Spatially Binned ROC: A Comprehensive Saliency Metric

Calden Wloka, John Tstotsos
Electrical Engineering and Computer Science Department
York University, Toronto
calden@cse.yorku.ca, tsotsos@cse.yorku.ca

Abstract

A recent trend in saliency algorithm development is large-scale benchmarking and algorithm ranking with ground truth provided by datasets of human fixations. In order to accommodate the strong bias humans have toward central fixations, it is common to replace traditional ROC metrics with a shuffled ROC metric which uses randomly sampled fixations from other images in the database as the negative set. However, the shuffled ROC introduces a number of problematic elements, including a fundamental assumption that it is possible to separate visual salience and image spatial arrangement.

We argue that it is more informative to directly measure the effect of spatial bias on algorithm performance rather than try to correct for it. To capture and quantify these known sources of bias, we propose a novel metric for measuring saliency algorithm performance: the spatially binned ROC (spROC). This metric provides direct insight into the spatial biases of a saliency algorithm without sacrificing the intuitive raw performance evaluation of traditional ROC measurements. By quantitatively measuring the bias in saliency algorithms, researchers will be better equipped to select and optimize the most appropriate algorithm for a given task. We use a baseline measure of inherent algorithm bias to show that Adaptive Whitening Saliency (AWS) [14], Attention by Information Maximization (AIM) [8], and Dynamic Visual Attention (DVA) [20] provide the least spatially biased results, suiting them for tasks in which there is no information about the underlying spatial bias of the stimuli, whereas algorithms such as Graph Based Visual Saliency (GBVS) [18] and Context-Aware Saliency (CAS) [15] have a significant inherent central bias.

1. Introduction

Saliency map algorithms are a popular class of algorithm originally designed to provide bottom-up attentional gating based on Koch and Ullman's architecture for attention selection [27], and heavily influenced by Treisman and Gelade's *Feature Integration Theory* [46]. One of the earliest and most popular saliency map models, referred to here as IKN, was developed by Itti *et al*. [21]. Since then an enormous variety of saliency map models have been developed and refined; the unifying feature of these disparate algorithms is the assignment of a *conspicuity value* to every location within a visual scene. A visual element which has a higher conspicuity value is something which can be considered interesting or important, and indicates a visual location which is worthy of allocating further processing resources. Examples of subsequently developed saliency algorithms include those based on information theory and sparse coding [8, 20], Bayesian reasoning over learned features [52], graph-based approaches [18], spectral analysis [19, 42], and machine learning techniques combined with pre-chosen object detectors (such as face detection) to create a salience classifier [25].

In addition to a rapidly expanding set of approaches, the concept of saliency has grown beyond just attentional gating and has been applied to a number of additional areas. The broadest category of models are largely still focused on understanding how humans allocate fixations when free-viewing scenes (*e.g*. [8, 18]). More recently, some algorithms forgo any modeling of the underlying computational structure of overt human attention and instead focus solely on predicting where in an image humans will fixate with the greatest possible accuracy (*e.g*. [25, 22]). Although potentially less informative to neuroscientists and psychologists interested in attentional eye movements, the focus on performance is motivated by potential commercial applications such as fixation-guided heterogeneous image compression [17]. Additionally, a third avenue of saliency research seeks to develop a system useful for prioritizing attentional resources (irrespective of human performance) for tasks such as mobile robot navigation [39, 10] and robotic visual search [37].

A continuing challenge in saliency modeling is the formulation of fair and informative metrics with which to evaluate and compare different saliency algorithms. Over the

years a number of metrics have been adapted from signal analysis or developed for measuring saliency performance. Several of the most common include Normalized-Scanpath Salience [36], Earth-Mover's Distance [40], Kullback-Liebler (KL) Divergence [28], and Receiver Operating Characteristic (ROC) curves [16]. Several recent benchmarking studies have provided summaries of these metrics and their role in evaluating saliency algorithms [4, 38]. Both Borji *et al*. [4] and Riche *et al*. [38] conclude that a robust evaluation of model performance is best obtained by combining complementary metrics, with the ROC class of metrics as a frequent focal point of algorithm analysis. Nevertheless, the central bias in human fixations remains a persistent issue in human fixation-based metrics, and in ROC metrics in particular. Both previously mentioned benchmarking efforts seek to correct for this spatial bias by advocating for the use of the *shuffled* area under the ROC curve (*s*AUC). However, in Section 2 we argue that *s*AUC fails to satisfactorily correct for a central bias, and further that what is referred to as center bias is better understood as an intrinsic aspect of active foveal vision.

Rather than attempting to separate the visual content of fixated locations from its spatial context, we propose in Section 3 a novel evaluation metric. This metric analyzes saliency algorithm prediction of human fixations within the context of their spatial distribution over the dataset. We do this by spatially binning the ground-truth fixation points and then deriving an ROC curve for each bin independently. The main contributions of our work are: First, we demonstrate that the current *s*AUC metric is problematic and may not provide the information implicitly assumed by its users (Section 2). Second, we provide an alternative metric which allows the explicit detection of algorithmic spatial bias while still providing the direct predictive power of traditional ROC methods (Section 3). Examples of metric application and discussion of its use are presented in Section 4.

