Gaze Fixation Prediction for Composable Image Transformations

James Youngblood
University of Utah
james@youngbloods.org

Rogelio Cardona-Rivera
University of Utah
r.cardona.rivera@utah.edu

***Abstract*—We contribute simple heuristic functions for estimating how much gaze fixation prediction models adhere to real gaze distributions under an arbitrary composition of image transformations. Notably, computing these heuristics does not require the collection of real gaze distributions. We analyze the effectiveness of the heuristics by computing their correlation to the divergence of predicted gaze distributions from real gaze distributions on a set of image transformations.**

***Index Terms*—gaze fixation prediction, image filters, stylistic post-processing**

I. Introduction

Visual media producers study the process of viewing. Studios which produce movies, games, illustration, or other visual media have a professional interest in how a viewer will direct their gaze over the presented image. With better knowledge, they can produce pleasing image compositions (i.e. the "rule of thirds"), or estimate the viewer's situational awareness (i.e. noticing a figure in the dark). In order to exert maximum control over the viewing experience, producers must intuit the complexities of human visual behavior, or gather a number of subjects and track their gazes directly in a lengthy study.

Recent research presents an alternative. Machine learning has greatly improved the adherence of computational models of gaze fixation to real human gaze behavior. The current state-of-the-art is DeepGaze IIE [1], boasting ~80% adherence to real gaze distributions on benchmarks[1]. These models may now provide more signal of human behavior than noise, and so might provide useful feedback in creative processes. Crucially, they do so in a matter of seconds. We believe that these models are an important step towards greater automation and more compelling experiences in visual media.

Alas, these models may not be applicable to many styles of visual media. Gaze prediction models are notably biased towards candid photography: Matthias Kümmerer et al. [2] show that all leading gaze prediction models utilize transfer learning from many other domains, including object recognition. This means that gaze prediction models use pre-trained object recognition models as a starting point in the training process for gaze prediction. The object recognition task prioritizes candid photography over stylistic representations due to the assumption that object recognition applications usually involve unaltered images captured with a camera, but this assumption does not hold true for the gaze prediction task. Because the assumption is implicitly encoded into most of the training data a gaze prediction model will see, the model's performance may suffer when generalizing to stylized images.

Furthermore, a study by Zhaohui Che et al. [3] shows that image transformations influence gaze fixations, sometimes in ways that are difficult to model. Modern visual productions use multiple steps of image processing and rendering for stylistic effect. These represent a chain of image transformations, for which gaze prediction models may not produce robust predictions. Demonstrating robustness, or the lack thereof, for a wide variety of transformations, not to mention combinations of transformations, is a daunting task if human trials are required; we hope to prove a more scalable approach.

We contribute simple heuristic functions, which are statistically significant signals for how adherent models are to gaze prediction under an arbitrary composition of image transformations. Notably, computing these heuristics does not require the collection of real gaze distributions. We analyze the effectiveness of the heuristics by checking how they correlate to the divergence of predicted gaze distributions from real gaze distributions on a set of image transformations. We prospect on the value heuristics might provide for explaining underlying mechanisms of gaze prediction models.

II. Related Work

The effects of image transformations–including cropping, rotation, skewing, and edge detection–on gaze behavior are studied by Zhaohui Che et al. [3]. Their paper is motivated by the possibility of augmenting gaze prediction datasets with transformed images. We extend their work by repeating their experiments using current state-of-the-art models which had not been published at the time. We describe this experiment in detail in the Method section. We use the results of these experiments to study the effectiveness of computed heuristics on model predictions for transformed images.

The effects of digital image alterations on gaze behavior have also been studied by Quian Quiroga et al. [4], who performed an experiment collecting gaze fixations before and after manual Photoshop edits were made to photographs of paintings. Insufficient explanation for the intention behind

[1]The DeepGaze IIE paper refers to this adherence percentage as the ratio of a model's "information gain" to that of a gold standard. This computation is described further in the Background section.

edits makes it difficult to formally and generally describe the image transformations they studied. Instead, we study computationally-modeled image transformations, such that we can describe the effects they produce on gaze behavior more formally and generally, and so that we can study a greater scale of data which does not require human decisions for each image.

