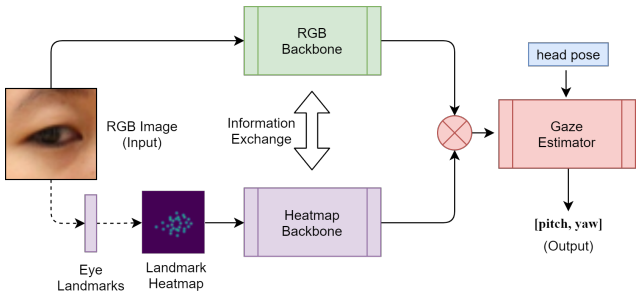


Figure 2: Method Summary. Landmark heatmap is first obtained from cropped eye images. Then, both are processed through a two-stream CNN-based network to extract gaze and anatomical information separately, which is further used along with head pose to predict gaze. Dotted line marks the pre-training extraction and solid line marks the training pipeline.

Key contributions of this work are as follows:

- A novel approach for **precise gaze estimation without calibration by incorporating anatomical features** of the eye seamlessly with only RGB image as input.
- A method for obtaining, and **using eye landmarks as heatmaps** to extract anatomical features and merge with RGB feature map.
- Extensive experimentation on public benchmarks to obtain **accurate gaze estimations in person-independent settings**.

In the rest of the paper, we discuss our proposed approach in detail. Section 2 describes the past work in this field, followed by complete details of the proposed methodology in Section 3. In Section 4 we discuss the implementation details, which is succeeded by results and ablations in Section 5. Section 6 concludes the work and discusses its future scope.

2. Related Work

3D gaze estimation can be broadly divided into 2 parts: geometric methods [13] and appearance-based methods [26]. Geometric methods [12, 13] focus on modelling the eye, extracting features in terms of geometrical parameters and use it to predict gaze. They have several constraints like near-frontal head-pose and high-resolution images, and thus have limited application as compared to appearance-based methods, which can directly estimate gaze from eye images, and thus have gained significant popularity in present days.

Among person-independent methods, [33] proposed a LeNet-based architecture using a single eye patch, [10] used parallel VGG-16 based architecture and fed eyes and face to estimate head pose and gaze angle, [4] used Dilated Convolutions, while [35] used a more complex CNN backbone to predict gaze. [34] did full-face gaze estimation by generating an attention map, and [18] implemented three parallel CNN backbones to process left eye, right eye and full face to predict gaze. [23] depicted a more robust use of head-pose for gaze estimation. Further, [29] proposed a multi-task CNN to predict eye-gaze and facial landmarks, and concluded that eye-landmarks are strongly co-related to gaze. As a developing trend, to remove person-specific bias of the models as shown by [12], person-specific models [30, 22] have been proposed for gaze estimation which calibrate the model on few samples of test subjects before evaluation. Unsupervised learning has also come into the picture with [31] proposing unsupervised gaze estimation model using gaze redirection mechanism. Recently, [6] proposed gaze estimation using transformers and [7] implemented gaze estimation by attention mechanism and also using difference layer to remove unwanted features from both eyes.

Multimodal fusion has been of significant interest across domains. With time, late fusion has gained significant popularity where each mode is processed separately and combined later. It can be achieved by element-wise summation, concatenation, bi-linear product [3], weighted average [21], rank minimization [28]. [15] involves using attention to pick the best mode for each input. Multimodal fusion module [16] was proposed which allows slow modality fusion in features of different dimensions.

Inspired by these works, we developed our own technique which we describe in-depth in the following sections.

3. Method Overview

The first part of our method is to extract the eye landmarks from the RGB image and generate the heatmap, followed by a two stream network having multiple components. From here on, for all discussions, r refers to RGB image and h represents eye landmark heatmap.

3.1. Eye Landmark Heatmap

To get the heatmap, we first extract the 2D eye landmarks from corresponding RGB images using facial landmark submodule [1, 32] of OpenFace 2.0 [2]. We prefer this because it can provide landmarks for the outline of the cornea and pupil along with outer eye, and is significantly precise and easy to use. OpenFace gives 28 2D landmarks for each eye as shown in Figure 3 in pixel coordinates, capturing the outline of the complete eye, cornea and pupil. These three components not only allow us to capture the shape of the outer eye but also the boundaries of inner regions, giving a large variety of descriptive information about the anatomy of the eye.