2. Center Bias

2.1. Center Bias and Shuffled ROC

Early in the study of saliency it was noticed that stimulus location had a strong effect on the likelihood of fixation, with regions closer to the image center being more commonly fixated than those near the image edge [34]. This topic was revisited by Zhang *et al*. [52], who discussed in detail the confounding effect center bias can have on saliency algorithm evaluation. As they pointed out, a saliency map consisting solely of a centered Gaussian (the cG model) outperforms many of the leading saliency models in predicting human fixations despite being independent of the actual image content. Likewise, particularly given the small image sizes being tested, differences in the thickness of the border region left undefined by filter convolution had a tendency to reward models with a greater undefined border due to a concentration of saliency values toward the image center. While acknowledging that *photographer bias* (the tendency to center pictures on interesting objects) might mean that the image centers are genuinely more likely to be salient than peripheral locations, they nevertheless advocated the use of shuffled metrics based on the work of Parkhurst and Neibur [35] and Tatler *et al*. [44] to rectify these two issues. Despite the fact that shuffling may reduce the raw numerical performance measured for each algorithm, they argue that the *relative* performance of algorithms should be unaffected, and thus shuffled metrics provide a fairer assessment. Although eliminating the effect of differing boundary region sizes could arguably have been accomplished in an alternative fashion by simply enlarging the undefined border (zeroing all saliency values) of all models to an equivalent size (as was done in [29]), such an approach would not penalize static maps (*e.g*. the cG model) which are independent of the underlying image.

Of course, while it is perhaps disappointing to have a static Gaussian center prior outperform one's algorithm in predicting fixation locations using traditional metrics, this does not necessarily mean that such metrics are wrong. Much of the debate over metrics seems to rest with an unclear definition of their goals [9]. If the motivation of a model is in producing the best possible predictor of human fixation locations in an image (*e.g*. for use in image compression), then it does not particularly matter whether a correct pixel label is based on a positional prior or the visual content of the image. This approach is exemplified in the benchmarking work of Judd *et al*. [23], whose saliency model is based on a machine learning classifier trained to label pixels in saliency space regardless of biological plausibility in the calculation [25]. As their focus is on producing the best prediction of human gaze location for applications in areas like human-computer interaction, they use a classical ROC metric and optimize a central prior and post-processing smoothing kernel for every algorithm. The argument follows that, since every algorithm has had these parameters optimized, the test is made fair. By contrast, shuffled metrics which penalize static contributions to fixation prediction (and which have dominated most recent benchmarking studies, *e.g*. [3, 38]) seek to rate a saliency algorithm's predictive ability solely on its interpretation of visual stimuli in isolation from any confounding factors introduced by spatial position.

In both the shuffled and classical ROC, the true positive rate is the percentage of human fixation points which are above a saliency threshold. However, whereas the false positive rate in the classical ROC is taken as the proportion of total image pixels which are above threshold (the proportion of non-fixated locations which are marked as salient), in the

shuffled ROC the false positive rate is calculated based on the number of fixation points, randomly sampled from other images in the same data set, which are above threshold. In this way regions of the image towards which viewers are spatially biased will more likely yield randomly sampled fixations when forming the false positive set, negating the benefits of a spatial bias prior.

However, the shuffled ROC also makes a fundamental assumption that it is actually possible to isolate the intrinsic salience of visual stimuli from its spatial context. We posit that this assumption is not valid, and that completely separating the visual and spatial properties of stimuli when seeking to predict human fixations is not possible (see Section 2.2). Furthermore, Bruce *et al*. [7] have recently shown that, through the discounting of centrally predicted fixations, shuffled metrics end up favoring algorithms with peripherally biased raw scores.

One final aspect to note regarding the *s*AUC metric is the lack of a clear physical interpretation of *s*AUC score. In classical ROC methods, the performance curve can be understood as a direct measure of the likelihood of successfully predicting a human fixation point at a given cutoff threshold. The ROC curve generated when calculating *s*AUC, however, does not explicitly notify the user how many fixations were discounted as false positives by overlap with the shuffled set. While this does not affect the utlity of *s*AUC in a relative comparison of algorithm performance, it does make it difficult to interpret the actual meaning of the numerical results.

2.2. The Persistence of Center Bias

While some center bias may be created by photographer bias toward centering objects of interest in a frame, this should have very little effect on algorithm performance in classical metrics once image border effects are controlled for (if the most interesting visual stimuli consistently appear near the image center a high performing saliency algorithm should likewise consistently detect the image center as most salient). However, we argue that compositional bias is not the only source of center bias, but rather that there exists an inherent central bias to eye movements which is independent of the stimuli. In fact, [45] have previously demonstrated the robustness of the underlying fixation biases inherent to human gaze patterns, showing that a model based on oculomotor patterns of movement (independent of the image itself) was more predictive of human gaze data than the IKN saliency model. Here, we concentrate specifically on the central bias aspect of human gaze, using eye tracking data from two different data sets: the **D**atabase **O**f **V**isual **E**ye Movement**S** (DOVES) produced by [31], and the MIT dataset of human eye-tracking produced by [25].

