GAZE FIXATION PREDICTION FOR COMPOSABLE IMAGE TRANSFORMATIONS

by
James Youngblood

A thesis submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Master of Science

School of Computing
The University of Utah
August 2025

The University of Utah Graduate School

STATEMENT OF THESIS APPROVAL

The thesis of *James Youngblood* has been approved by the following supervisory committee members:

Rogelio Cardona-Rivera,	Chair,	*(Date Approved)*
Paul Rosen,	Member,	*(Date Approved)*
Cem Yuksel,	Member,	*(Date Approved)*

by *Mary Hall*, the Chair of the School of Computing,
and by *Darryl P. Butt*, the Dean of the Graduate School.

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.

CONTENTS

LIST OF FIGURES

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 and Matthias Bethge [2] show that all leading gaze prediction models utilize transfer learning from other problem domains, primarily object recognition. Researchers do this becase of the lack of training data for the gaze recognition task, relative to the object recognition task. 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. This assumption is implicitly encoded into most of the training data a gaze prediction model will see, and so the model's performance may suffer when generalizing to stylized images.

[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.

Furthermore, a study by Zhaohui Che, Ali Borji, Guangtao Zhai, Xiongkuo Min, Guodong Guo and Patrick Le Callet [3] shows that image transformations influence gaze fixations, and effects of transformations on gaze behavior may be difficult to model. This fact is problematic, because modern visual productions use multiple steps of image processing and rendering for stylistic effect. Each step represents an image transformation, and a possible degradation of gaze prediction performance. Demonstrating robust performance, 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 recognize three classifications for common image transformations, and we contribute simple heuristic functions for each class of transformation. These heuristics are meant to be statistically significant signals for how well a model's prediction adheres to a real gaze distribution. Notably, computing these heuristics does not require the collection of real gaze distributions, making them cheap to run. Our analysis of the heuristics utilizes real gaze distribution data on images before and after transformation provided by Che et al.

While contributing these heuristics, we establish a loose mathematical definition for prediction adherence, which affords us the ability to determine whether a transformation is "composable". We argue that composable transformations can be applied to an image with lesser concern for gaze prediction degradation.

II. RELATED WORK

The effects of image transformations–including cropping, rotation, skewing, and edge detection–on gaze behavior are studied by 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 Rodrigo Quian Quiroga and Carlos Pedreira [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 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 foundations we build upon are reviewed comprehensively by Kümmerer et al. [2]. We will briefly describe key terms and concepts.

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 Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge [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 gaze densities. 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 complementary information. Compared to a normal probability distribution, a log probability distribution lowers the influence of shared or differing area between two gaze densities when computing divergence, and raises the influence of shared or differing global maxima. See figure 1.

Figure 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 right contains much higher intensity values than the brightest pixels on the left.

IV. METHOD

We wish to produce a method which shows whether a predicted gaze density adheres to the real gaze density when considering an image that is transformed with one of three classes of transformation.

First, let us define some terms. We will refer to the gaze density of an image before a transformation as the “prior density”, and the gaze density of the image after a transformation as the “subsequent density”. We will refer to KL-divergence as “divergence” for brevity. We define “adherence” to be a condition for a transformation, such that the divergence of a prior density to its appropriate real gaze density is greater than or equal (to within 5% of the divergence value) to the corresponding subsequent density to its real gaze density, for 95% of images tested under the transformation.

We call such a transformation a “composable” transformation. Assuming a transformation has been tested for adherence over a large number and large variety of images, we can be confident that the gaze prediction model understands the transformation and its effects on gaze behavior. Composable transformations may be applied freely to an image without concern for degrading the model’s performance.

TODO:

- Obviously, if divergence is low, then it stands to reason that greater subsequent adherence is likely.
- If divergence is higher, but information asymmetry is low, then greater subsequent adherence might be the case: we should study.
- Check divergence, information asymmetry, and a product of both metrics for binary classification and for error bound/trend
- Figure out where to set confidence bounds for adherence

We wish to provide evidence that two metrics, computed on the prior density and the subsequent density, are statistically significant heuristics for whether the adherence of the subsequent density is roughly equal or greater than the adherence of the prior density to its corresponding real gaze density.

We will argue that the condition of greater subsequent adherence implies that the transformation does not degrade the performance of a gaze prediction model. We will refer to this condition as “composability”. Transformations that have been proven composable for a gaze prediction model can be applied to images freely, without the need for consideration of the effects on the model’s performance.

The first metric is KL-divergence, which will referred to henceforth simply as “divergence”. The second is the difference in information gain between the two distributions. We we will invent the term “information asymmetry” to refer to this difference in information gain.

We apply the stylization at relative strengths, such that we can study whether a stylization has non-linear effects on the prediction.

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.

TODO:

- add problem statement to the end of background
- add the paper references to MIT benchmark

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: 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.