Gaze Fixation Prediction for Stylized Images

James Youngblood
james@youngbloods.org
University of Utah

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

2025-03-06

Abstract

We study the effects of stylistic post-processing image filters on gaze fixation prediction.

Keywords gaze fixation prediction · image filters · stylistic post-processing

1 Introduction

We are trying to apply gaze fixation prediction to games, shows, interactive applications, and virtual environments in general. This will allow designers to have a better understanding of how users interact with their work, as well as what elements grab users' attention. This can allow them to iterate and improve their work quickly, without as much human testing.

Gaze fixation prediction has been shown to generalize well to a few domains, including still life images, movies, and TV. *Citation needed.* It may be reasonable to assume that it will generalize to photo-realistic environments and games as well. *Citation desired.* There is concern, however, that gaze fixation prediction models may not generalize to stylistic renders as well, because of the lack of training data.

Add section about the difference between image-space post-processing stylization and world-space geometry stylization.

Gathering more training data is a difficult task, due to eye-tracking equipment and the time required from many human subjects. Further, the domain of stylized images is large and varied, and training without prior knowledge of potential weaknesses of the model will be time-consuming.

We contribute an experiment which investigates the effects of stylistic post-processing on images on the gaze fixation prediction of models. Our experiment attempts to isolate those effects which perturb the prediction compared to the non-stylized image the most. Using perceptual theory, we speculate on the reasons for why these effects might be so destructive to the existing prediction, and posit that these effects should recieve highest priority for additional training data.

Our experiment can also be used as a model for discovery of lacking training data for other image effects which we have not yet studied.

2 Related Work

The models used in this experiment, and their previous training data. *Citation needed.* In contrast to our method, the training data used for the original methods use traditional photography and video shoots, which are not very representative of the range of styles in illustration, animation, or computer graphics.

The study on where people look at for artwork in museums. *Citation needed.* In contrast to our method, they have gathered a set of human eye-tracking data over oil paintings in museums. This can be used as training data for gaze fixation prediction models, and potentially address some pitfalls of current models, but does not cover the breadth of styles we want to study.

3 Background

Annual review of vision prediction. *Citation needed.* This is a review of the field of gaze fixation prediction, and it provides the theoretical foundation from which our experiment builds. The following terms and operations are as defined in the review. *Also include the original papers these techniques were published in.*

We define the term gaze densities: the probability distribution of gaze fixations over an image, as opposed to a saliency map which has an underdefined scale. We can computer the gaze density from a saliency map by dividing by the sum of the saliency map. There exists a "baseline" gaze density that is not a function of the image, which predicts human visual behavior as best possible without reading the image data. It is called the "center bias", because humans will typically look to the center of an image more than the edges.

We define two operations on gaze densities: KL-divergence, which measures the difference between two probability distributions in terms of the number of bits required to describe the difference between them (and maybe the Jensen-Shannon divergence, which is a symmetrized and smoothed version of the KL divergence). We also describe the information gain, which is the KL-divergence from the center bias, which describes the complexity of the prediction, or the amount of information that the model was able to extract from the image.

Visual perception theory. *Citation desired.* (Debating on whether to include this.)

4 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.

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.

5 Experiment

We use MIT300 and MIT1003 image datasets. *Citation needed.* (Will try to add more datasets later.) 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.

6 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).