III. Background

The review of the field of gaze fixation prediction written by Kümmerer et al. [2] is a comprehensive summary of the foundations we build upon. We will describe key terms and concepts in this section.

In the process of studying an image, the human eye will occasionally jump to a new fixation point in what is called a "saccade". Both measured and predicted gaze can be represented by a two-dimensional probability distribution, analagous to a monotone image. The image is sometimes referred to as a "saliency map", but we will invent the term "gaze density" in order to disambiguate from the many different uses of the term "saliency map".

In gaze densities, pixels with the highest intensity are those most likely to be the next fixation target after a saccade for an arbitrary person under "free-viewing" conditions (which means the person has not been instructed to search for an element of the image).

In order the measure how gaze densities diverge, we use the Kullback-Leibler divergence, which measures the entropy between two probability distributions in terms of the number of bits required to describe the difference between those distributions.

We will also use another measure of divergence, called "information gain". Information gain is defined by Kümmerer et al. [5] to be the KL-divergence from a baseline prediction called the "center bias", which is produced by averaging all gaze densities for the dataset of interest. The center bias is the best prediction possible for any selected image from the dataset of interest, if one has no knowledge of the image's contents. It is called the center bias because viewers tend to look towards the center of an image.

We use information gain in our method because it describes the complexity of a gaze density, and when combined with KL-divergence, it can tell us whether the divergence between two gaze densities is due to a loss or gain in complexity, or neither.

The most widely adopted benchmark for gaze prediction model performance is the MIT/Tuebingen Saliency Benchmark [6], which utilizes information gain, KL-divergence, and several other metrics to compare the performance of submitted models. The benchmark also compares models to a gold standard: they leave a random subject out of the real gaze data from the validation set, and produce gaze densities using the remaining data. These gaze densities diverge slightly from the validation data because of the omission of a subject, but they approximate a "best possible" prediction. The current top contender on the benchmark, DeepGaze IIE [1], achieves roughly 80% of the information gain of the gold standard, with the center bias being defined as 0% information gain. The model achieves roughly equivalent performance in other metrics.

We will perform experiments using the DeepGaze IIE model and the UNISAL model [7], which achieves rougly 70% of the information gain of the gold standard (less than DeepGaze IIE), but which has the advantages of running at or near real-time for video and being much smaller in memory. The models studied previously by Che et al. [3] for image transformations achieved roughly 60% of information gain of the gold standard on the MIT/Tuebingen Saliency Benchmark.

The outputs of the particular gaze prediction models we will be experimenting with, as well as the gaze data collected by Che et al., are unnormalized probability distributions. In order to compare divergence between gaze densities, we may normalize to a probability distribution or a log probability distribution. For our experiment, we will be performing our calculations for both of these normalizations in parallel, because they each provide differing information. Compared to a normal probability distribution, a log probability distribution lowers the influence of shared area between two gaze densities when computing divergence, and raises the influence of shared maxima between the two gaze densities. See figure 1.

Fig. 1: In this figure, we see a small section of the same gaze density, normalized to a probability distribution (left) and a log probability distribution (right). The intensity of pixels have been clipped in order to improve visibility: the bright spot on the left contains much higher intensity values than the brightest pixels on the right.

IV. Method

We formalize our method as computing the KL divergence and difference of information gain between two gaze densities produced by the model from two images, an original and a stylized post-processing of the original. We apply the stylization at relative strengths, such that we can study whether a stylization has non-linear effects on the prediction.

We will invent the term "information asymmetry", simply defined as the information gain of one gaze density minus that of another.

So that we can compare the effects of different stylizations, we normalize the KL-divergence and information gain difference by the size of the image and the least-squares difference between the stylization and the original image. The resulting metric tells us the effect size per pixel of the image, per pixel intensity value altered by the stylization. We compute the mean, median, and standard deviation of these normalized metrics across all images for a given stylization.