It is important to note that eye fixations in both the MIT and DOVES datasets were captured during free-viewing. It has long been established that task can have a profound effect on fixation patterns; this was first suggested by the seminal work of Yarbus [51] and recently explored more systematically by Borji and Itti [2]. Although some saliency work has attempted to incorporate task bias [12, 26], the majority of saliency modeling is nevertheless done under the assumption of free-viewing. For a set of recently developed task-controlled eye-tracking datasets see [32, 49], as well as [33] which characterizes the spatiotemporal ordering of human fixations under two different tasks. Given that the present work is specific to free-viewing scenarios, further discussion or comparison with datasets based on task bias would be inappropriate.

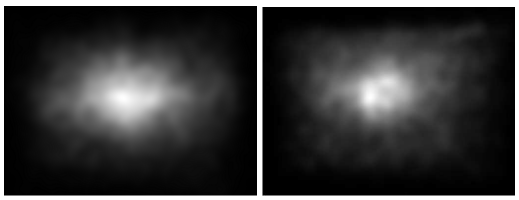
Statistical properties of the free-viewing fixation patterns for human observers of these datasets are presented in Table 1. All values have been normalized with respect to the image dimensions, and therefore, although the proportional variance of the DOVES fixations is nearly identical in both the x- and y-directions (0.14 and 0.13, respectively), the fixations along the x-direction actually do have a greater spread in terms of raw pixel distances. The MIT dataset, available at [24], is composed of 1003 images sampled from Flickr creative commons and LabelMe [41] with eye tracking data for fifteen observers. Although it is never possible to have a completely representative dataset of images, the MIT set provides a decent attempt to capture a cross-section of the types of photographs people take and share with others (*e.g*. Figure 1). Human fixations over this dataset are strongly biased toward the image center; at least a portion of this bias likely arises due to photographic composition. In order to compile distribution statistics shown in Table 1 for the MIT dataset, which includes images of different dimensions, we limited those included in our analysis to only those 463 which were 1024×768 pixels (landscape) and 123 which were 768×1024 pixels (portrait) in size (the most common sizes in the set).

The DOVES dataset, available at [30], consists of 101 grayscale images cropped from the dataset originally created by [48]. All images in the DOVES set are of landscape orientation with dimensions 1024×768 pixels. In contrast to the MIT dataset, the DOVES dataset provides a strong attempt to mitigate any bias inherent in photographic composition; most images in the dataset have no clearly framed central object or creature (*e.g*. Figure 2). Despite this lack of compositional bias in the image stimuli, aggregate fixation statistics over the dataset shown in Table 1 display that human fixations remain distinctly biased toward the image center (albeit to a lesser extent than in the MIT dataset). Given the lack of strong central objects, this centrally biased distribution pattern most likely corresponds to factors independent of the visual qualities of the stimulus.

We can formulate a spatial prior for eye fixations in the following manner: At the most basic level of abstraction we

Figure 1: Typical image in the MIT dataset [25]. As with many of the images, there is a strong central subject with little peripheral content.

Figure 2: Typical image in the DOVES dataset [31]. Most images have no central object of interest.



(a) MIT Landscape (b) DOVES

Figure 3: Fixation cloud images formed by smoothing over all human fixations in the dataset. On the left is shown the fixation cloud for landscape-oriented images in the MIT dataset, and on the right the fixation cloud for the DOVES dataset.

consider eye fixation data over a visual field as a sequence of points constrained to the 2D plane of the image. Without any knowledge of the underlying visual stimulus (given that we are formulating a prior), an initial best guess for a fixation will be a drawn from a random distribution $p(x, y)$, where p is the probability distribution and (x, y) are the current pixel coordinates of gaze. Each subsequent fixation is dependent only on the previous location in the chain (for now ignoring, for the sake of simplicity, inhibition of return), and thus the t-th fixation takes the form

		Mean	Variance
DOVES Database	x	0.00	0.14
	y	0.00	0.13
MIT Database Portrait	x	0.00	0.027
	y	0.00	0.040
MIT Database Landscape	x	0.00	0.035
	y	0.00	0.028

Table 1: Distribution statistics over all human fixations collected in psychophysical eye-tracking datasets. Values are normalized with respect to image dimensions, and show that fixations consistently cluster around the center of the image (0 mean) rather than off-center, but range in variance. Thus, while degree of bias varies, the bias itself remains consistent.

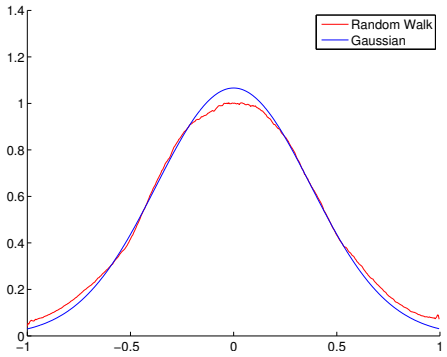


Figure 4: A comparison of the approximate distribution curve for fixations produced by a random walk plotted against a Gaussian of identical variance.

$$(x_t, y_t) \sim p(x_{t-1}, y_{t-1}) \tag{1}$$

which is the definition of a random walk.

Each specific image in a dataset corresponds to a single independent sampling of the random walk. As one would expect by the Central Limit Theorem, it can be shown that the distribution of the point conglomerate produced by this process will tend toward that of a Gaussian distribution [5]. Empirically, we demonstrate this in one dimension by generating random walk trials with sequences of five fixations over a uniform subinterval of the normalized domain $[-1, 1]$. The approximate distribution for this fixation set is formed from the smoothed histogram of the fixation locations. Figure 4 shows how after 1000 trials this approximate probability distribution very closely matches a Normal distribution of identical variance.

