Channel Selection for Faster Deep Learning-based Gaze Estimation in the Frequency Domain

Author: Thijs Penning T.J.Penning@student.tudelft.nl

Supervisors: Dr. G. (Guohao) Lan, Lingyu Du

1. Background

- Human gaze is an important non-verbal cue that communicates valuable information used in various fields of study.
- Estimating gaze using deep learning convolutional neural networks (CNN) is computationally expensive, leading to latency.
- Reducing the input data size can lower these computational costs.
- By converting the data to the frequency domain and applying channel selection, the input data can be reduced up to 87.5% with marginal compromise to accuracy [1].

2. Research Question

"What is the impact of channel selection on the latency and accuracy of frequency domain gaze estimation?"

Additionally, the goal was to find channel selections that provide an optimal trade-off between a maximal speedup with marginal accuracy loss.

3. Methodology

- The images were transformed from RGB to the YCbCr color space.
- The YCbCr color space images were then converted to the frequency domain using the discrete cosine transform (DCT).
- In the frequency domain, each component (Y, Cb, and Cr) consists of 64 channels.
- The data within these channels was analyzed and visualized, as illustrated in Figure 1.

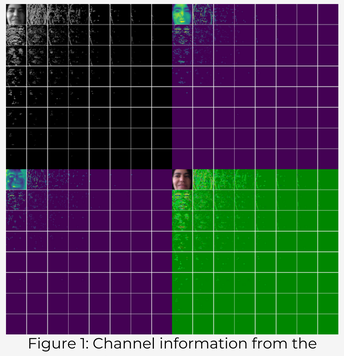

Figure 1: Channel information from the frequency domain.
Y (top-left), Cb (top-right), Cr (bottom-left), and the bottom-right quadrant represents the combined channels.

- Various models using a selection of these channels were tested. Two methods were used for the channel selection process:
  - Static channel selection: Manual selections based on channel analysis (like Figure 1) were made for all subjects.
  - Dynamic channel selection: The model learns to select channels, varying its selection based on the provided image. (Inspired by [1])

4. Experiment & Results

- The experiment utilized the MPIIFaceGaze dataset [2] and the AlexNet [3] and ResNet-18 architectures [4].
- In the experiment frequency domain models using the channel selections displayed in Table 1 were compared to models using RGB and YCbCr images.
- Each model went through training, calibration and inference ten times from scratch.
- During each phase a time and accuracy component was measured. Of the experiment, the mean and standard deviation of these measurements were recorded.

Table 1: Channel selections for the experimental models.
The indices are related to the layout in Figure 1 where 0 is the top left, and 63 the bottom right. The rest of the indices are distributed row by row.

Selection ID's	Y Channel	Cb/Cr Channel
FD0	all	all
FD1	[0]	[0]
FD2	[0, 1, 8]	[0, 1, 8]
FD3	[0, 1, 2, 8, 9, 16]	[0, 1, 8]
FD4	[0, 1, 2, 3, 8, 9, 10, 16, 17, 24]	[0, 1, 3, 8, 24]
FD5	[0, 1, 2, 3, 8, 9, 10, 16, 17, 24]	[0, 1, 2, 8, 9, 16]
FD6	[0, 1, 2, 3, 4, 8, 9, 10, 11, 16, 17, 18, 24, 25, 32]	[0, 1, 2, 3, 8, 9, 10, 16, 17, 24]
FDD7	dynamic	dynamic

- The best results achieved for each architecture along with their speedups and error increase:
  - For the AlexNet, selection FD3 achieved speedups of 3.3, 4.0, and 1.35 with only a marginal 0.05 degrees error increase.
  - In case of the ResNet-18, selection FD4 came out on top with speedups of 1.5, 1.7, and 1.35 and a 0.44 degrees error increase.
- A full overview of the results in the inference phase is illustrated in Figure 2.

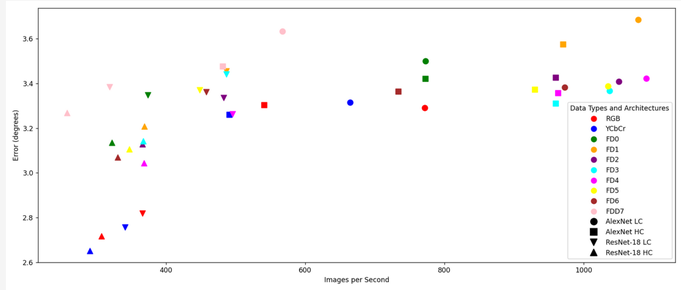

Figure 2: Inference results of the data types combined with the architectures.
The color represent the data types and the shapes the architectures. The LC or HC indicates the regular version of the model or a modified double channel version which was also tested.

5. Conclusion & Limitations

- Applying channel selection to the frequency domain data resulted in faster models with speedups ranging from 1.35 to 4.0, with marginal to slight compromise in accuracy of 0.05 and 0.44 degrees.
- The structure of the network played an important part in the results indicated by the differences observed.
- Furthermore, considering the data used and the specific experiment setup, the models with the static selections outperformed the model using the dynamic selection.

- The dynamic channel selection model suffered from a lack of data variety.
- The architectures used were designed for 224x224 images, limiting the frequency models accuracy.

References

- [1] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. Learning in the frequency domain. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1740–1749, 2020.
- [2] https://perceptualui.org/research/datasets/MPIIFaceGaze/
- [3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6):84–90, 2017.
- [4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016