Exploring the Influence of Facial Features Beyond the Eyes on Gaze Estimation

Tan M. Nguyen[1]

Supervisor(s): Dr. G. (Guohao) Lan[1]**, Lingyu Du**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Tan M. Nguyen
Final project course: CSE3000 Research Project
Thesis committee: Dr. G. (Guohao) Lan, Dr. Xucong Zhang, Lingyu Du.

Abstract

Gaze estimation holds significant importance in various applications. Pioneering research has demonstrated state-of-the-art performance in gaze estimation models by utilizing deep Convolutional Neural Networks (CNNs) and incorporating full facial images as input, instead of or in addition to solely using one or both eye images. Facial images encode crucial cues that can enhance the accuracy of gaze regression models. However, it remains unclear which specific facial features contribute and to what extent they contribute to the overall estimation accuracy. In this research, we aim to shed light on identifying the influential facial regions and quantifying their contributions to gaze estimation accuracy.

1 Introduction

Background. Eye gaze direction holds importance in many applications, including human-computer interaction [1], driver monitoring systems [2], and behavioral analysis [3] among others. Accurate estimation of gaze direction has been shown to significantly enhance human-computer interaction efficiency [4], for example in head-mounted display (HMD) systems [5]. Gaze tracking is a crucial component in driver monitoring safety, as it enables the detection of a driver's focused attention to prevent accidents [6], [7], [8]. The works by Hessels et al. [3] and Holmqvist et al. [9] have further evidenced the importance of eye-tracking models in the studying of human perceptual, cognitive processes, and behavioral analysis. These findings highlight the crucial role that gaze estimation plays in our world and emphasize its importance in various research areas.

In recent years, the inclusion of full facial images as input to deep Convolutional Neural Networks (CNNs) has emerged as a crucial factor in improving the accuracy of eye-gaze regression models, significantly highlighting the importance of capturing the entire facial context [10], [11], [12]. Pioneering research by Zhang et el., Krafka et al., and Zemblys et al. have made a significant contribution by incorporating full facial images as input to convolutional neural networks (CNNs) instead of, or in addition to, solely focusing on one or both eye regions. Their studies have shown that while eye regions are crucial for gaze estimation, the inclusion of full facial images is also essential as they encompass valuable cues that significantly leverage the performance of gaze estimation models.

While previous studies have highlighted the importance of including full facial features in gaze regressor models, as they provide valuable cues for accurately predicting gaze direction. However, despite these advancements, a fundamental question remains unanswered: Which specific facial features contribute significantly to improving the accuracy of gaze estimation models?

Research questions and main contributions. The objective of this study aims to answer the aforementioned question by conducting an investigation into the significant regions of the face and quantifying their contribution toward their impact on the accuracy of gaze estimation. The contributions of this research are two folds. Firstly, it aims to establish and conduct a comprehensive analysis of the performance disparity between convolutional neural network (CNN) models that utilize only eye-only regions and those that incorporate full facial input. Second, this research focuses on detecting and assessing the influence of specific facial regions that are significant for gaze estimation.

Roadmap. This paper includes the following sections. Section 2 provides an overview of relevant literature. Section 3 explains the research approach for analyzing eye and full-face models and their facial feature contribution. Section 4 presents the conducted experiments and their results. Section 5 discusses and interprets the findings. Section 6 addresses the study's ethical implications. Finally, Section 7 summarizes the results, limitations, and future improvements.

2 Related Work

This section provides an overview of relevant research on the role of facial features in gaze estimation. Gaze estimation encompasses two main approaches: model-based [13], [14], [15] and appearance-based [16], [12], [11] methods. Model-based techniques employ mathematical models that capture the geometry and position models and their correlation with observed eye images. Appearance-based methods utilize machine learning to establish a mapping between the visual appearance of the eye including images and corresponding gaze directions. In this study, we employ appearance-based gaze estimation techniques to explore the impact of distinct facial regions on gaze estimation accuracy. Our study is related to previous work as discussed below.

Zhang et al. [16] explored the significance of various facial features in gaze estimation. The research involved occluding different facial regions using a grey-colored mask to assess the resulting decrease in estimation accuracy. Comparisons were also made between the performance of blocking the eye regions and utilizing full-face input. Their findings indicated that full-face input provides more information than head pose direction, highlighting the importance of incorporating full-facial input for improved gaze estimation accuracy.

Krafka et al. [12], in addition, introduced iTracker, a multi-region gaze regressor that inputs both eyes and full-face images and face grid. Their study showed the inclusion of full-face images is beneficial as it contributes to reducing the error rate and outperformed state-of-the-art approaches.