Thus, we see that the Gaussian central prior, which is prevalent in improving saliency model scores with traditional ROC metrics, can be derived by a simple translating saccadic model [50]. Additional efforts to model the

dynamic process of saccadic eye movements with a random walk includes both Brockman and Geisel's [6] and Boccignone and Ferraro's [1] work showing that saccadic movements can be well captured as stochastic sequences over a saliency field which correspond well to Levy flight random walks. Therefore, we suggest that rather than a confounding artifact which must be corrected for, the center bias of human fixations can be seen to derive, at least in part, from the mechanics of how people look. Likewise, the improvement in fixation prediction seen by the addition of a Gaussian center prior is due to the fact that a Gaussian functions as a first-order approximation to the actual spatial biases which are introduced through active gaze mechanics. As a result, a model of saliency should inherently account for these effects rather than view their manifestation as a nuisance which must be separately corrected for.

Nevertheless, we still seek a fair method of evaluating human fixation prediction for algorithms with varying degrees of spatial bias representation, and would like this metric to represent algorithm performance across the entire image rather than have the measure of performance be overwhelmed by the central signal. Our solution is to construct a spatially binned ROC (spROC) metric, presented in Section 3.

3. Spatially Binned ROC

The spROC metric seeks to preserve a useful degree of spatial information while still yielding a clear evaluation of saliency algorithm performance. The metric is constructed in the following manner:

1. Partition the image into a set of non-overlapping spatial regions (bins). Each bin is an annulus (except the central bin, which is an ellipse, and the final outer bin) centered on the image center. Because of the tendency for human fixations to vary in proportion to the height and width of the image, bin dimensions are determined by the aspect ratio of the image (see Figure 5 for examples)

2. For a given image, determine into which bin each ground-truth human fixation falls

3. Calculate a traditional ROC curve for each bin

The selection of the most appropriate size and number of the spatial bins may be application specific. We elected to use ten bins and allocate the bins such that each bin had an equal portion of the total set of human fixations (see Figure 5). To ease comparison among methods, such a configuration might be considered as the 'standard' one. However, it is possible for some specific applications that one may wish to investigate the performance of an algorithm according to an alternative distribution of bins which is independent of fixational set, such as one which is determined by relative image area.

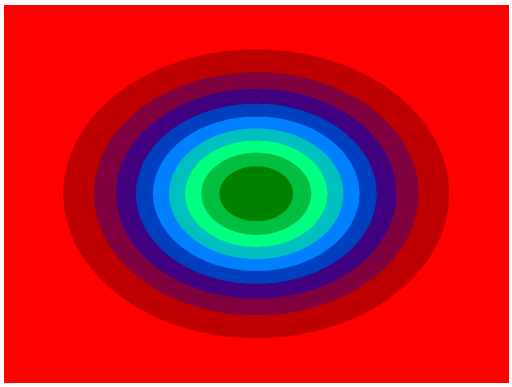


Figure 5: Example showing bins distributed on a 4:3 aspect ratio image proportional to the number of fixations from the MIT dataset falling into each bin. Each band of color represents a different spatial bin.

One of the advantages of the spROC method is that algorithm performance can be analyzed at a number of levels. The traditional ROC curve can be straightforwardly calculated by taking the weighted sum of the individual spatial bins according to the equation:

$$PR_j = \sum_{i=1}^{n} c_i PR_{ij} \tag{2}$$

where PR_j is the positive rate at threshold j, c_i is the count of fixations falling into bin i, and PR_{ij} is the positive rate in the ith bin at threshold j. Likewise, the traditional AUC score can be calculated by finding the area under this curve. When using a proportional distribution of bins Equation 2 simplifies to the average across all bins.

Alternatively, however, one can also examine a spatial profile of the algorithm performance by plotting the AUC score for each individual spatial bin (see Figure 6). Algorithms with a spatial bias will exhibit deviations from a horizontal line, and the degree of deviation can be used to quantify the extent of bias. An unbiased algorithm will form a flat line (every bin will have the same AUC score), while a well-performing algorithm will have the best combined score across all bins. It depends on the application which is more important; although a highly biased algorithm might end up giving the best overall score, the spatial bias exhibited suggests that at least part of its performance is based on an overemphasis (either implicitly or explicitly) on the spatial tendencies of human fixations (the ability to predict less frequent peripheral fixation is sacrificed to improve the chances of predicting central fixations). Adjustment or 'correction' for the center bias of human fixations can be performed through a re-weighting of the ROC points or AUC

score between the bins. This will have an effect similar in outcome to shuffled ROC, but with the added transparency of knowing precisely how fixations have been re-weighted rather than relying on a hidden stochastic process. An example of this type of analysis is shown in the comparison of Tables 2 and 3, where Table 2 shows results using classical AUC, and Table 3 displays instead AUC scores weighted by the image area covered by each bin.