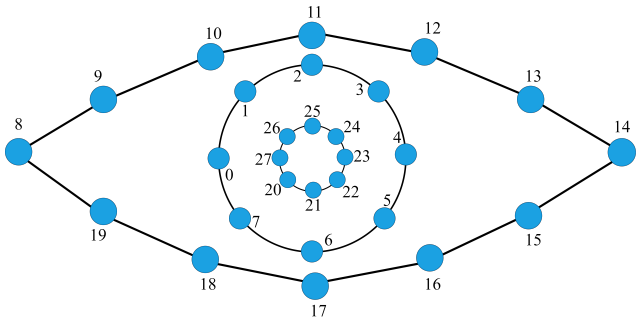


Figure 3: Eye landmarks from Openface 2.0

For a d-dimensional input, Gaussian probability distribution heatmap is defined by

$$P(X) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}\left((X-\mu)^\intercal \Sigma^{-1} (X-\mu)\right)\right) \tag{1}$$

with mean μ and covariance matrix Σ. We use it to obtain the Gaussian probability distribution heatmap of the eye landmark coordinates having the same dimensions as the input image.

3.2. Transfer Function

The transfer function $T(r, h)$ is the function that defines the exchange of information between the two streams. It is based on Multimodal Transfer Module (MMTM) [16], which is a slow modality fusion block used to re-calibrate channel-wise features based on squeeze and excitation mechanism between any two feature maps of arbitrary dimension. The complete set of steps can be visualized from Figure 4.

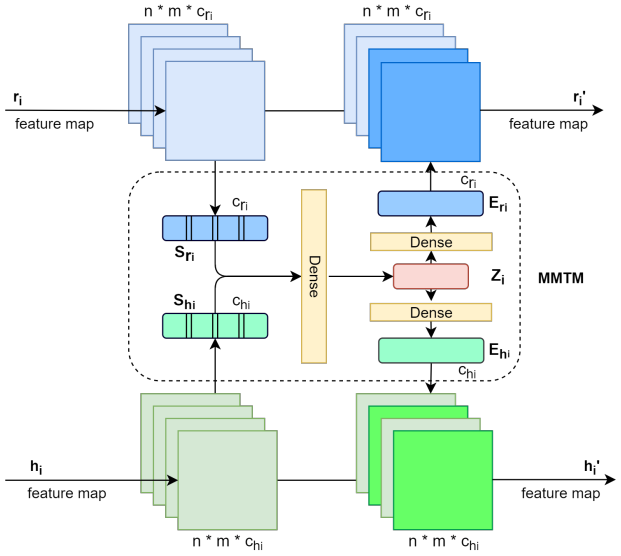


Figure 4: Multimodal Transfer Module [16]

Assume an input feature vector of height n and width m. Let $\mathbf{r_i} \in \mathbb{R}^{n \times m \times c_{r_i}}$ be the i-th RGB feature map of c_{r_i} channels and $\mathbf{h_i} \in \mathbb{R}^{n \times m \times c_{h_i}}$ the i-th heatmap feature vector of c_{h_i} channels. We first squeeze them into $\mathbf{S_{r_i}} \in \mathbb{R}^{c_{r_i}}$ and $\mathbf{S_{h_i}} \in \mathbb{R}^{c_{h_i}}$, the channel-wise descriptors, using global average pooling across the spatial dimensions:

$$S_{r_i}(c) = \frac{1}{n \cdot m} \sum_{p=1}^{n} \sum_{q=1}^{m} r_i(p, q, c) \tag{2}$$

$$S_{h_i}(c) = \frac{1}{n \cdot m} \sum_{p=1}^{n} \sum_{q=1}^{m} h_i(p, q, c) \tag{3}$$

Then, we concatenate these two, and pass it through a dense layer to obtain a joint representation $\mathbf{Z_i}$. This $\mathbf{Z_i}$ is passed through separate dense layers to obtain excitation signals $\mathbf{E_{r_i}} \in \mathbb{R}^{c_{r_i}}$ and $\mathbf{E_{h_i}} \in \mathbb{R}^{c_{h_i}}$ for both RGB and heatmap modalities.