In a study by Palmero et al. [17], it was demonstrated that whole facial images contain richer information beyond just the eye regions, including illumination and pose direction. The researchers explored the influence of different facial components, such as eyes, full facial images, and facial landmarks, on the development of a multi-stream recurrent convolutional gaze estimation network. The findings indicated that incorporating geometric facial landmarks into appearance-based methods had a beneficial regularization effect on the

accuracy of gaze estimation. This highlights the significance of considering the entire facial context in improving the performance of gaze estimation models.

Sakurai et al. [18] introduced a study that investigated the correlation between facial and eye movements, highlighting the value of integrating visual cues from both the face and eyes for accurate gaze tracking. The research demonstrated the efficacy of incorporating facial direction estimation in achieving precise eye-tracking results, particularly when subjects have the freedom to move their eyes and head. This work emphasizes the importance of considering both facial and eye movements for robust gaze estimation.

3 Methodology

In this section, we propose the following experimental procedure as shown in Figure 1 to investigate the efficacy of different facial regions on eye gaze estimation. It aims to throughout identify and quantify the contribution of different important facial areas toward the accuracy of eye gaze estimation.

Figure 1. The research methodology employed to investigate the influence of different facial regions on eye gaze estimation.

The experimental methodology is divided into several subsections, each addressing a specific stage of the research process. Subsection 3.1 provides comprehensive details regarding developing deep convolutional neural network (CNN) models for gaze estimation. These models serve as the foundational results for further analysis. Subsection 3.2 focuses on the evaluation and analysis of two CNN models: the baseline eye-only CNN model and the full-face CNN model. The assessment aims to compare their performances and determine the impact of incorporating full-face information. To investigate the significance of specific facial areas, subsection 3.3 outlines the procedure used for region importance analysis. This analysis aims to identify which facial regions play a crucial role in gaze estimation. The final subsection 3.4 explains the quantitative analysis of the contribution of each identified facial landmark in subsection 3.3, conducting a more in-depth investigation into their actual influence on the accuracy of gaze estimation.

3.1 Model Development

This stage aims to create two eye-gaze regression models: a baseline model that takes two eye images as input, and a comparative model that processes a complete facial image. Specifically, both models implement a deep CNN architecture consisting of two main procedures: feature extraction using convolutional blocks and eye-gaze estimation using a series of fully connected layers to produce pitch and yaw values. To allow our baseline model to incorporate left and right eye images as its input features, we decided on a multi-input architecture that employs shared weight for the convolutional layers to receive two images of size $43 \times 73 \times 3$. This design allows the model to learn relevant features from both eyes simultaneously. The feature maps extracted from the convolutional blocks are then merged by concatenation and passed through the fully connected layers to make predictions. Similarly, our comparative model is developed utilizing the same convolutional blocks and fully connected layers to infer gaze direction from the full facial image of size $224 \times 224 \times 3$. By employing identical convolutional and fully connected architectures, we can perform a comprehensive performance comparison between different input features while mitigating the impact of network complexities.

3.2 Performance Analysis

The performance analysis is conducted to examine the disparity in estimation accuracy between using only eye images and utilizing full facial images. The analysis encompasses three stages: pre-tuning, fine-tuning, and testing. During the pre-tuning stage, the two models are trained on a dataset to enable them to learn relevant features that can be generalized across multiple test subjects. Fine-tuning, on the other hand, replicates the calibration techniques inspired by the works of Chen et al. [19] and Bandyopadhyay et al. [20]. This fine-tuning process is essential in acquainting the model with the specific instances of a given test subject, thus leading to improved performance. The testing phase utilizes the fine-tuned model and assesses its accuracy with regard to the angular loss by first converting the predicted pitch and yaw into the 3-dimensional XYZ domain, Equation 1, and computing the loss in degree as shown in Equation 2.

$$\begin{aligned} x &= cos(pitch) * sin(yaw) \\ y &= sin(pitch) \\ z &= cos(pitch) * cos(yaw) \end{aligned} \tag{1}$$

Equation 1 shows the conversion from pitch yaw vector to 3-dimensional vector in XYZ plane.

$$\begin{aligned} cosine\_similarity &= \frac{XYZ_{pred} \cdot XYZ_{truth}}{||XYZ_{pred}|| * ||XYZ_{truth}||} \\ angular\_loss &= cos^{-1}(cosine\_similarity) * \frac{180}{\pi} \end{aligned} \tag{2}$$

Equation 2 computes the angular loss between 2 vectors in a 3-dimensional XYZ plane in degree unit.