We use the divergence and information gain difference resulting from random noise as a control group for the comparison of effects of stylizations. If a stylization produces lower metrics than random noise, the stylization has little effect on the prediction produced by the model, and vice versa for higher metrics.

If a stylization produces metrics significantly different from random noise, whether lower or higher, and if an explanation for the effect based on human visual behavior can't be produced, it warrants further study and training for the model. The same can be said for metrics which are not significantly different, contrary to expectation from human visual behavior.

V. Experiment

The benchmark with the widest adoption for gaze fixation prediction models is the MIT/Tuebingen Saliency Benchmark [6], which compiles a list of community-submitted models and analytically computes the fairest saliency map predicted by that model [8] when evaluated for each of a set of metrics [9] on a couple of image datasets [10], [11].

The top contenders on the MIT/Tuebingen saliency benchmark are all deep-learning approaches to the gaze fixation prediction problem [1]. Deep learning models are prone to overfitting to training data.

We use MIT300 and MIT1003 image datasets. *Citation needed.* We apply our selected stylization effects to each of these images, in varying levels from 1 to 10 "strength", where strength is an arbtrary measure for applying a filter with greater pixel difference. Because we recognize that strength is not on a well-defined scale, we normalize our results by the pixel difference as mentioned in the Method section.

We use the UNISAL model for saliency prediction. It is one of the top performing models on the MIT/Tuebingen saliency benchmark. *Citation needed.* (Will try to add more models later.) We had to convert the saliency maps we generated with UNISAL to gaze densities by dividing by the sum of the saliency map, before performing any metric computations.

We focus on post-processing effects that are particularly relevant to our motivating use cases, including games, shows, and virtual environments. Therefore, we gather a few classes of effects from the paper on NPR rendering and the open source image editing software GIMP. *Citation needed.* We gather from the NPR rendering paper that edge-enhancement, color-space adjustment, and frequency-filtering (similar to texture filtering) are important effects. We gather from GIMP that digital distortions are also important stylization effects.

The effects included in edge-enhancement are difference-of-gaussians edge darkening (for two different sizes of gaussian kernels), and the Kuwahara filter. *Citation needed. Should put some visual examples of these effects.*

The effects included in color-space adjustment are color-quantization, hue shift, saturation shift, contrast shift, color inversion, shadow darkening, and vignette. *Should put some visual examples of these effects.*

The effects included in frequency-filtering are gaussian blur, gaussian high-pass, horizontal blur, vertical blur, focus blur, and bloom. *Should put some visual examples of these effects.*

The effects included in digital distortions are pixelation, row shift, screen-door effect, and chromatic aberration. *Should put some visual examples of these effects.*

After applying all stylization effects to all images and producing gaze densities for each, we compute the KL divergence and information gain difference between each pair of original images and stylized counterparts, along with their gaze densities. We normalize these two metrics by the pixel difference between the original and stylized images as mentioned in the Method section.

Finally, after computing the two metrics for each stylization, we aggregate the data by computing the mean, median, and standard deviation of the metrics across all images for a given stylization by dataset. We also compute a general mean, median, and standard deviation across all datasets and all stylizations.

VI. Results

As mentioned in the Method and Experiment sections, we measure two metrics on each stylization: KL-divergence between original and stylized, and information gain difference between original and stylized. We normalize these both by the number of pixels, as well as the total difference between the original and stylized images. From here on, we will call these two metrics "Divergence per Difference" (DPD) and "Information Asymmetry per Difference" (IAPD), respectively.

We aggregate these metrics by filter type, dataset, and "strength level" (as described in the Experiment section). The following graphs will plot curves for each filter type, with the x-axis being the strength level, and the y-axis being the metric (DPD or IAPD). Additionally, we will shade the area one standard deviation above and below the mean for both the Gaussian noise filter (high frequency noise, in blue) and the Perlin noise filter (low frequency noise, in orange), as a reference point.

First, we find that divergence per difference trends upwards as the strength level increases. We normalize by the pixel difference between the original and stylized images, so this shows that the asymptotic behavior of the divergence is greater than linear with respect to the the modification of the image (strength level).