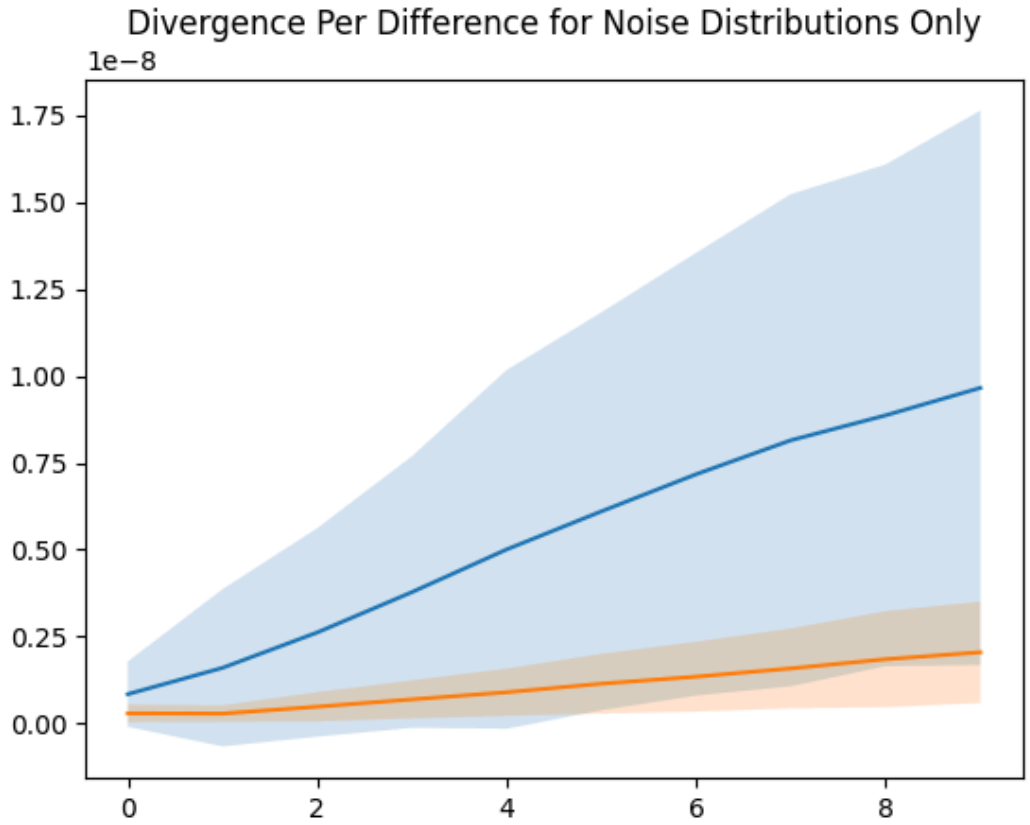

Figure 1: DPD for the Gaussian noise filter (high frequency noise, in blue) and the Perlin noise filter (low frequency noise, in orange). We see that high frequency noise causes greater divergence of the distribution than low-frequency noise, but both are greater than a constant effect based on the modification 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.