3.3 Region Importance Analysis

This stage adapts the Region Important Analysis approach, as proposed by the works of Zeiler et al. [21] and Zhang et al. [16], to gain insights into the contribution of different facial areas towards the accuracy of gaze estimation models. This approach allows us to identify and analyze the specific regions of the face that play a significant role in influencing the performance of these models. The approach employs a sliding window to occlude different regions of the face with a gray-colored mask allowing for the evaluation of the resulting decrease in accuracy. The occluded image is then fed into the

full-face model for gaze prediction, and the corresponding accuracy drop is measured. To construct the heat map for each image, a blur filter is applied, and the values are normalized to the range of [0,1]. A significant distinction in this work compared to the work proposed by Zhang et al. [16] is the utilization of the mean of max operation, illustrated by Algorithm 1, to obtain the region importance heat map for each test subject. This derivation is established to put significance to facial regions that benefit the estimation accuracy even when minor variations and rotations of facial images occur.

Algorithm 1 Calculate Heat Map using Mean of Max Operation for batches of heat map images.

for $i \leftarrow 0$ to $\#batches$ **do**
    $max[i] \leftarrow \max(batches[i], \text{axis} = 0)$
end for
$mean \leftarrow \text{mean}(max, \text{axis} = 0)$
$heatmap \leftarrow mean$
return $heatmap$

3.4 Region Contribution Measurement

To understand the relative importance of different facial regions in gaze estimation, this stage identifies and quantifies the accuracy contribution of each influential region. The process of identifying the significant facial regions involves localizing and segmenting areas that exhibit substantial contributions to the error based on the region importance heat map obtained in the previous Subsection 3.3. These regions are then encoded into rectangular masks. Each mask is subsequently utilized to fine-tune the full face model and evaluate the corresponding test subject, following the cross-validation procedure discussed in Subsection 3.2. By observing the resulting accuracy changes for all test subjects against the full-face model performance benchmark acquired by evaluating the comparative model, we gain valuable insights into the impact of the masked regions on gaze estimation accuracy.

4 Experiments

In this section, we present a series of experiments conducted based on the methodology outlined in Section 3. Our aim is to determine and quantify the accuracy contribution of different facial input features on the accuracy of gaze estimation. We begin by providing implementation details of the gaze estimation models, which serve as the fundamental components of our analysis. Next, in Subsection 4.2, we delve into the specifics of the dataset and metrics utilized for model construction and evaluation. Lastly, Subsection 4.3 showcases and examines the experimental work and its outcomes.

4.1 Model Architectures

This study introduces two distinct models with different specifications. The first baseline model is specifically designed to process two eye images, which are cropped from the original full facial image, as its input. On the other hand, the subsequent comparative model takes the entire full facial image as its input. The implementation details of these two models are provided below. The reference implementations used in this paper are publicly accessible here [GitHub].

The baseline eye-only model architecture. The inclusion of two eye areas is a crucial descriptor in gaze regressor models as the direct gaze direction visual information they provided. To establish the baseline benchmark model, a deep CNN based on the ResNet18 architecture [22] is adapted to predict the 3D gaze direction solely from the left and right eye regions.

Figure 2 illustrates the architecture of the baseline eye model, which incorporates two eye images as input. Each image is processed through the Convolutional layers of ResNet18 to extract relevant features. These features are then concatenated and passed through an Average Pooling layer, followed by a final fully connected layer that predicts the pitch and yaw values. It is worth noting that the ResNet18 convolutional layers used to learn features from both eye images are shared weights.

The comparative Full-Face model architecture. To examine the impact of facial areas beyond the eye regions, a CNN model that utilizes full facial images as input is developed. The model architecture adopts the ResNet18 architecture as its backbone. By utilizing the same architectural design as the baseline eye model, we can obtain a throughout understanding of how different input features influence the performance of gaze estimation while minimizing the influence of varying model complexities.

Figure 3 provides a detailed depiction of the implementation of the full face model. This model utilizes ResNet18 Convolutional layers to extract relevant features from the input full-face image. One distinction from the baseline eye model is that the features extracted from the convolutional layers directly pass through the Average Pooling and Fully Connected layers, eliminating the need for a concatenation layer.

4.2 Experimental Setup

Dataset. In this study, the MPIIFaceGaze Dataset [23], [24] is utilized to develop and assess the performance of both the baseline, shown by Figure 2, and comparative gaze estimation models, shown by Figure 3. The evaluation is conducted using cross-validation with 15 test subjects. Examples of the dataset are depicted in Figures 4, 5, and 6. The dataset comprises 3000 facial images and annotation vectors with 28 dimensions for each of the 15 test subjects. The first two dimensions of the annotation vectors encode the ground-truth gaze direction information, while the coordinates of the left and right eye corners are represented in 8 dimensions at index 4, 5, 6, 7, 8, 9, 10, and 11. For this study, the mentioned dimensions are crucial in constructing and assessing the models. Both models undergo training using data from 14 test subjects and are subsequently evaluated on the remaining subject, incorporating calibration techniques, inspired by a higher accuracy performance as discussed in the works of Chen et al. [19] and Bandyopadhyay et al. [20]. This iterative process is repeated 15 times, covering all 15 test subjects. The accuracy achieved by each model is then averaged across all test subjects to assess their performance.

