Learning-based Region Selection for End-to-End Gaze Estimation

Supplementary Materials

Xucong Zhang[1]
xucong.zhang@inf.ethz.ch
Yusuke Sugano[2]
sugano@iis.u-tokyo.ac.jp
Andreas Bulling[3]
andreas.bulling@vis.uni-stuttgart.de
Otmar Hilliges[1]
otmar.hilliges@inf.ethz.ch

[1] Department of Computer Science, ETH Zurich, Zürich, Switzerland
[2] Institute of Industrial Science, The University of Tokyo, Tokyo, Japan
[3] Institute for Visualisation and Interactive Systems, University of Stuttgart, Stuttgart, Germany

1 Introduction

This document contains additional materials supplementing the experimental results in the main paper. We first investigate alternative region averaging strategies. We then provide additional evaluations to assess when our method is beneficial over traditional approaches. We finally discuss the impact of allowing the *RSN* to not only choose a sub-region from the input image but also to change the size of the crop.

1.1 Effect of region averaging

Although we show that increasing number of regions tends to improve performance, it is computationally expensive to add more regions to the network. The possible combinations of regions grows exponentially with more selection steps, and each region will increase the total length of the concatenated feature vector for *gaze net*. These two factors also make the initial training of the *gaze net* with random regions more difficult.

Inspired by [1] which uses averaged image features from multiple point of views to estimate human pose, we evaluate the effect of using averaged features from multiple regions. We used our single-region model and take the top-k regions with the highest probabilities as input to the *gaze net*. We trained one model with each different k for evaluation.

The results are shown in Figure 2, together with the performance of our two-region and three-region models as references. We can see that, even with the averaged features, the use of more regions results in lower gaze estimation error. The model achieves better performance than the two-region model when the model uses the averaged feature from six regions. However, the performance is still worse than the three-region model until adding a seventh

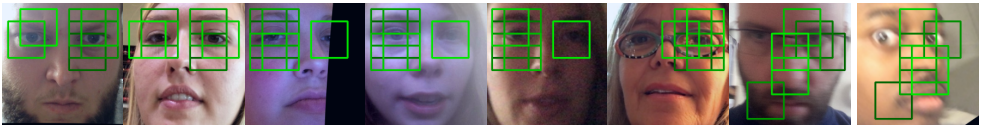

Figure 1: Examples of the top seven regions with highest probabilities from our single region model. The green rectangle indicates the selected region. Our model selects region around two eyes (a,b), focuses on one eye due to visibility (c,d) or brightness (e,f). Extreme motion blur can cause failure on selecting eyes (g,h).

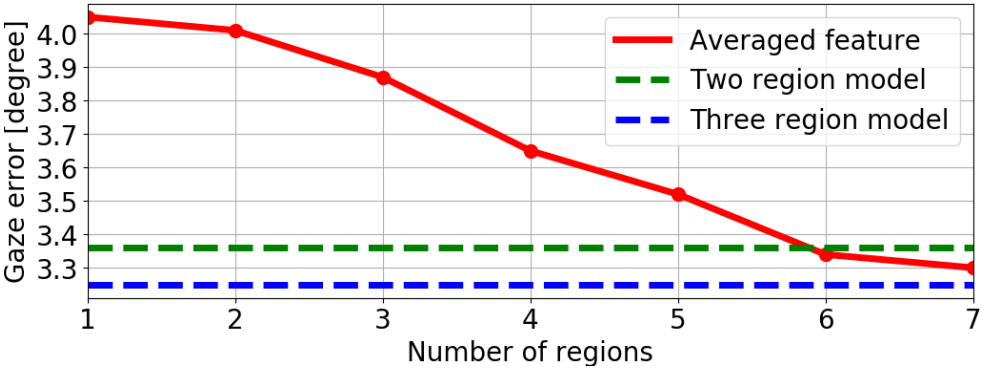

Figure 2: Experiments on averaging features from multiple regions. The x-axis indicates numbers of regions and y-axis is gaze estimation errors in degree. We also show the gaze estimation error from two-region model and three-region model.

region. This is because these regions from the single-region model are independent with each other therefore can have overlap with each other which contain redundant information.

Figure 1 shows some examples of the top seven regions proposed by the one-region *RSN*. The figure shows that regions are located around two eye (Figure 1 (a,b)), and focus on one eye due to visibility (Figure 1 (c,d)) or brightness (Figure 1 (e,f)). Again, region selection is difficult and more spread out for blurry faces (Figure 1 (g,h)).

2 Impact of different head poses and illumination conditions

In this section, we unpack the effect of our method in the presence of difficult environmental conditions. In particular, we provide more evidence that the method selects meaningful sub-regions across different head poses and illumination conditions. In Figure 3 we select sample images based on their head-poses, varying across the entire range of horizontal rotation and compare our method to the baseline (without dynamic region selection). The top row shows examples from the baseline method, always using the left eye region as input, and the second row shows examples from *our* method which dynamically picks the sub-region based on the content of the input image. The green rectangles indicate the selected region. We furthermore show the gaze estimation error in an inset in each image. While both meth-