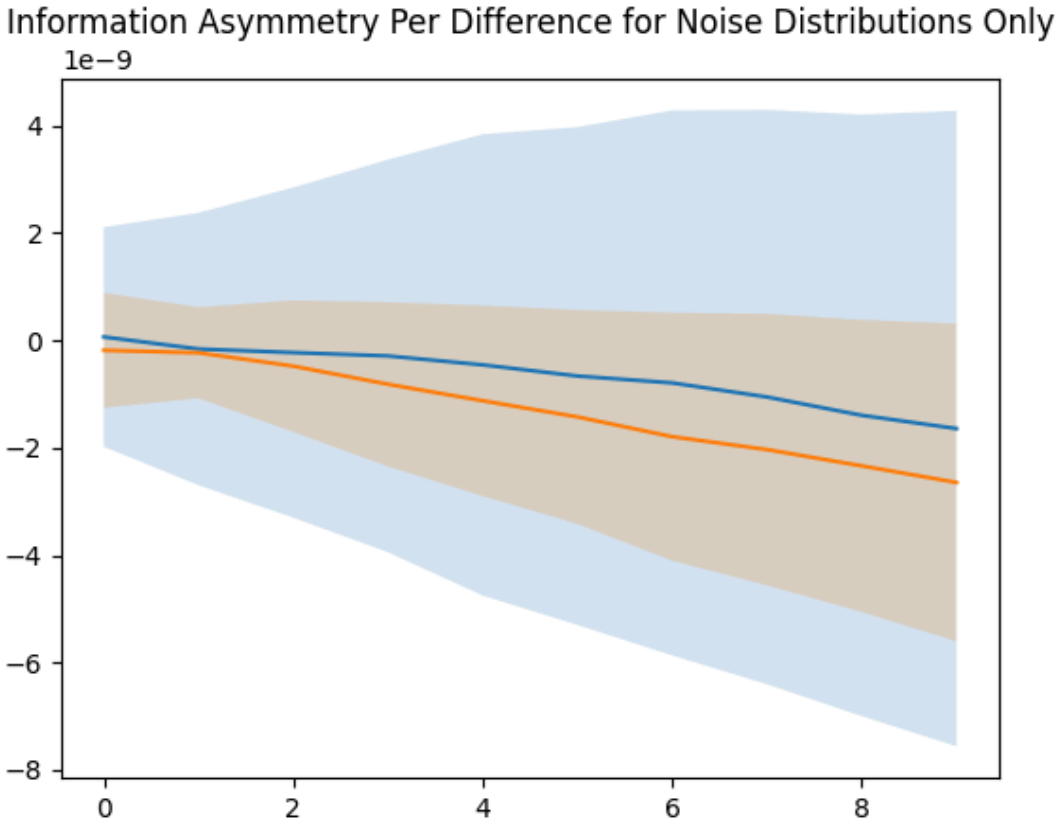

Figure 2: IAPD for the Gaussian noise filter (high frequency noise, in blue) and the Perlin noise filter (low frequency noise, in orange). We see that both trend downwards, showing that the model cannot extract a more complex prediction from noisy images than the original.

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.

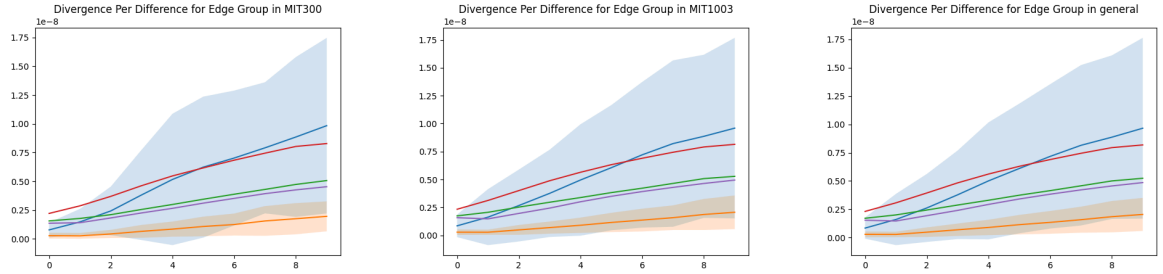

Figure 3: DPD for the Edge-Enhancement group of filters for MIT300 and MIT1003 datasets, as well as the general average between the two. We see similar behavior across all filter types.

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

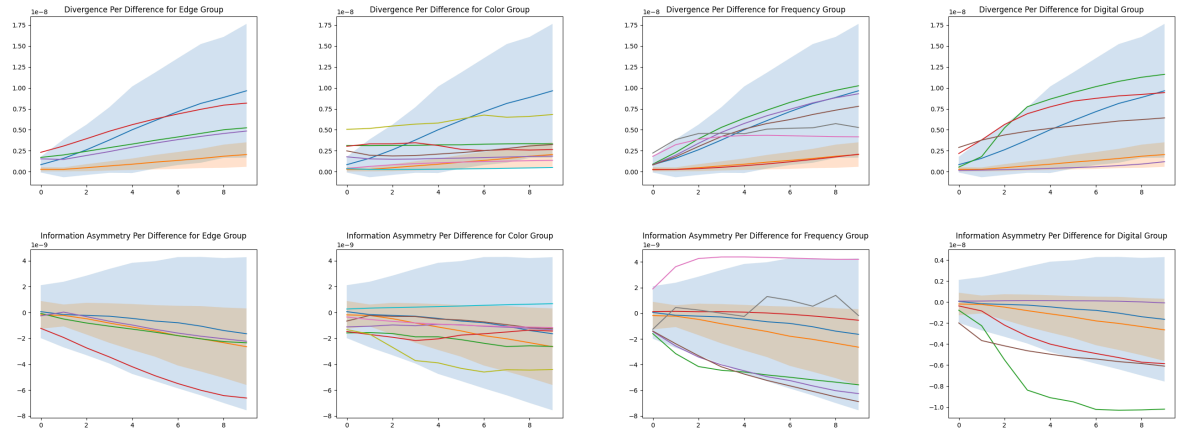

Figure 4: Plotting averages for all datasets. The top row is DPD, the bottom row is IAPD. The first column is the Edge-Enhancement group, the second column is the Color-Space Adjustment group, the third column is the Frequency Filtering group, and the fourth column is the Digital Distortion group. We 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.

7 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.