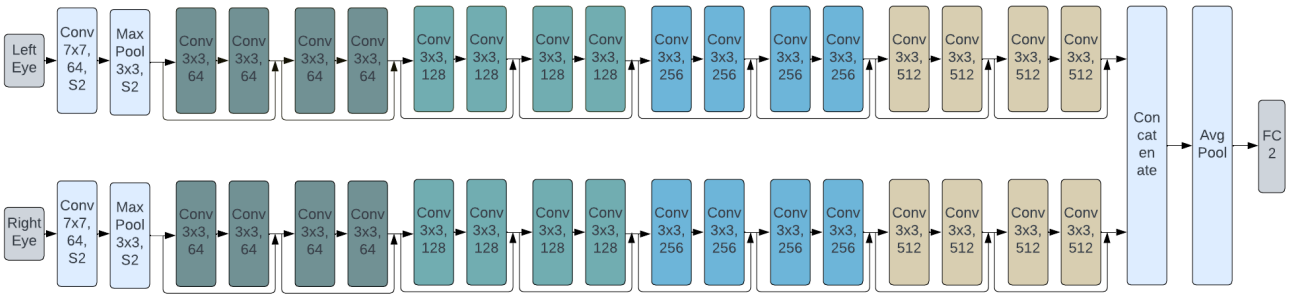


Figure 2. CNN Architecture for Eye baseline model based on ResNet18 Convolutional backbone.

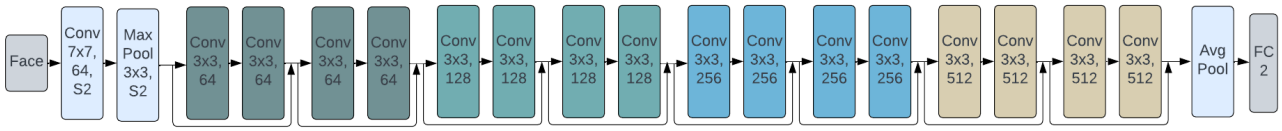
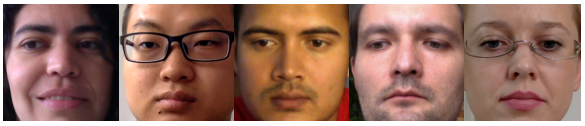


Figure 3. CNN Architecture for full face input model based on ResNet18 Convolutional backbone.



(a) (b) (c) (d) (e)

Figure 4. Examples of MPIIFaceGaze Dataset: First image index, test subject IDs 0, 1, 2, 3, and 4 from left to right.

(a) (b) (c) (d) (e)

Figure 5. Examples of MPIIFaceGaze Dataset: First image index, test subject IDs 5, 6, 7, 8, and 9 from left to right.

(a) (b) (c) (d) (e)

Figure 6. Examples of MPIIFaceGaze Dataset: First image index, test subject IDs 10, 11, 12, 13, and 14 from left to right.

Metrics. To train and evaluate the performance of the models, Mean Absolute Loss (L1) and Angular Loss are utilized as the chosen metrics. First, the L1 loss is utilized to minimize the loss of the regression models shown in Equation 3. Second, the angular loss given by Equation 2 is used for the evaluation procedure, including the determination of the best-performing model and result comparisons. This angular loss metric is particularly valuable as it measures the gaze direction loss in degrees, providing a more intuitive understanding of the model's performance. By employing the angular loss, we can effectively assess how accurately the models capture the desired gaze direction, thus gaining insights into their overall performance.

$$L1 = \frac{1}{N}\sum_{i=1}^{N} |gaze_{true}^{(i)} - gaze_{pred}^{(i)}| \tag{3}$$

Equation 3 explains the Mean Absolute Loss (or L1 Loss) function used to train the gaze estimation models.

Training Specification. The training phase pre-tunes both models adhering to the following specifications. The procedure utilizes a dataset comprising 14 test subjects, with each subject contributing 3000 images. The total of $42,000$ images is divided into pre-tune and evaluation datasets using a ratio of 0.9 and 0.1, resulting in $37,800$ and $4,200$ images respectively for the pre-tune and evaluation sets. The models are trained for 20 epochs, and the model with the best performance based on the angular error (Equation 2) is saved at each epoch, which is utilized to test the model. The training uses a batch size of 32, a learning rate of 10^{-4}, and Adam optimizer [25]. The model is optimized to minimize the L1 loss function, Equation 3, between the predicted and ground-truth 2-dimensional pitch, yaw vectors.

