On Computational Modeling of Visual Saliency: Examining What's Right, and What's Left

Article *in* Vision Research · February 2015

DOI: 10.1016/j.visres.2015.01.010 · Source: PubMed

Vision Research xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

On computational modeling of visual saliency: Examining what's right, and what's left

Neil D.B. Bruce [a,*], Calden Wloka [b,c], Nick Frosst [b], Shafin Rahman [a], John K. Tsotsos [b,c]

[a] Department of Computer Science, University of Manitoba, 66 Chancellors Cir, Winnipeg, Manitoba, Canada
[b] Dept. of Electrical Engineering and Computer Science, Lassonde School of Engineering, York University, 4700 Keele Street, Toronto, Ontario, Canada
[c] Centre for Vision Research, York University, 4700 Keele Street, Toronto, Ontario, Canada

ARTICLE INFO

Article history:
Received 26 July 2014
Received in revised form 16 December 2014
Available online xxxx

Keywords:
Saliency
Modeling
Evaluation
Eye tracking
Visual search
Computer vision

ABSTRACT

In the past decade, a large number of computational models of visual saliency have been proposed. Recently a number of comprehensive benchmark studies have been presented, with the goal of assessing the performance landscape of saliency models under varying conditions. This has been accomplished by considering fixation data, annotated image regions, and stimulus patterns inspired by psychophysics. In this paper, we present a high-level examination of challenges in computational modeling of visual saliency, with a heavy emphasis on human vision and neural computation. This includes careful assessment of different metrics for performance of visual saliency models, and identification of remaining difficulties in assessing model performance. We also consider the importance of a number of issues relevant to all saliency models including scale-space, the impact of border effects, and spatial or central bias. Additionally, we consider the biological plausibility of models in stepping away from exemplar input patterns towards a set of more general theoretical principles consistent with behavioral experiments. As a whole, this presentation establishes important obstacles that remain in visual saliency modeling, in addition to identifying a number of important avenues for further investigation.

1. Introduction

Attention and visual search are the product of a complex set of processes involving a myriad of different neural mechanisms working in concert. Understanding the nature of these processes and their interplay is a significant challenge. Even in light of decades of research involving visual psychophysics and an accumulating body of data from single cell recording and various brain imaging modalities, a consensus on the specific mechanisms involved at the level of architectural, or even relatively high-level philosophical points remains elusive (Carrasco, 2011). One important element of this complex concert of neural processes are the mechanisms that drive visual saliency. Certain visual patterns tend to draw covert or overt attention on the basis of the stimulus properties subject to the surrounding context, or are influenced by overall scene composition (Torralba et al., 2006). Significant progress in the characterization of visual salience has been made in the past decade, specifically in the domain of computational modeling and benchmarking (e.g. Borji, Sihite, & Itti, 2012; Judd, Durand, & Torralba, 2012; Andreopoulos & Tsotsos, 2012; Borji et al., 2013; Winkler & Ramanathan, 2013; Riche et al., 2013; Meur et al., 2013; Borji, Sihite, & Itti, 2013).

The body of research pointing to the importance of task driven visual attention also suggests that greater benefit may be had in placing more emphasis on modeling the neural substrates of attention that are driven by task or contextual factors (Hayhoe & Ballard, 2005; Frintrop, Backer, & Rome, 2005; Tatler, 2009; Ballard & Hayhoe, 2009; Koehler et al., 2014; Chen & Zelinsky, 2006; Henderson et al., 2007). It is evident that a fully comprehensive understanding of attention, gaze and visual search will only be had in developing sophisticated models that include all of these considerations as central elements. That said, visual salience remains an important factor in itself, in the complex interplay involved in the neural guidance of attention. In this paper, we therefore attempt to frame the role of visual salience insofar as it may contribute to the overall understanding of attention. The subject matter of this paper also seeks to highlight some of the shortcomings and challenges in assessing model validity for visual salience in considering gaze data, and behavioral observations from the psychophysics literature. We also attempt to highlight subproblems that may benefit most from receiving greater emphasis in future efforts, and address the need to disentangle salience from the many other competing factors inherent in forming an objective assessment and analysis of models of visual saliency.

* Corresponding author.
E-mail address: bruce@cs.umanitoba.ca (N.D.B. Bruce).

http://dx.doi.org/10.1016/j.visres.2015.01.010
0042-6989/

Visual saliency presents a useful characterization for many applications in image processing (Avidan & Shamir, 2007; Achanta & Susstrunk, 2009), computer vision (Kadir & Brady, 2001; Achanta et al., 2008; Liu et al., 2011), robotics (Frintrop, Jensfelt, & Christensen, 2007; Butko et al., 2008; Chang, Siagian, & Itti, 2010; Klein & Frintrop, 2011), graphics (Lee, Varshney, & Jacobs, 2005; Kim & Varshney, 2006; Longhurst, Debattista, & Chalmers, 2006) and human–computer interaction (Halverson & Hornof, 2007; Suh et al., 2003) among other problem domains. A consequence of this is fragmentation in the literature, with modeling efforts residing at several different levels of abstraction and towards various disparate goals. For example, a model that is relatively successful at predicting fixation patterns may have relatively little connection to biology at the level of neural circuitry but provide a strong functional characterization. Alternatively, prediction of fixated locations for a certain class of stimuli (such as real world scenes) might be successfully achieved, but in a fashion that is relatively detached from biology even at a functional level and thereby violates some of the core behavioral observations that are observed in basic visual psychophysics tasks.