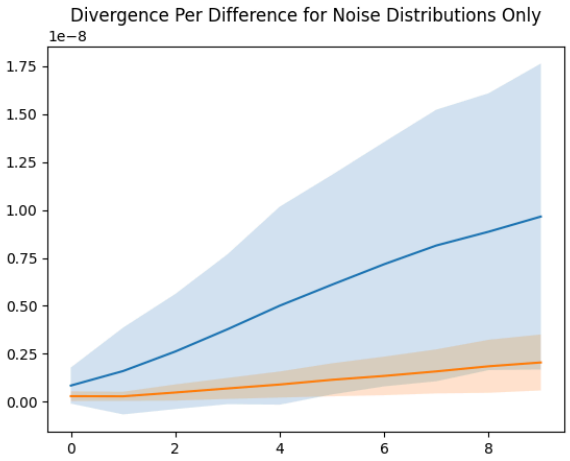

Fig. 2: DPD for the Gaussian noise filter (high frequency noise, in blue) and the Perlin noise filter (low frequency noise, in orange). We see that high frequency noise causes greater divergence of the distribution than low-frequency noise, but both are greater than a constant effect based on the modification of the original image (strength level).

Second, we find that the information asymmetry per difference trends downwards as the strength level increases, for all filter types. Because we are subtracting the original information gain from the stylized information gain to obtain this metric, we have shown that all filters are destructive compared to the original image, and the model cannot extract a more complex prediction from the stylized image than the original.

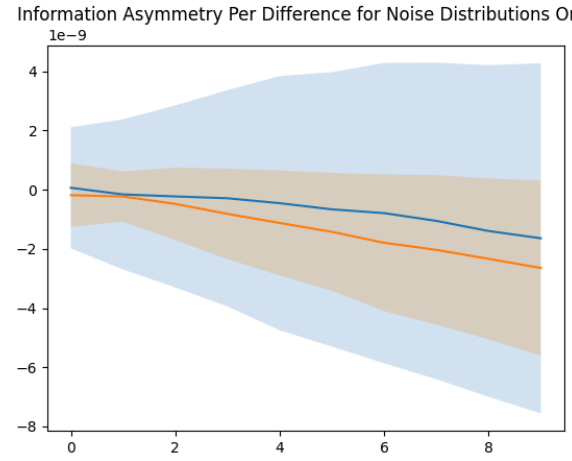

Fig. 3: IAPD for the Gaussian noise filter (high frequency noise, in blue) and the Perlin noise filter (low frequency noise, in orange). We see that both trend downwards, showing that the model cannot extract a more complex prediction from noisy images than the original.

We find that between the datasets measured, similar behavior is observed, across all filter types. Below is an example plotting the Edge-Enhancement filter group.

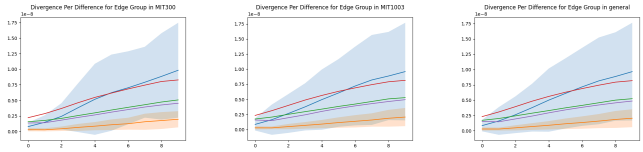

Fig. 4: DPD for the Edge-Enhancement group of filters for MIT300 and MIT1003 datasets, as well as the general average between the two. We see similar behavior across all filter types.

We can see that certain filters are notably more destructive on the prediction than others.

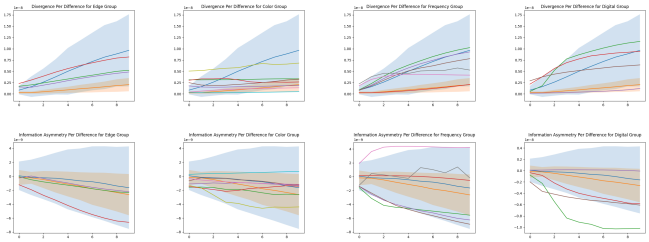

Fig. 5: Plotting averages for all datasets. The top row is DPD, the bottom row is IAPD. The first column is the Edge-Enhancement group, the second column is the Color-Space Adjustment group, the third column is the Frequency Filtering group, and the fourth column is the Digital Distortion group. We see that notable outliers exist in DPD for the Color group, and in IAPD for the Frequency and Digital groups.