Testing Specification. Having obtained the best model from the training procedure, the testing stage begins by

initially fine-tuning the model and subsequently evaluating its performance using the following approach. The testing is conducted for the one remaining test subject, with a total of 3000 images. The test dataset is divided into fine-tuning, also known as calibration, and evaluation sets, with a ratio of 0.1 and 0.9 respectively. This division results in 300 images allocated for fine-tuning and 2700 images designated for evaluation purposes. The adoption of fine-tuning, or calibration in other words, is inspired by the works of Chen et al. [19] and Bandyopadhyay et al.[20]. The calibration dataset, 300 images, is used to fine-tune the best pre-tuned model obtained in the training phase. In order to prevent overfitting, the Batch Normalization layers of the models are frozen before the fine-tuning process takes place. The hyperparameters used during the calibration training phase remain similar to those of the pre-tuning phase. The models are trained for 30 epochs, and the evaluation results against the evaluation dataset, which consists of 2700 images, are averaged over the last 10 epochs to obtain the final evaluation result. The average operation is utilized to compute a robust final evaluation result, mitigating the impact of fluctuation around the convergence point. This approach helps ensure stability and reliability in the obtained evaluation outcome.

4.3 Results

The cross-validation results of the baseline eye-exclusive input model are depicted in Figure 7. The angular error results for the full-face model are provided in Figure 8. The histograms provide the angular error in degree per test subject and the average loss among all subjects. The average angular loss across the 15 test subjects is calculated to be 3.316 ± 0.603 for the eye-only model and 2.924 ± 0.594 for the full-face model. These results evidence that utilizing the full facial input significantly improves the error of gaze estimation, with the full-face model outperforming the eye-only model by over 11%, given by Equation 4.

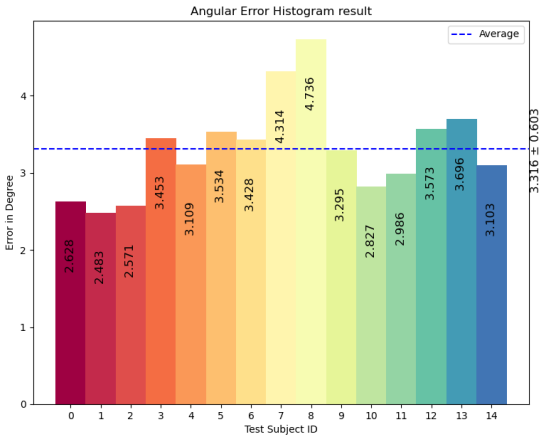


Figure 7. Cross-Validation results for the baseline eye-only model over 15 MPIIFaceGaze test subjects.

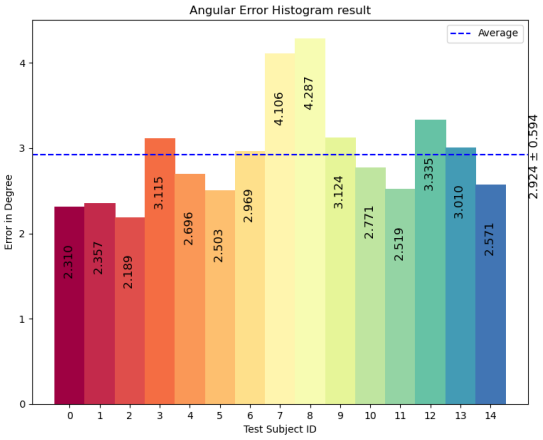


Figure 8. Cross-Validation results for the full face model over 15 MPIIFaceGaze test subjects.

4.4 Influential Facial Regions

The improved performance by utilizing full facial images implies that facial landmarks beyond the eye areas are contributing important cues that can leverage gaze estimation accuracy. To further investigate the contributive facial factor, this experiment applies the regional importance analysis approach (described in Subsection 3.3) to identify the significant facial regions encoded in the heat map representation for 15 test subjects. The experiment employs specific hyperparameters: a filter size of 32×32, a stride of 16, a mask value of 127, a blur filter of size 32, and a batch size of 85. Examples of the applied sliding windows at different positions are shown in Figure 9. The acquired 3000 heat maps using the occluded facial images for each subject are divided into batches of size 85 and applied the mean of max Algorithm 1 to obtain the final region importance heat map. The experiment is conducted on all 15 test subjects, using the respective fine-tuned full facial models trained on their individual data. The resulting heat maps are illustrated in Figures 10, 11, and 12. The heat maps assign significance to the error proportionally according to the corresponding facial area, emphasizing the regions that are valuable for gaze estimation.