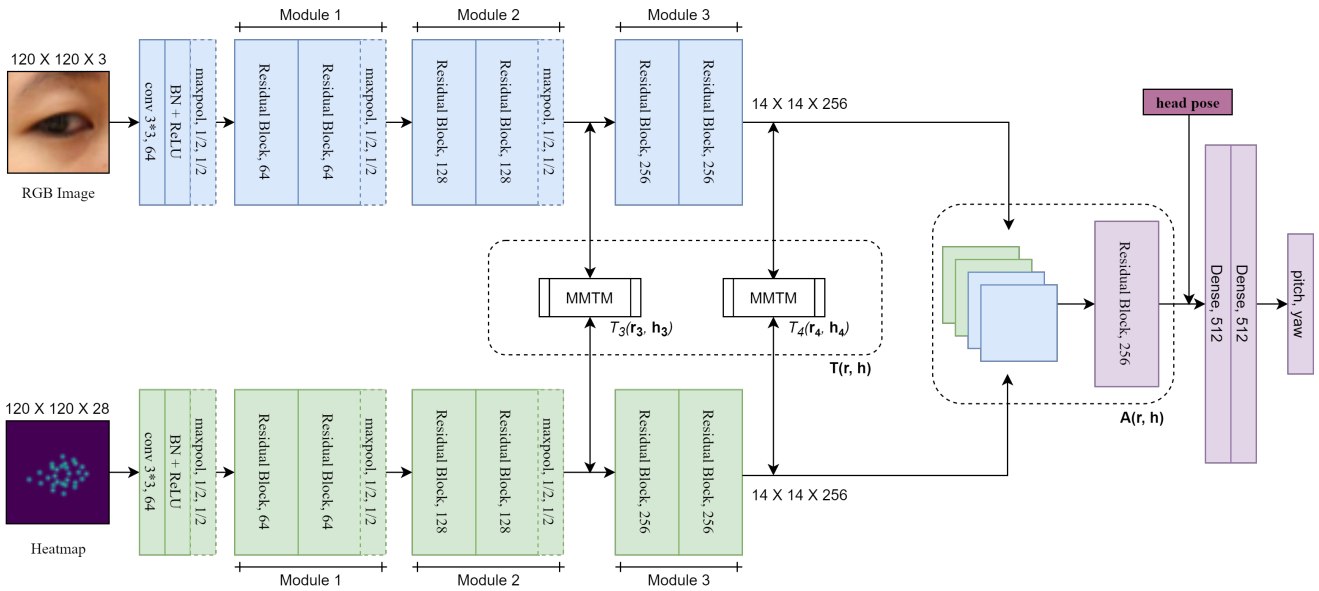


Figure 5: Complete model architecture. Two identical CNN backbones process the RGB and heatmap modality separately, with exchange of information at intermediate layers as per transfer function and later fusing them using aggregation function. This hybrid feature map along with head pose is passed to a fully-connected regression network to predict gaze angles.

The final re-calibrated feature map can be obtained by taking sigmoid σ of the excitation signal and performing channel-wise multiplication with the actual feature map:

$$\mathbf{r}'_{\mathbf{i}}(c) = 2 \cdot \sigma\left(E_{r_i}(c)\right) \cdot \mathbf{r}_{\mathbf{i}}(c) \tag{4}$$

$$\mathbf{h}'_{\mathbf{i}}(c) = 2 \cdot \sigma\left(E_{h_i}(c)\right) \cdot \mathbf{h}_{\mathbf{i}}(c) \tag{5}$$

We take $dim(\mathbf{Z_i}) = \frac{1}{4}\left(dim\left(\mathbf{S_{r_i}}\right) + dim\left(\mathbf{S_{h_i}}\right)\right)$, as suggested in [16]. In our case, $dim\left(\mathbf{S_{r_i}}\right) = dim\left(\mathbf{S_{h_i}}\right)$.