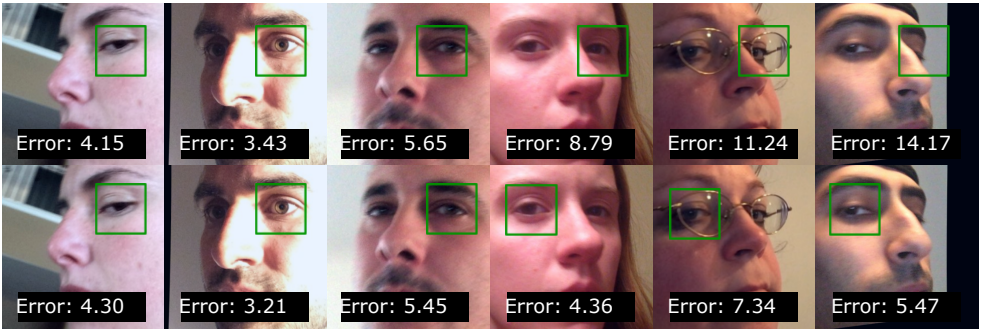

Figure 3: Examples of selected regions and gaze estimation error (in inset). *Top row:* results from the static region baseline. *Second row:* results from our method across different head poses. Our method can dynamically select the visible eye to achieve better gaze estimation performance.

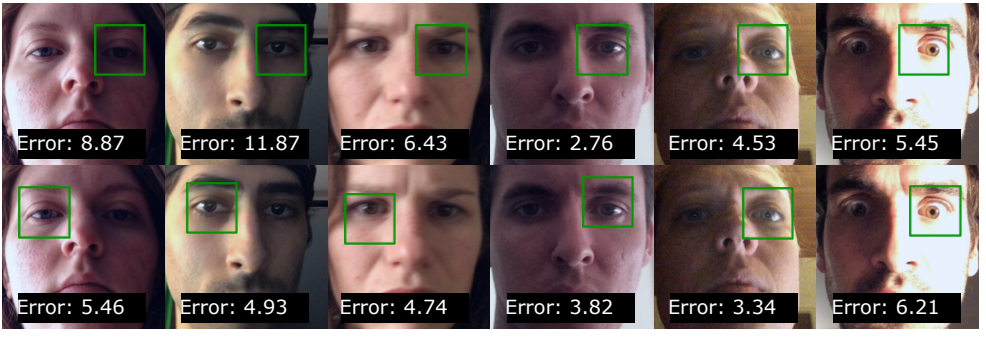

Figure 4: Examples of selected regions and gaze estimation error (in inset). *Top row:* results from the static region baseline. *Second row:* results from our method across different illumination conditions. Our method can dynamically select the brighter eye to achieve better gaze estimation performance compared to the baseline.

ods attain comparable results under good visibility of the left eye (Figure 3, left half), our method clearly produces better results once the left eye is at least partially occluded (right part of Figure 3).

A similar pattern occurs when unpacking the effect of varying lighting conditions. Figure 4 shows how our method selects region compared to the baseline. Analogously here the baseline always chooses the left eye as input. Depending on which side of the face is brighter this can have a significant impact on the gaze estimation performance. In contrast, our method can pick the brighter eye region to yield better performance compared to the baseline method (left part of Figure 3). In general, it can be seen that the gaze error remains more stable across different head-poses and lighting conditions. This is also reflected in quantitative results. Our method achieves both a lower mean gaze estimation error *and* a lower standard deviation compared to the baseline. (Evaluated on the GazeCapture testset. Ours: 4.05° mean error, $SD = 3.15°$ versus baseline: 4.35° mean error and $SD = 3.34°$).

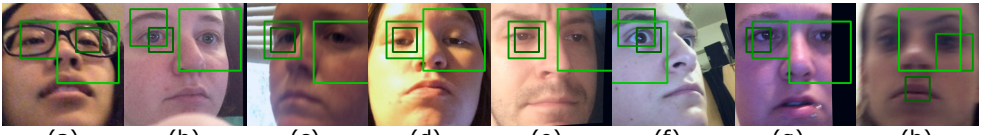

Figure 5: Examples of the selected three regions with different sizes. The green rectangle indicates the selected region. Our model selects regions contain both eyes for most of samples (a-c), focuses on a single eye due to brightness (d,e) or visibility (f,g). Extreme motion blur is a typical failure case (h).

3 Region size

In most of our experiments we use a single, fixed size for the selected sub-regions (0.3× the input image). This size was choosen to limit the search space during training. However, this is an additional dimension that potentially impacts the *gaze net* performance.

To further understand the impact of crop size, we performed an experiment with the three-region + face model as reported in the main paper, but allowed different sizes for each of the three regions (0.2, 0.3 and 0.5 times the input size respectively). The experiment was conducted as within GazeCapture dataset evaluation. The model with fixed region size achieves 3.22 degrees gaze estimation error, and the model that is allowed to vary the crop size achieves even better results (3.14 degrees). This suggests that there is additional potential in selecting the pixels that are informative for the final task. It would be of interest to explore even more sophisticated strategies, potentially including soft-attention mechanisms, in future work.

Figure 5 illustrates that the network leverages the differently sized crops as in a focus+context fashion. For most of samples, two regions are located on one eye but are chosen at different scales (Figure 5(a-c)). However, depending on the lighting condition (Figure 5 (d,e)) and visibility (Figure 5 (f,g)), the same dynamic region selection behavior as in the single region case can be observed. Extreme blur remains a challenging setting (Figure 5 (h)).

References

[1] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1175–1186, 2019.