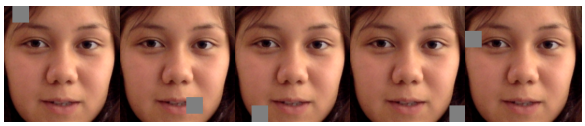


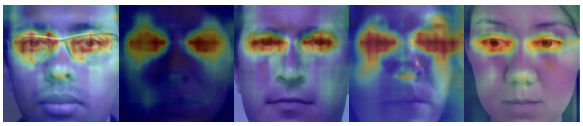
Figure 9. Region Importance Analysis sliding windows applied to the first image of test subject 14. Shown from left to right: $(16, 32)$, $(192, 144)$, $(208, 48)$, $(208, 208)$, and $(64, 16)$ positions with the box filter implemented

(a) (b) (c) (d) (e)

Figure 10. Regional importance heat maps for test subject ids 0, 1, 2, 3, and 4 respectively from left to right.

(a) (b) (c) (d) (e)

Figure 11. Regional importance heat maps for test subject ids 5, 6, 7, 8, and 9 respectively from left to right.



(a) (b) (c) (d) (e)

Figure 12. Regional importance heat maps for test subject ids 10, 11, 12, 13, and 14 respectively from left to right.

The acquired individual region importance heat maps are averaged to identify the facial regions that consistently hold significance across all 15 test subjects. The resulting heat map, Figure 13, highlights the importance of different facial regions in accurately estimating gaze.

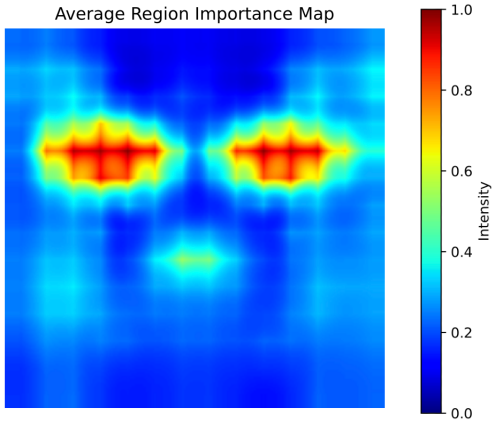
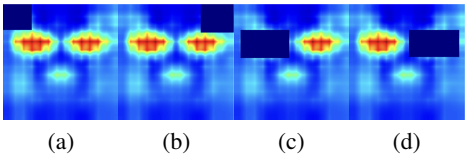


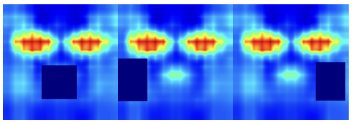
Figure 13. Unweighted Averaged Region Importance Heatmap computed with all 15 test subjects' heatmaps

4.5 Region Accuracy Contribution

We have decoded the influential areas depicted in the heat map Figure 13. The analysis has led us to identify seven specific facial areas that play a crucial role in accurately estimating gaze. These areas are highlighted in Figure 14 below.



(a) (b) (c) (d)



(e) (f) (g)

Figure 14. The identified regions contribute the most to the overall accuracy of gaze estimation. The regions are highlighted by a dark rectangular shape.

The regions are identified as the right eyebrow (a), left eyebrow (b), right eye (c), left eye (d), nose (e), right cheek (f), and left cheek (g).

To quantify the contribution toward gaze estimation accuracy of the identified facial landmarks, we conducted an experimental approach discussed in subsection 3.4 to occlude each landmark encoded by a rectangular mask and compare the performance with the benchmark data illustrated in Figure 8. The mask rectangular coordinates are given in Table 1. The regions excluded by the rectangular masks are illustrated in Figure 15. The experiment involved the use of occluded image data to fine-tune and evaluate, following a similar experimental setup as described in Subsection 4.4 using all test subjects for each occluded region. The results of this evaluation are presented in Table 2.

Region	x_0	y_0	x_1	y_1
Right Eyebrow	0	0	56	50
Left Eyebrow	160	0	224	55
Right Eye	13	50	107	104
Left Eye	116	50	213	102
Nose	75	118	143	183
Right Cheek	0	105	57	187
Left Cheek	160	112	217	186

Table 1. Rectangular coordinates in pixels for 7 identified facial landmarks, depicted relative to a width and height of (224, 224).

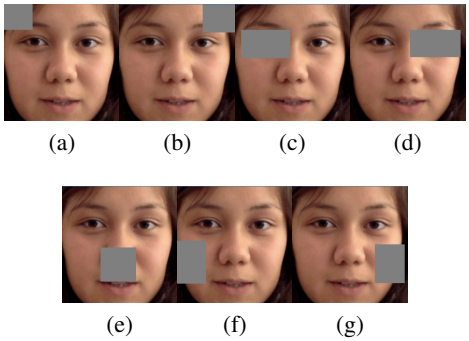


Figure 15. From left to right, the figure shows examples of occluded facial regions: right eyebrow, left eye-brow, right eye, left eye, nose, right cheek, and left cheek for the first image of test subject 14.