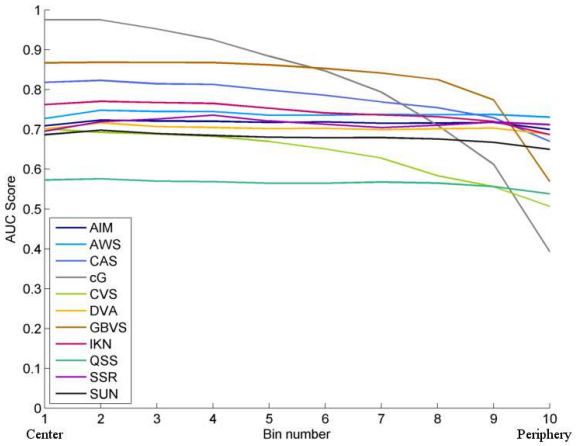
4. Results

Here we present the quantitative clarity achieved by using spROC for a selection of algorithms which have publicly available MATLAB code. All algorithms have been run without the application of post-processing smoothing (also referred to as blurring). Although smoothing is a standard practice and is well-known to have a strong effect on the performance of an algorithm's fixation prediction, convolution will introduce an additional bias against peripheral saliency values proportional to the size of the Gaussian kernel used to perform the smoothing. Since algorithms will frequently exhibit different optimal sizes of smoothing kernel (*e.g.* see [23]), we felt it was useful to look at the inherent degrees of algorithm spatial bias which exists prior to applying any post-processing smoothing. Note, however, that while post-processing smoothing was removed, some algorithms still implicitly smooth their output through image resizing. This step is required for efficient processing speed (*e.g.* GBVS) and thus was retained, but does generally lead to improved scores for these algorithms versus those which have no built-in smoothing. Therefore, it is important to reiterate that the scores presented here are not an optimized benchmark (as in [23, 4, 38]), but rather serve as a baseline characterization of the inherent spatial bias for each algorithm.

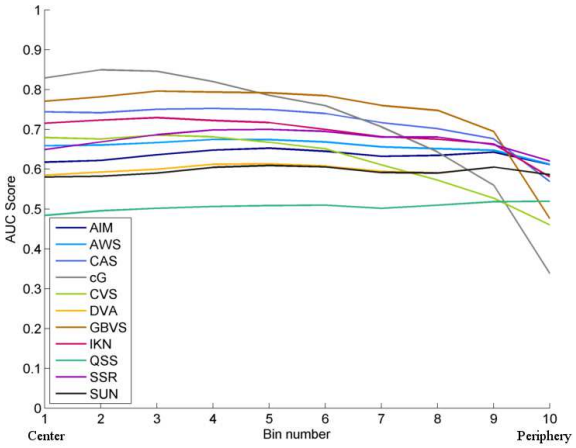
We demonstrate the spROC metric using the following algorithms:

- Attention by Information Maximization (AIM) [8]
- Adaptive Whitening Saliency (AWS) [14]
- Context Aware Saliency (CAS) [15]
- A centered Gaussian prior (cG)
- Covariance-based Saliency (CVS) [11]
- Dynamic Visual Attention (DVA) [20]
- Graph-Based Visual Saliency (GBVS) [18]
- The Itti-Koch-Niebur Saliency Model (IKN) [21]
- Quaternion-Based Spectral Saliency (QSS) [42]
- Saliency Detection by Self-Resemblance (SSR) [43]
- Saliency Using Natural statistics (SUN) [52]

which we ran on two widely used benchmarking datasets: the MIT dataset already discussed in Section 2.2, and the ImgSal dataset [29], which was the basis of the benchmarking work by Riche *et al*. [38]. Note that we used the implementation of CAS created by Tsai and Chang [47] to ensure control over post-processing, as the original study authors released only a binary implementation.



(a) MIT

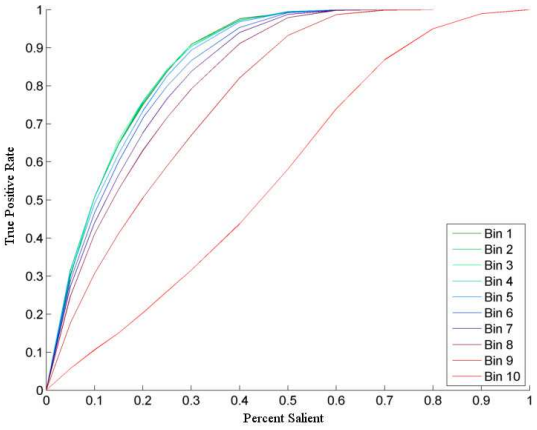


(b) ImgSal

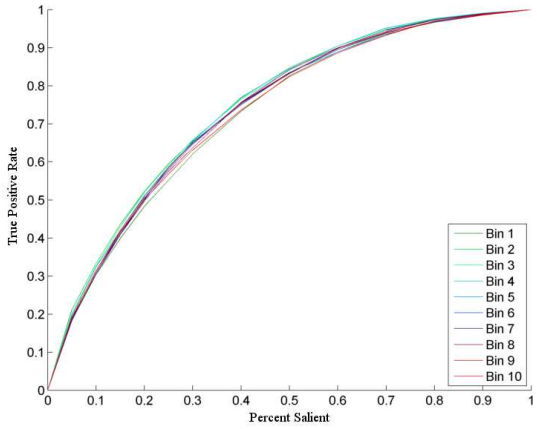
Figure 6: AUC Scores by bin number for a selection of algorithms. (a) presents results over the MIT dataset, and (b) presents results over the ImgSal dataset. All algorithm saliency maps were unsmoothed

As expected, the most extreme spatial bias is exhibited by the cG model (this is, after all, a prediction based solely on a spatial location), with an AUC very close to 1 for the central bins which then rapidly falls off to nearly zero in the more peripheral bins. Of the models tested, GBVS exhibits the strongest degree of spatial bias. Surprisingly, although identified in [7] to have a peripheral bias in terms of raw saliency scores, in terms of predictive performance AWS is actually one of the least biased models.