In Section 2, we discuss a number of issues in fixation-based benchmarking that arguably warrant further consideration when weighed against the broader set of sensory and motor systems that interact with visual saliency. Specifically, the following points are considered:

1. Scale: Scale is discussed in detail including its crucial importance in the prediction of salient regions. This also includes consideration of some facets of computation common to models involved in fixation prediction, including post-processing of raw saliency maps in the form of Gaussian blurring and implications for neural information processing.
2. Border effects: The lack of spatial support for a filter positioned at the boundary of an image, and the spatial support among cells that represent visual input in the far periphery is important in its implications for behavior and benchmarking. We provide demonstrations of the importance of border effects, with emphasis on implications within biological systems.
3. Spatial bias: In considering spatial bias, there has been much debate as to the most appropriate means of evaluating models that target prediction of gaze data. One important factor in this analysis is the observed spatial (central) bias in fixation data (Zhang et al., 2008; Tseng et al., 2009; Judd, 2011). In this paper we also present results on the impact of spatial bias, implications for metrics, and discussion of why this is an issue that requires careful consideration.
4. Context and scene composition: Some prior efforts in fixation prediction have demonstrated the utility of considering context (Torralba et al., 2006), with even relatively coarse spatial guidance providing a means of improving performance in characterizing visual saliency. We also discuss existing efforts in this domain, and consider additional questions relating to scene composition that may present important targets for future work in characterizing visual saliency.
5. Oculomotor constraints: While visual saliency is often considered in isolation, there is a great degree of interaction with other systems involved in visual sensing including cortical and sub-cortical mechanisms (Hopfinger, Buonocore, & Mangun, 2000) involved in directing gaze. This carries additional implications for the nature of observed fixation patterns, and also arguably for the appropriateness of metrics for benchmarking in gaze prediction (Foulsham & Underwood, 2008).

In Section 3, we shift focus to considering visual saliency, search efficiency and pop-out viewed through the lens of the large corpus of visual psychophysics literature that has accumulated over the past several decades. We discuss the shortcomings of purely quantitative tests of biological plausibility, and aim at establishing a base set of constraints that might be applied as a test of biological consistency of model behavior. This includes the following areas of focus:

1. A core focus of this section lies in drawing out a base set of patterns from classic psychophysics targeting visual search that serves to constrain expected model behavior. This analysis is directed at identifying critical behavioral patterns that are pervasive in the psychophysics literature, and attempts to translate these observations to constraints on visual saliency models.
2. We also consider the problem at the level of physiology, and discuss possible implications for computational modeling such as receptive field size, the role of recurrence, and involvement of higher level cognitive function in determining fixation behavior.

Finally, Section 4 presents general discussion of important points that emerge from the analysis appearing in this paper and we also highlight a number of potentially important avenues for further investigation. Given differences in emphasis between Sections 2 and 3, both of these sections are mostly self-contained and may be read in accord with the reader's level of interest.

2. An appraisal of benchmarking efforts: Challenges and potential confounds

The past decade has been marked by significant growth in data and computational capabilities available for pattern analysis, combined with progress in techniques in computer vision, image processing and machine learning. Gaze data, and the analysis of such data has also followed this trend with many sources of fixation data and models to choose from Winkler and Ramanathan (2013). This wealth of data has also brought with it a much greater capacity to examine the behavior of modeling efforts in a quantitative manner. While such efforts are eminently useful, they also carry the risk of drawing focus away from more systematic and careful examination of some of the fundamental theoretical questions that underlie the problem, or discourse on appropriate methods for measuring success beyond quantitative measures.

Given that several comprehensive and recent benchmark studies have been presented (Judd et al., 2012; Borji et al., 2013; Riche et al., 2013; Borji, Sihite, & Itti, 2013), the target of this section instead lies in examining what benchmarks say about the general problem of saliency modeling, and also focuses on how saliency relates to the larger domain of visual attention. This is done by highlighting facets of the problem domain that have been well studied, as well as those that have been relatively neglected. This also includes a critical appraisal of common practices in fixation based benchmarking. In particular, the argument is made that weaknesses associated with fixation based benchmarking are of sufficient concern to challenge the notion that these benchmarks are suitable for gauging model validity.

In the balance of this section we focus on a number of specific considerations in modeling and assessment. In Section 2.1 we consider similarity among models, the suitability of various metrics, and the role of central bias in evaluation. In Section 2.2 we examine the critical role of scale in determining visual saliency. This also includes important additional elements that interact with scale such as spatial bias in fixation patterns, and the role of blurring

Please cite this article in press as: Bruce, N. D. B., et al. On computational modeling of visual saliency: Examining what's right, and what's left. *Vision Research* (2015), http://dx.doi.org/10.1016/j.visres.2015.01.010

within the processing pipeline. Finally, in Section 2.5 we discuss the role of *high level* factors in determining fixation behavior, including context and scene structure.