To design the complete transfer function $T(r, h)$, we use this MMTM block for feature reactivation at the last and second last feature maps, as given in Figure 5. This is because we want the network to learn some initial representation and then, the higher-level features of both streams can be utilized by each other. At the deeper layers, using this transfer function can help both individual unimodal streams to learn better representation by benefiting from the information extracted by the other stream.

3.3. CNN Backbone

We believe that gaze is a relatively low-level feature and thus, a shallow architecture is effective to learn relevant features. So, we designed a custom CNN backbone based on Residual Blocks [14] as given in Figure 5.

It starts with a 3×3 convolution layer followed by Batch Normalization, ReLU and maxpool with kernel size $(2, 2)$ and stride $(2, 2)$. This is followed by 3 modules, each having 2 residual blocks followed by a maxpool layer. The intermediate feature output at the end of each module has 64, 128 and 256 channels respectively. Both RGB and heatmap modalities are processed using identical backbones to maintain symmetry, as it creates identical feature dimensions to ease and balance other operations in the network.

3.4. Aggregation Function

The aggregation function $A(r, h)$ is used to fuse the final feature map of both streams in order to get a hybrid representation. This is done by channel-wise concatenation of the individual feature maps, and then passing this through a residual block. The output of this is a hybrid feature map having relevant information from both RGB and heatmap modalities.

3.5. FLAME

The complete architecture of FLAME, as depicted in Figure 5, is designed from the previously described individual units. The two separate CNN backbones take 120×120 RGB eye patch and $120\times120\times28$ eye landmark heatmap. Both learn features independent of each other until the last two modules where both the features are fused into one another by the transfer function $T(r, h)$, which consists of two MMTM blocks. Thereafter, using the two final feature maps, a hybrid feature map is obtained by the aggregation function $A(r, h)$. This feature map contains the information of the gaze direction as well as the anatomical information of the eye and is a better descriptor of the gaze direction.

It is then flattened and concatenated with the head pose angles and is passed through two dense layers of 512 neu-

rons each, followed by a two-dimensional output layer for the pitch and yaw gaze angles directly in the World Coordinate System (WCS). In the dense layers, dropout is used with $p_{dropout} = 0.2$ and ReLU is used as activation function. The output layer, however, is kept as linear with no activation function.

3.6. Loss Function

This model is end-to-end trainable using the gaze angle ground truth. The 3D angular loss can be defined as

$$\mathcal{L}_{ang}\left(\hat{\boldsymbol{g_p}}, \hat{\boldsymbol{g_t}}\right) = \arccos \frac{\hat{\boldsymbol{g_p}} . \hat{\boldsymbol{g_t}}}{|\hat{\boldsymbol{g_p}}||\hat{\boldsymbol{g_t}}|}, \tag{6}$$

where $\hat{\boldsymbol{g_p}}$ and $\hat{\boldsymbol{g_t}}$ represent the predicted and true gaze vectors respectively. This loss, given by Equation 6, directly gives the angular deviation of predicted gaze vector $\hat{\boldsymbol{g_p}}$ from the true gaze vector $\hat{\boldsymbol{g_t}}$ in 3D space.

However, training directly on this loss function, as used by a few models [30], is not ideal because the gradient of this loss, given by

$$\frac{\partial \mathcal{L}_{ang}\left(x\right)}{\partial x} = \frac{1}{\sqrt{1-x^2}} \quad \text{where} \quad x = \frac{\hat{\boldsymbol{g_p}} . \hat{\boldsymbol{g_t}}}{|\hat{\boldsymbol{g_p}}||\hat{\boldsymbol{g_t}}|}, \tag{7}$$

increases sharply with decrease in loss and approaches ∞ as the loss approaches 0 (x approaching 1). So, training on this loss makes the network unstable with decrease in loss.

Thus, we define a vector difference loss for training, which is the sum of squared errors of each individual component of the gaze vector $g_k = (g_k^x, g_k^y, g_k^z)$ where $k = \{p, t\}$, given by

$$\mathcal{L}_{vec}\left(\hat{\boldsymbol{g_p}}, \hat{\boldsymbol{g_t}}\right) = \left(g_p^x - g_t^x\right)^2 + \left(g_p^y - g_t^y\right)^2 + \left(g_p^z - g_t^z\right)^2. \tag{8}$$

For evaluation, we use the 3D angular loss itself as per Equation 6.