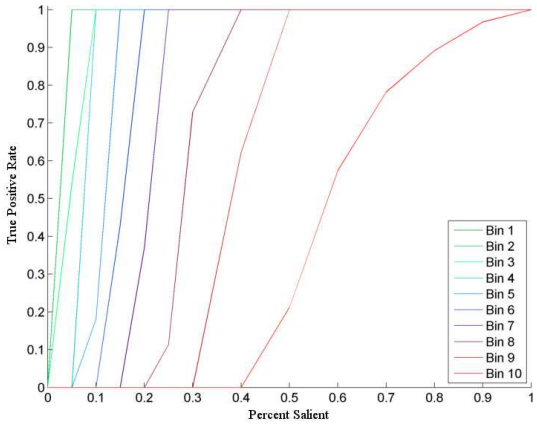
Figure 7 shows the bin by bin ROC curves for GBVS (7a), AWS (7b), and a Gaussian center prior (7c) for the MIT dataset. These figures show a more detailed view of the nature of the spatial bias in these various models, and these



(a) GBVS



(b) AWS



(c) cG

Figure 7: ROC scores by bin for GBVS, AWS, and a Gaussian center prior on the MIT dataset. GBVS is the most spatially biased model tested, while AWS represents the most spatially consistent model tested

MIT			ImgSal		
Model	AUC	σ	Model	AUC	σ
GBVS	0.82	0.093	GBVS	0.74	0.098
cG	0.81	0.188	CAS	0.71	0.057
CAS	0.78	0.049	cG	0.71	0.163
IKN	0.74	0.026	IKN	0.69	0.045
AWS	0.74	0.007	SSR	0.67	0.025
AIM	0.72	0.007	AWS	0.66	0.018
SSR	0.72	0.013	AIM	0.63	0.014
DVA	0.70	0.007	CVS	0.62	0.078
SUN	0.68	0.013	DVA	0.60	0.011
CVS	0.64	0.067	SUN	0.59	0.011
QSS	0.56	0.011	QSS	0.51	0.011

Table 2: Algorithms ranked by AUC score for the MIT and ImgSal datasets, presented along with the standard deviation calculated over bin scores representing the degree of inherent spatial bias. High performance on both data sets appears to be correlated with spatial bias

specific models were chosen for presentation in Figure 7 as they represent the most biased (GBVS) and most consistent (AWS) performance of the algorithms tested, as well as a representation of performance for a model which is only based on spatial location (cG).

As mentioned in Section 3, one can calculate traditional AUC scores in a straightforward manner from the binned ROC results. We present the ordered ranking of unsmoothed algorithm performance over the MIT dataset in Table 2, along with the standard deviation of their binned AUC scores as a measure of the inherent spatial bias in each model. This provides a user with a direct performance measure (AUC score) which gives them a clear sense of algorithm performance operating over natural scenes, which is useful for any application in which choice of algorithm is solely dependent on its ability to predict human fixations in these environments. At the same time, we also have a quantifiable measure of how much of this performance is likely based on simple spatial bias versus an ability to identify salient visual stimuli, which is important for future scientific pursuits into saliency and saliency algorithm design.

We also present in Table 3 the AUC scores from the MIT dataset which have been weighted according to the relative image area occupied by each bin. The intention here is to provide a reasonable form of spatial correction, but which is transparent and deterministic in its source.

One such example of exploration into aspects of saliency algorithm performance is in the effect of smoothing kernel size. To explore this issue, we focused our efforts on the AIM algorithm as it has previously been shown to typically achieve maximum performance at relatively large smoothing kernel sizes [23, 19]. However, as kernel size increases

MIT		ImgSal	
Model	wAUC	Model	wAUC
GBVS	0.78	GBVS	0.70
CAS	0.76	CAS	0.69
cG	0.74	IKN	0.67
AWS	0.74	SSR	0.66
IKN	0.73	cG	0.65
SSR	0.71	AWS	0.65
AIM	0.71	AIM	0.63
DVA	0.70	DVA	0.60
SUN	0.67	CVS	0.59
CVS	0.61	SUN	0.59
QSS	0.56	QSS	0.51

Table 3: Algorithms ranked by AUC score weighted by relative bin area for the MIT and ImgSal datasets. For models with low spatial bias (like AWS and AIM), there is little change in AUC score, while there is a significant drop in score for highly biased models (such as GBVS and CAS)

the numerical effects of border padding likewise increase, suggesting that at least some of these gains are due to the introduction of an implicit center bias [13]. The exact degree to which improvements are due to the direct act of smoothing versus the introduction of spatial bias have previously not been quantified. Using spROC, however, we can directly explore this issue. Figure 8 displays the AUC scores by bin number for a range of different smoothing kernels acting on the AIM algorithm.