Occluded Region	Averaged Angular Error (Deg)
Right Eyebrow	3.007 ± 0.604
Left Eyebrow	2.973 ± 0.649
Right Eye	3.502 ± 0.754
Left Eye	3.518 ± 0.720
Nose	3.033 ± 0.626
Right Cheek	3.000 ± 0.628
Left Cheek	2.968 ± 0.614

Table 2. Averaged angular error evaluated for each occluded facial region in degree.

The ranked importance of facial features, determined by the averaged angular loss, is as follows: eyes, nose, eyebrows, and cheeks. To quantitatively assess the contribution value of each identified region, we utilize Equation 4 and compare it against the benchmark accuracy of 2.924 degrees achieved with full facial input.

$$\text{contribution} = \frac{err_{\text{region}} - err_{\text{benchmark}}}{err_{\text{benchmark}}} \cdot 100\% \qquad (4)$$

The relative error improvement contributions are as follows: Right Eyebrow (2.7%), Left Eyebrow (1.6%), Right Eye (16.5%), Left Eye (16.8%), Nose (3.5%), Right Cheek (2.5%), and Left Cheek (1.4%). These results demonstrate the relative impact on accuracy for each specific facial region.

5 Discussion

Performance comparison. Our results, Figures 7 and 8, highlight the superior performance of the comparative model that incorporates a full facial image as its input, surpassing the baseline model that solely relies on the eye regions. Our findings reveal a significant improvement in estimation accuracy, by more than 11%, when utilizing a full facial image. This finding emphasizes the crucial role played by other facial regions, in addition to the left and right eye areas, in enhancing gaze estimation accuracy.

Facial region contributions. The experiments utilizing region-important analysis and contribution study have identified the facial areas: right eyebrow, left eyebrow, right eye, left eye, nose, right cheek, and left cheek that plays an important role in improving the estimation accuracy. The obtained results clearly demonstrate the relative impact of different facial regions, with the nose emerging as a particularly influential feature, just behind the eye regions. The experiments show that the nose region can contribute up to 3.5% improvement in relative accuracy. The high contribution of the nose region can be attributed to its ability to provide valuable cues related to head pose direction. The eyebrow area contains specific features that are related to eye movements and gaze direction such as the position of the eyebrow arch or the presence of wrinkles can provide clues about the upward or downward gaze. During our dataset analysis, we discovered that the left and right cheek regions can be helpful in determining the direction of a test subject's gaze. In our observations, we noticed that when the face image leaned in a particular direction, the cheek areas exhibited variations and revealed background information as well, offering valuable information about the test subject's likely gaze direction, as illustrated in Figure 16. These findings reinforce the notion that the cheek areas can improve gaze prediction.

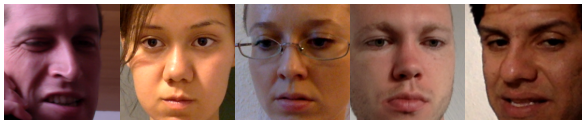


Figure 16. From left to right, the figure shows instances where the test subject leans toward one direction, and the cheek regions can provide an important cue for estimating gaze direction. From left to right, the figure shows examples of test subjects 12, 14, 4, 5, and 8 respectively.

6 Responsible Research

This study utilizes deep learning techniques and a dataset containing facial information. Throughout our research, we have implemented several measures to ensure the integrity and reproducibility of our findings. In Section 6.1, we discuss the ethical considerations involved in employing the deep learning model and dataset. Additionally, Section 6.2 elaborates on the reproducibility of our research results and implementation.

6.1 Scientific Integrity

The MPIIFaceGaze dataset [23], [24], which contains facial images and encoded landmark information, is employed in this study. It is essential to acknowledge the potential risk of re-identification faced by the dataset participants. To address this risk, our research takes measures to anonymize the participants by withholding personal information such as names, ages, and addresses.

In terms of deep learning, our research upholds ethical principles and ensures the credibility and dependability of our findings. We provide transparent explanations of the

methodology, algorithms, and implementation details in Section 3, and offer reference implementations used in this study [GitHub]. To eliminate discrimination and biases towards specific user groups, the dataset utilized in constructing and evaluating our models incorporates the data of all test subjects through a cross-validation approach.

6.2 Reproducibility

Ensuring the reproducibility of our research is an important aspect to allow for further exploration and improvement of the proposed methods. We publicly open-source our code on GitHub, including the baseline and comparative models architecture and necessary evaluation programs to obtain our findings. The methodology and algorithms ideas can be found in Section 3 and hyper-parameter setups are presented in Subsection 4.2. It is important to acknowledge the stochastic property caused by the random weight initiation procedure when training the provided model, which can cause the results to fluctuate and not fully equal to the results illustrated in this work. We have performed average operations through many epochs to mitigate these fluctuations and improve the robustness of our obtained results. Our reference implementation includes the necessary packages and version information that can be deployed in any environment to ensure the program behaves deterministically in reproduced setups. Further improvements or feedback to our work can be requested or reported via pull request features.