4. Implementation Details

4.1. Datasets

We perform the experiments with two benchmark datasets: ColumbiaGaze [25], EYEDIAP [20]. EYEDIAP is a video dataset with 16 subjects, with different head motion and gaze target settings. We select every 20-th frame of the HD screen target videos (continuous and discrete) under both static and moving head pose of all subjects for our experiments. ColumbiaGaze is a dataset with discrete values of head pose and targets, having 56 subjects, each with a combination of 7 horizontal and 3 vertical gaze targets. We select all the images from this dataset. For all these images, the ones on which OpenFace outputs could not be obtained were left out of the final dataset. For our experiments, we use the raw images without any re-alignment.

4.2. Preprocessing

We first extract face crops of the ColumbiaGaze and EYEDIAP images using RetinaFace [8], and zero-pad it to a uniform size of 384×480 (4:5 ratio). As ColumbiaGaze dataset was of higher resolution, we rescale the padded 4:5 crops to this dimension. Then, we run OpenFace 2.0 to obtain eye landmarks, and generate the heatmap for all the data with dimension $d = 2$, $\Sigma = \mathbb{I}_2$ (2D identity matrix) and $\mu = (x_i, y_i), i = 1, 2, \ldots, 28$ where (x_i, y_i) are the obtained eye landmark coordinates, following the method discussed in Section 3.1. We then divide the complete dataset into cross-subject split of 8:1:1 for train, validation and test sets, respectively.

4.3. Training and Optimization

The network is trained with randomly initialized weights for 200 epochs with a batch size of 8, using Adam Optimizer with an initial learning rate of 10^{-4}, and is decreased by a factor of 0.5 after 85-th, 120-th and 175-th epoch. During training, 120×120 pixels eye patch of either left or the right eye is picked randomly for each image sample along with the corresponding landmark heatmap of that eye having same resolution, and fed to the network. The eye crop is obtained by calculating the centre of the eye using the four corner landmarks (8, 14, 11, 17 in Figure 3), and taking a 120×120 patch across it with the obtained centre. The same operation is done with landmark heatmap to maintain alignment between both inputs. Head pose is used as available with the datasets, but can also be obtained from facial analysis toolkits like OpenFace 2.0 with sufficient precision. The model is then trained against ground truth gaze angles using the vector difference loss given by Equation 8.

4.4. Experimental Settings

In order to conduct an ablation study of the proposed method, we implemented three other models in addition to our main approach of FLAME. All four experimental settings are described below:

FLAME: Our original approach as proposed in Section 3.5 with settings as described in Section 4.3 on all the datasets.

FLAME-aggregation-only (F-AO): The original architecture without the transfer function $T(r, h)$. This is to evaluate the impact of using multimodal transfer.

FLAME-additive-fusion (F-AF): Another approach to fuse both modalities, where aggregation function is the element-wise addition of the final feature map of both modalities. This is to analyse the effectiveness of aggregation function $A(r, h)$. There is no $T(r, h)$ here.

FLAME-baseline (F-B): This is the baseline approach that performs gaze estimation using only the RGB images and proposed CNN backbone. It is to see the contribution of the entire heatmap modality and proposed novelty.

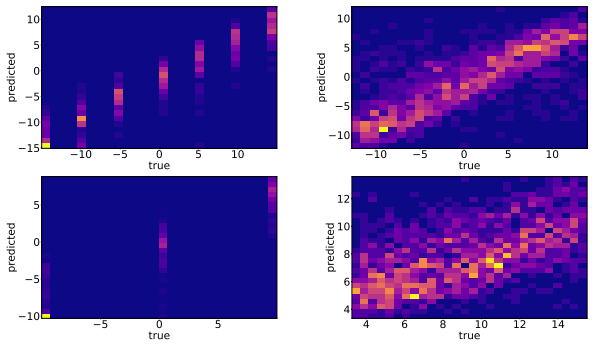
5. Results

The performance of our method can be evaluated by Equation 6, the angular error between true and predicted gaze vectors in 3D space. Results are given in Table 1 for both cross-subject and cross-dataset validation to see how well the model performs on unencountered subjects (cross-subject evaluation), and also on completely unknown setup (distance of subjects, camera settings, etc.) and range of ground truth (cross-dataset evaluation).