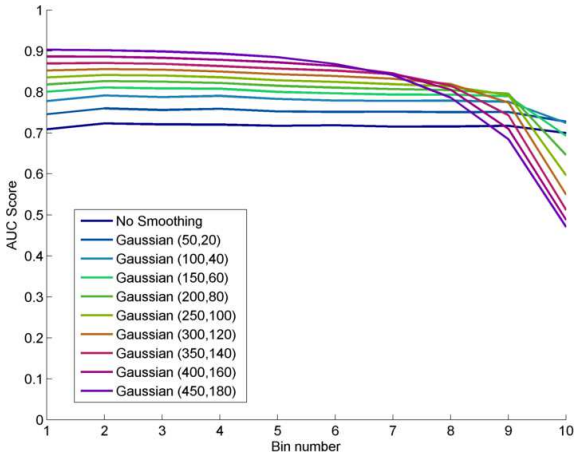


Figure 8: AUC score by bin for different degrees of smoothing applied to the AIM algorithm applied to the MIT dataset. Kernel properties are reported as $(size, \sigma)$. Initial smoothing boosts performance overall without appreciable increases in bias, but very large smoothing kernels sacrifice peripheral performance for central gains

At the smallest smoothing kernel tested, algorithm performance is almost uniformly boosted across all bins, including in the periphery. Subsequent smoothing initially boosts central scores without affecting peripheral performance, but a trade-off quickly develops thereafter between central gains and peripheral losses. Thus, we are able to begin to quantify the complex interactions smoothing has on the saliency signal, which opens the doors to further research into generally optimized post-processing techniques.

Although we have concentrated here on one particular form of spatial binning, it should be straightforward to extend this methodology to explore other interesting aspects of saliency model performance. Of particular interest may be temporal binning, in which fixation points are binned by temporal order rather than spatial location.

5. Conclusion

Saliency algorithms are applied to a steadily increasing range of problems, and the pertinent aspects of performance will often change with the specific requirements of an application area. A primary difficulty in evaluating algorithm performance differences is the complicated interaction which visual appearance and spatial location have on salience. While it is true that traditional ROC metrics have a hard time fairly evaluating an algorithm's ability to identify visually distinct image elements given the sometimes overwhelming spatial component of the ground-truth set, discounting the role of spatial location in saliency can likewise lead to misleading conclusions regarding relative algorithm performance. This is particularly true for applications (such as image compression) in which gross predictive performance is more important than the underlying reason for why an element is salient.

We have presented here a novel evaluative method which provides insight into the impact of spatial location on algorithm performance. The method is flexible enough to be tailored for analyzing a wide range of aspects of algorithm performance, but can nevertheless be easily collapsed back into a straightforward measure of performance. We demonstrated a similar rank-ordering as found in the benchmark work of Judd *et al*. [23], but with added information specifying the spatial bias inherent to the tested algorithms. Further, we were able to directly explore the role of Gaussian smoothing on the spatial bias of an algorithm's performance. This provides us with the ability to begin quantifying *how* rather than simply *how much* smoothing modulates the saliency signal, which opens up a novel avenue of research into saliency algorithm optimization.

References

[1] G. Boccignone and M. Ferraro. Modelling gaze shift as a constrained random walk. *Physica A: Statistical Mechanics and its Applications*, 331:207 – 218, 2004. 5

[2] A. Borji and L. Itti. Defending yarbus: Eye movements reveal observers' task. *Journal of Vision*, 14(3):29, 2014. 3

[3] A. Borji, D. N. Sihite, and L. Itti. Objects do not predict fixations better than early saliency: A re-analysis of einhuser et al.'s data. *Journal of Vision*, 13:1–4, 2013. 2

[4] A. Borji, D. N. Sihite, and L. Itti. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *Image Processing, IEEE Transactions on*, 22(1):55–69, 2013. 2, 6

[5] A. N. Borodin and I. A. Ibragimov. Limit theorems for functionals of random walks. In V. N. Sudakov, editor, *Proceedings of the Steklov Institute of Mathematics*, volume 195, 1995. 4

[6] D. Brockmann and T. Geisel. The ecology of gaze shifts. *Neurocomputing*, 32-33:643–650, 2000. 5

[7] N. D. Bruce, C. Wloka, N. Frosst, S. Rahman, and J. K. Tsotsos. On computational modeling of visual saliency: Examining whats right, and whats left. *Vision Research*, 116, Part B:95 – 112, 2015. Computational Models of Visual Attention. 3, 6

[8] N. D. B. Bruce and J. K. Tsotsos. Saliency based on information maximization. *Advances in Neural Information Processing*, 18:155–162, 2006. 1, 6

[9] Z. Bylinskii, E. M. DeGennaro, R. Rajalingham, H. Ruda, J. Zhang, and J. Tsotsos. Towards the quantitative evaluation of visual attention models. *Vision Research*, In Press, 2015. 2

[10] C.-K. Chang, C. Siagian, and L. Itti. Mobile robot vision navigation and localization using gist and saliency. In *Proc. Intelligent Robots and Systems (IROS)*, 2010. 1

[11] E. Erdem and A. Erdem. Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of Vision*, 13(4):11, 2013. 6

[12] S. Frintrop, G. Backer, and E. Rome. Goal-directed search with a top-down modulated computational attention system. In *Pattern Recognition*, volume 3663, pages 117–124. Springer Berlin Heidelberg, 2005. 3