7 Conclusions and Future Work

In this work, we studied the importance of facial features in the gaze estimation problem. We proposed the experimental methodology and implementation that demonstrated the improved performance, to more than 11%, with the inclusion of full facial input to predict gaze direction in comparison to using only eye regions. By adapting the region importance analysis method [21], [16] and cross-validation, we have identified the 7 facial landmarks: right eyebrow, left eyebrow, right eye, left eye, nose, right cheek, and left and their perspective contribution values in improving relative accuracy of gaze estimation.

It is important to acknowledge that while the performed experiment captures the independent important facial regions, it does not show the correlation impacts. In future work, we aim to deploy combination feature analysis to gain insight into how combined facial features impact accuracy. Furthermore, we aim to study the important facial feature by deep learning interpretation methods such as GradCam [26] and FullGradient representation [27] to understand which features are learned by the CNN models. This information can explain what are the important features considered by the networks.

References

[1] Jaeyeon Hwang, Juyoung Park, and Kuk-Jin Yoon. Gaze estimation for human-computer interaction. *Computer Vision and Image Understanding*, 165:24–37, 2017.

[2] SangHun Han, Xuan Zhang, Yusuke Sugano, and Mario Fritz. Real-time eye gaze tracking with a consumer-grade rgb-d sensor. *ACM Transactions on Graphics (TOG)*, 36(6):192, 2017.

[3] Roy S. Hessels and Ignace T.C. Hooge. Eye tracking in developmental cognitive neuroscience – the good, the bad and the ugly. *Developmental Cognitive Neuroscience*, 40:100710, 2019.

[4] Heiko Drewes. Eye gaze tracking for human computer interaction. 03 2010.

[5] W. X. Chen, X. Y. Cui, J. Zheng, J. M. Zhang, S. Chen, and Y. D. Yao. Gaze gestures and their applications in human-computer interaction with a head-mounted display, 2019.

[6] Francisco Vicente, Zehua Huang, Xuehan Xiong, Fernando De la Torre, Wende Zhang, and Dan Levi. Driver gaze tracking and eyes off the road detection system. *IEEE Transactions on Intelligent Transportation Systems*, 16:1–14, 08 2015.

[7] Sayyed Mudassar Shah, Zhaoyun Sun, Khalid Zaman, Altaf Hussain, Muhammad Shoaib, and Lili Pei. A driver gaze estimation method based on deep learning. *Sensors*, 22(10), 2022.

[8] Akshay Rangesh, Bowen Zhang, and Mohan M. Trivedi. Driver gaze estimation in the real world: Overcoming the eyeglass challenge. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1054–1059, 2020.

[9] Kenneth Holmqvist and Richard Andersson. *Eye-tracking: A comprehensive guide to methods, paradigms and measures*. 11 2017.

[10] 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.

[11] Raimondas Zemblys, Diederick C Niehorster, and Kenneth Holmqvist. gazenet: End-to-end eye-movement event detection with deep neural networks. *Behavior research methods*, 2018.

[12] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone, 2016.

[13] R. Stiefelhagen, Jie Yang, and A. Waibel. A model-based gaze tracking system. In *Proceedings IEEE International Joint Symposia on Intelligence and Systems*, pages 304–310, 1996.

[14] 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, 2012.

[15] James O'Reilly, Ahmed Shehab Khan, Zhiyuan Li, Jie Cai, Xiangyu Hu, Min Chen, and Yan Tong. A novel remote eye gaze tracking system using line illumination sources. pages 449–454, 03 2019.

[16] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, jul 2017.

[17] Cristina Palmero, Javier Selva, Mohammad Ali Bagheri, and Sergio Escalera. Recurrent cnn for 3d gaze estimation using appearance and shape cues, 2018.

[18] Keiko Sakurai, Mingmin Yan, Hiroki Tamura, and Koichi Tanno. A study on gaze estimation system using the direction of eyes and face. In *2016 World Automation Congress (WAC)*, pages 1–6, 2016.

[19] Zhaokang Chen and Bertram E. Shi. Offset calibration for appearance-based gaze estimation via gaze decomposition, 2020.

[20] Nairit Bandyopadhyay, Sébastien Riou, and Didier Schwab. Effect of personalized calibration on gaze estimation using deep-learning, 2021.

[21] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, pages 818–833, 2014.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[23] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2299–2308, 2017.

[24] 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 (TPAMI)*, 41(1):162–175, 2019.

[25] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[26] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, oct 2019.

[27] Suraj Srinivas and Francois Fleuret. Full-gradient representation for neural network visualization, 2019.