Table 1: Results on cross-subject, cross-dataset evaluation.

Train \ Test	ColumbiaGaze	EYEDIAP
ColumbiaGaze	4.64°	12.53°
EYEDIAP	12.83°	4.62°

Figure 6 represents the relation between prediction and truth values for both datasets. They pretty much follow a linear relationship which further proves the correctness of our methodology.



(a) ColumbiaGaze (b) EYEDIAP

Figure 6: Analysis of predictions versus true labels. Top heatmaps are evaluating the yaw angle and bottom heatmaps are evaluating the pitch angles.

In Table 2 we can see that when the model was not provided with any previously encountered subjects, it was able to extract the information from eye landmarks and reduce the mean error by 1.29° and 0.7° as compared to RGB only representing the baseline (F-B). However, to still analyse the factors contributing to errors, it can be understood that we are using 2D facial landmarks, and not 3D. This might fail to capture the precise orientation and shape of the eye in 3D space, and may not describe accurate anatomical structure. Additionally, there can be significant error in eye landmark extraction, particularly when the eye is not fully visible, or when the person is looking directly into the camera. These can be larger for the cornea and pupil landmarks which have finer boundaries.

Cross-dataset error is predominantly because the two datasets are very different to one another in many parameters such as range of gaze angles, distance of subject, or camera parameters. However, although these values are not directly comparable to other methods due to difference in datasets used for cross-dataset evaluation, the performance is competitive to them.

5.1. Ablation Study

It is also important to understand the contribution of all the components proposed in our methodology. For this, we implemented three other architectures as described in Section 4.4. The results obtained are given in Table 2.

Table 2: Mean and standard deviation (std) of angular error for different experimental settings

Experiment Setup	ColumbiaGaze		EYEDIAP	
	mean	std	mean	std
F-B	5.93°	3.20°	5.32°	3.08°
F-AF	5.88°	3.06°	5.30°	3.03°
F-AO	5.06°	3.13°	4.80°	3.02°
FLAME	**4.64°**	**2.86°**	**4.62°**	**2.93°**

We can see that FLAME outperforms all other experimental settings. Figure 7 depicts the distribution of error for EYEDIAP and ColumbiaGaze datasets on all four experimental settings through a box plot.

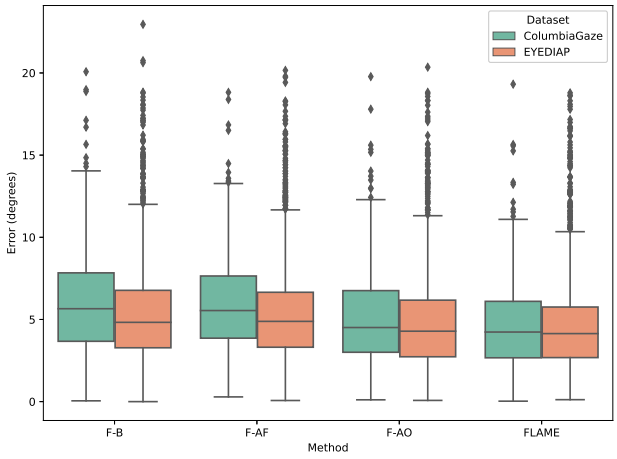


Figure 7: Distribution of errors of all experimental settings

There is a difference in mean error of 0.18° for EYEDIAP and 0.42° for ColumbiaGaze between F-AO and FLAME, where the absence of transfer function was the only difference. This shows its contribution.

Further, F-AO and F-AF differ in aggregation function $A(r, h)$. We compare our approach with element-wise addition method in F-AF, where $A(r, h) = \mathbf{r_4} + \mathbf{h_4}$. The error difference highlights the effectiveness of our aggregation function for creating a hybrid feature map.