[13] N. M. W. Frosst, C. Wloka, and J. K. Tsotsos. The effects of image padding in saliency algorithms. *Perception*, 43 ECVP Abstract Supplement:106, 2014. 8

[14] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil. Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing*, 30(1):51 – 64, 2012. 1, 6

[15] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34:1915–1926, 2012. 1, 6

[16] D. M. Green, J. A. Swets, et al. *Signal detection theory and psychophysics*, volume 1. Wiley New York, 1966. 2

[17] C. Guo and L. Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Transactions on Image Processing*, 19:185–198, 2010. 1

[18] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. *Advances in Neural Information Processing*, 19:545–552, 2007. 1, 6

[19] X. Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34:194–201, 2012. 1, 7

[20] X. Hou and L. Zhang. Dynamic visual attention: Searcing for coding length increments. *Neural Information Processing Systems*, 21:681–688, 2008. 1, 6

[21] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:1254–1259, 1998. 1, 6

[22] M. Jiang, S. Huang, J. Duan, and Q. Zhao. Salicon: Saliency in context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1

[23] T. Judd, F. Durand, and A. Torralba. A benchmark of computational models of saliency to predict human fixations. Technical report, Massachusetts Institute of Technology, 2012. 2, 6, 7, 8

[24] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. `http://people.csail.mit.edu/tjudd/WherePeopleLook/`. 3

[25] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *International Conference on Computer Vision*, 2009. 1, 2, 3, 4

[26] C. Kanan, M. H. Tong, L. Zhang, and G. W. Cottrell. Sun: Top-down saliency using natural statistics. *Visual Cognition*, 17:979–1003, 2009. 3

[27] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985. 1

[28] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, pages 79–86, 1951. 2

[29] J. Li, M. D. Levine, X. An, X. Xu, and H. He. Visual saliency based on scale-space analysis in the frequency domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:996–1010, 2013. 2, 6

[30] I. v. d. Linde, U. Rajashekar, A. C. Bovik, and L. K. Cormack. DOVES: A database of visual eye movements. Retrieved from `http://live.ece.utexas.edu/research/doves`. 3

[31] I. v. d. Linde, U. Rajashekar, A. C. Bovik, and L. K. Cormack. DOVES: A database of visual eye movements. *Spatial Vision*, 22:161–177, 2009. 3, 4

[32] S. Mathe and C. Sminchisescu. Dynamic eye movement datasets and learnt saliency models for visual action recognition. *European Conference on Computer Vision*, 2012. 3

[33] S. Mathe and C. Sminchisescu. Action from still image dataset and inverse optimal control to learn task specific visual scanpaths. *Advances in Neural Information Processing*, pages 1923–1931, 2013. 3

[34] D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42:107–123, 2002. 2

[35] D. Parkhurst and E. Niebur. Scene content selected by active vision. *Spatial Vision*, 16:125–154, 2003. 2

[36] R. J. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45:2397–2416, 2005. 2

[37] A. Rasouli and J. K. Tsotsos. Visual saliency improves autonomous visual search. In *Computer and Robot Vision (CRV)*, 2014. 1

[38] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit. Saliency and human fixations: State-of-the-art and study of comparison metrics. In *International Conference on Computer Vision (ICCV)*, pages 1153–1160, 2013. 2, 6

[39] R. Roberts, D.-N. Ta, J. Straub, K. Ok, and F. Dellaert. Saliency detection and model-based tracking: A two part vision system for small robot navigation in forested environments. In *Proc. SPIE 8387*, 2012. 1

[40] Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *Computer Vision, 1998. Sixth International Conference on*, pages 59–66. IEEE, 1998. 2

[41] B. Russel, A. Torralba, K. Murphy, and W. T. Freeman. LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 2007. 3

[42] B. Schauerte and R. Stiefelhagen. Quaternion-based spectral saliency detection for eye fixation prediction. In *European Conference on Computer Vision*, pages 116–129, 2012. 1, 6

[43] H. J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9(12):1–27, 2012. 6

[44] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist. Visual correlates of fixation selection: effects of scale and time. *Vision Research*, 45:643–659, 2005. 2

[45] B. W. Tatler and B. T. Vincent. The prominence of behavioural biases in eye guidance. *Visual Cognition*, 17:1029–1054, 2009. 3

[46] A. Treisman and G. Gelade. A feature integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980. 1

[47] J.-F. Tsai and K.-J. Chang. Opensource implementation of context-aware saliency detection. https://sites.google.com/a/jyunfan.co.cc/site/opensource-1/contextsaliency. 6

[48] J. H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings: Biological Sciences*, 265:359–366, 1998. 3

[49] E. Vig, M. Dorr, and D. Cox. Space-variant descriptor sampling for action recognition based on saliency and eye movements. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *ECCV 2012*, volume 7578 of *Lecture Notes in Computer Science*, pages 84–97. Springer Berlin Heidelberg, 2012. 3

[50] C. Wloka and J. K. Tsotsos. Overt fixations reflect a natural central bias. *Journal of Vision*, 13:239, 2013. 4

[51] A. Yarbus. *Eye Movements and Vision*. Plenum Press, 1967. 3

[52] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7:32):1–20, 2008. 1, 2, 6