These particularly destructive filters come from the Color-Space adjustment group, the Frequency Filtering group, and the Digital Distortion group. *Need to describe the specific filters and their differences in more detail.* Notably, Edge-Enhancement filters remain consistent with the trends observed in noise filters.

We posit that the human visual behavior for color-space, frequency-filtering, and digital distortion effects should have greatest priority for further study so that we can confirm whether our gaze prediction models are accurate under these effects.

Add plots about error tolerance.

VII. Conclusion

We have shown general trends that hold true across datasets and filter types: all stylizations cause a gaze fixation divergence from the original image, and all stylizations cause a decrease in information gain. Given that many effects which do not destroy relative pixel intensity information (such as hue shift) still display this effect, it would either indicate that the gaze prediction model has insufficient training data for these effects (and thus cannot produce complex predictions), or that human visual behavior is not well equiped for these effects and the model follows this trend in human visual behavior.

We have also compared each effect studied, and shown that some effects in particular are destructive to the prediction. These effects should be given highest priority for further study, as they are most likely linked (whether true to human visual behavior or not) to image characteristics which are crucial to the model's prediction.

For other effects that do not fall far outside of the distribution produced by noise filters, we have not found evidence to suggest that they are important to the model's prediction. We may potentially find the risk of poor predictions acceptable for these effects, and begin using them in our applications to predict human visual behavior.

Our work has provided guidance for future human trials, such that they can avoid costly search for significant effects, and instead focus on explanation of significant effects.

References

[1] A. Linardos, M. Kümmerer, O. Press, and M. Bethge, " DeepGaze IIE: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling ." [Online]. Available: https://arxiv.org/abs/2105.12441

[2] M. Kümmerer and M. Bethge, "Predicting Visual Fixations," *Annual Review of Vision Science*, vol. 9, no. Volume9, 2023, pp. 269–291, 2023, doi: https://doi.org/10.1146/annurev-vision-120822-072528.

[3] Z. Che, A. Borji, G. Zhai, X. Min, G. Guo, and P. L. Callet, "How is Gaze Influenced by Image Transformations? Dataset and Model," *IEEE Transactions on Image Processing*, vol. 29, pp. 2287–2300, 2020, doi: 10.1109/tip.2019.2945857.

[4] R. Quian Quiroga and C. Pedreira, "How Do We See Art: An Eye-Tracker Study," *Frontiers in Human Neuroscience*, vol. 5, 2011, doi: 10.3389/fnhum.2011.00098.

[5] M. Kümmerer, T. S. A. Wallis, and M. Bethge, "Information-theoretic model comparison unifies saliency metrics," *Proceedings of the National Academy of Sciences*, vol. 112, no. 52, pp. 16054–16059, 2015, doi: 10.1073/pnas.1510393112.

[6] M. Kümmerer *et al.*, "MIT/Tübingen Saliency Benchmark." Accessed: Mar. 15, 2025. [Online]. Available: https://saliency.tuebingen.ai/

[7] R. Droste, J. Jiao, and J. A. Noble, "Unified Image and Video Saliency Modeling," in *Computer Vision – ECCV 2020*, Springer International Publishing, 2020, pp. 419–435. doi: 10.1007/978-3-030-58558-7_25.

[8] M. Kümmerer, T. S. A. Wallis, and M. Bethge, " Saliency Benchmarking Made Easy: Separating Models, Maps and Metrics ," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., in Lecture Notes in Computer Science. Springer International Publishing, 2018, pp. 798–814.

[9] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?," *arXiv preprint arXiv:1604.03605*, 2016.

[10] T. Judd, F. Durand, and A. Torralba, " A Benchmark of Computational Models of Saliency to Predict Human Fixations ," in *MIT Technical Report*, 2012.

[11] A. Borji and L. Itti, " CAT2000: A Large Scale Fixation Dataset for Boosting Saliency Research ," 2015, [Online]. Available: https://arxiv.org/abs/1505.03581