At last, the contribution of our complete methodology can be seen through the difference of FLAME with F-B.

We also compared the performance of FLAME with different input resolutions. In practical settings, having high-resolution eye images is a constraint due to limited camera resolution, distance of subject from it, etc. On ColumbiaGaze, we obtained error of $4.79°$ in 60×60 pixels and $5.5°$ on 30×30 pixels image resolution. Based on this, we would propose 120×120 pixels as optimum and 60×60 pixels as the minimum threshold of input resolution with FLAME.

5.2. Comparison with State-of-the-art

We also compare our method with various state-of-the-art. The results are given in Table 3. All of them are for leave-one-subject-out settings (or cross-subject evaluation).

Table 3: Mean 3D angular error as reported by different SOTA methods: person-specific methods (top), person-independent methods (bottom)

Method	ColumbiaGaze	EYEDIAP
Yu *et al.* (sup. HCS) [31]	5.25°	7.09°
Yu *et al.* (sup. WCS) [31]	3.54°	6.79°
Yu *et al.* (unsup.) [31]	7.15°	8.2°
Zhang *et al.* [34]	-	6.0°
Zhang *et al.* [33]#	-	7.37°
Zhang *et al.* [35]#	-	6.79°
Cheng *et al.* [5]	-	5.3°
Cheng *et al.* [6]	-	5.17°
Fischer *et al.* [10]*	-	6.4°
Chen *et al.* [4]*	-	5.9°
Kellnhofer *et al.* [17]#	-	5.36°
Yu *et al.* [29]	-	8.54°
FLAME (ours)	**4.64°**	**4.62°**

*error metric as reported by [5], #error metric as reported by [6]

It can be seen that our method performs significantly better than all of the person-independent methods for gaze estimation. This shows the impact of using eye landmark heatmaps to utilize anatomical information of the eye along with the RGB image. When comparing to the models of [31], a person-specific method (based on few-shot calibration), our approach performs better than all 3 methods on both datasets, except for one model in ColumbiaGaze. A larger error than person-specific models can be because when the model gets the test samples directly, it can extract additional information like orientation in 3D space, PERCLOS (% of eye closed), and other salient anatomical features, which are not possible to capture using 2D eye landmarks. For ColumbiaGaze, SOTA benchmarks are not available for many models.

Thus, with all these analyses and experimentation, we can justify that our approach holds key to a precise gaze estimation.

6. Conclusion

We presented an approach of using 2D eye landmarks along with RGB images to perform 3D gaze estimation by incorporating anatomical features of the eye without any person-specific calibration. In our approach, we obtained a 2D Gaussian probability distribution heatmap of eye landmarks from the RGB image itself and then used a two-stream CNN-based network to effectively extract relevant features. We used this along with head pose to predict gaze angles. Our method gives better results than all state-of-the-art methods on two benchmark datasets. This proves that eye landmarks can play a vital role in incorporating anatomical information to predict gaze more accurately.

We believe more precise systems can be designed if we can incorporate more information related to the anatomical aspects of the eye, and use eye landmarks in other ways. As of now, we leave that for further study.

References

[1] T. Baltrusaitis, P. Robinson, and L.-P. Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *ICCVW*, 12 2013. 3

[2] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66, 2018. 3

[3] H. Ben-younes, R. Cadène, M. Cord, and N. Thome. Mutan: Multimodal tucker fusion for visual question answering. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2631–2639, 2017. 3

[4] Z. Chen and B. E. Shi. Appearance-based gaze estimation using dilated-convolutions. In *Asian Conference on Computer Vision*, pages 309–324. Springer, 2018. 2, 7

[5] Y. Cheng, S. Huang, F. Wang, C. Qian, and F. Lu. A coarse-to-fine adaptive network for appearance-based gaze estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):10623–10630, Apr. 2020. 7

[6] Y. Cheng and F. Lu. Gaze estimation using transformer. *CoRR*, abs/2105.14424, 2021. 2, 7

[7] M. L. R. D and P. Biswas. Appearance-based gaze estimation using attention and difference mechanism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3143–3152, June 2021. 2

[8] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, 2020. 5

[9] H. Duan, Y. Zhao, K. Chen, D. Shao, D. Lin, and B. Dai. Revisiting skeleton-based action recognition. *CoRR*, abs/2104.13586, 2021. 2

[10] T. Fischer, H. J. Chang, and Y. Demiris. RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments. In *European Conference on Computer Vision*, pages 339–357, September 2018. 1, 2, 7

[11] J. H. Goldberg, M. J. Stimson, M. Lewenstein, N. Scott, and A. M. Wichansky. Eye tracking in web search tasks: Design implications. In *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications*, ETRA '02, page 51–58, New York, NY, USA, 2002. Association for Computing Machinery. 1

[12] E. D. Guestrin and M. Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on Biomedical Engineering*, 53:1124–1133, 2006. 1, 2

[13] D. Hansen and Q. 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, mar 2010. 1, 2

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[15] R. Jacobs, M. Jordan, S. Nowlan, and G. Hinton. Adaptive mixture of local expert. *Neural Computation*, 3:78–88, 02 1991. 3

[16] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13289–13299, 2020. 3, 4

[17] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, , and A. Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *IEEE International Conference on Computer Vision (ICCV)*, October 2019. 7

[18] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba. Eye tracking for everyone. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[19] M. Michalowski, S. Sabanovic, and R. Simmons. A spatial model of engagement for a social robot. *9th IEEE International Workshop on Advanced Motion Control, 2006.*, pages 762–767, 2006. 1

[20] K. Mora, F. Monay, and J. Odobez. Eyediap: a database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. *Proceedings of the Symposium on Eye Tracking Research and Applications*, 2014. 5

[21] P. Natarajan, S. Wu, S. Vitaladevuni, X. Zhuang, S. Tsakalidis, U. Park, R. Prasad, and P. Natarajan. Multimodal feature fusion for robust event detection in web videos. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1298–1305, 2012. 3

[22] S. Park, S. D. Mello, P. Molchanov, U. Iqbal, O. Hilliges, and J. Kautz. Few-shot adaptive gaze estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9368–9377, 2019. 1, 2

[23] R. Ranjan, S. D. Mello, and J. Kautz. Light-weight head pose invariant gaze tracking. *CoRR*, abs/1804.08572, 2018. 2

[24] G. E. Raptis, C. Katsini, M. Belk, C. Fidas, G. Samaras, and N. Avouris. Using eye gaze data and visual activities to infer human cognitive styles. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*. ACM, July 2017. 1

[25] B. Smith, Q. Yin, S. Feiner, and S. Nayar. Gaze Locking: Passive Eye Contact Detection for Human?Object Interaction. In *ACM Symposium on User Interface Software and Technology (UIST)*, pages 271–280, Oct 2013. 5

[26] K.-H. Tan, D. J. Kriegman, and N. Ahuja. Appearance-based eye gaze estimation. In *Proceedings of the Sixth IEEE Workshop on Applications of Computer Vision*, WACV '02, page 191, USA, 2002. IEEE Computer Society. 2

[27] M. Vidal, J. Turner, A. Bulling, and H. Gellersen. Wearable eye tracking for mental health monitoring. *Computer Communications*, 35(11):1306–1311, June 2012. 1

[28] G. Ye, D. Liu, I.-H. Jhuo, and S.-F. Chang. Robust late fusion with rank minimization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3021–3028, 2012. 3

[29] Y. Yu, G. Liu, and J.-M. Odobez. Deep multitask gaze estimation with a constrained landmark-gaze model. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018. 2, 7

[30] Y. Yu, G. Liu, and J.-M. Odobez. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11937–11946, 2019. 1, 2, 5

[31] Y. Yu and J.-M. Odobez. Unsupervised representation learning for gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7314–7324, 2020. 2, 7

[32] A. Zadeh, T. Baltrusaitis, and L. Morency. Deep constrained local models for facial landmark detection. *CoRR*, abs/1611.08657, 2016. 3

[33] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Appearance-based gaze estimation in the wild. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4511–4520, June 2015. 1, 2, 7

[34] X. Zhang, Y. Sugano, M. Fritz, and A. 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. 2, 7

[35] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):162–175, 2017. 1, 2, 7