2.1. Model similarity

There exist numerous different metrics that have been exercised for the purposes of characterizing the performance of visual saliency algorithms. Metrics include Pearson and Spearman correlation coefficients (Kendall et al., 1946), Normalized-Scanpath Salience (Peters & Itti, 2007), Earth-Mover's Distance (Rubner, Tomasi, & Guibas, 1998), Kullback–Leibler Divergence (Kullback & Leibler, 1951), and Precision-Recall (Ziman, 1969) and Receiver Operating Characteristic (ROC) curves (Green et al., 1966). For a detailed discussion of these metrics in visual saliency evaluation, the reader may refer to (Borji, Sihite, & Itti, 2013; Riche et al., 2013). In what follows, the discussion of these metrics resides at the more general level of implications of influence from regions of the visual cortex that have likely involvement with visual saliency computation. In the seminal work of Itti, Koch, and Niebur (1998), Itti and Koch (2001) evaluation was based primarily on simulation of fixations given an underlying representation of visual saliency, with fixation points determined on the basis of a Winner-Take-All selection strategy (Koch & Ullman, 1987; Tsotsos et al., 1995; Lee et al., 1999). Such a strategy in not uncommon among computational models of attention (Itti, Koch, & Niebur, 1998; Tsotsos et al., 1995), and generally there is consensus that attention involves some form of *focal* selection strategy in guiding fixations. It is therefore worth considering what some of the implications of such a selection strategy may be, as it relates to visual saliency metrics:

1. Absolute versus relative saliency: Assuming overt shifts of attention may occur according to a strategy that shifts gaze on the basis of an underlying representation of visual saliency, the absolute magnitude of values in a saliency map might be considered a relatively less important consideration than relative values.
2. Contrast dependence: Similar to previous point, there are significant differences in the dynamic range of output produced by different algorithms. Assuming a selection strategy in overt attention, the contrast of saliency maps should have little impact on observed fixations.
3. Value versus rank: For metrics sensitive to numeric values in algorithmic saliency output, relative performance of algorithms may be challenging to judge given that differences in output contrast may appreciably affect numeric scores.

It is important to note that while the contrast of saliency maps and absolute magnitude of values (as opposed to relative values/rank order) are unimportant in the context of some benchmarking metrics, the magnitude of these values are important when considering human behavior. The strongest value in a saliency map remains unchanged subject to a change of contrast, or change in the range of values. However, the dynamics of overt attention may be influenced by the absolute magnitude of activity associated with saliency computation. Duncan and Humphreys (1989) show that increasing target-distractor similarity (thereby reducing the magnitude of the saliency difference) increases visual search times, even though the target remains overall more salient. Likewise, at a basic level, saccade dynamics may be altered subject to an absolute measure of activation (Schütz, TrommershSuser, & Gegenfurtner, 2012). This gains even greater importance in recognizing that representations associated with visual saliency and those tied to visuomotor control may have differing temporal dynamics. The time scale of activation associated with salient patterns, versus timing in fixation control and/or inhibition of return may give rise to a complex interplay between the underlying representation of saliency, and representations driving the control of overt attention. This hints at one shortcoming of existing benchmarking paradigms associated with the static nature of prevailing fixation based benchmarking efforts. This point is discussed in greater detail later in this section.

For the grouping of metrics presented by Riche et al. (2013) and grouping of models given by Judd, Durand, and Toralba (2012), some of the aforementioned considerations may have an impact on the determination of metric similarity (or also model similarity). Metric similarity may result from similarity in output produced by saliency algorithms, but also the sensitivity of metrics to absolute numeric factors. For example, contrast of saliency maps may bias the determination of metric similarity. An additional factor that may have bearing on this point, is the impact of spatial bias in fixation data used for evaluation. This point is further discussed in Section 2.2.1. It is therefore important to view such clustering of metrics with due consideration for nuisance factors that may lead to similarities due to numeric factors (e.g. contrast of algorithm output) rather than similarity in metric or model behavior.

One metric that is relatively insensitive to the aforementioned considerations is ROC score which is insensitive to the absolute numeric values or contrast of saliency maps. Points along the ROC curve are calculated as an (x, y) coordinate pair where the y-coordinate is the *sensitivity* (proportional to the true positive rate (TPR)), and the x-coordinate is equal to $1 - specificity$ (proportional to the false positive rate (FPR)). Even in the realm of computing ROC scores, there has been considerable debate on the specific methodology for ROC based evaluation. Although there is general agreement on how to calculate TPR (the proportion of pixels fixated by humans which are also marked as salient by the saliency algorithm), there is more disagreement on the best way to determine the FPR of an algorithm.

The different methodologies employed largely fall into two main groups: what are referred to as *traditional* ROC assessments and *shuffled* ROC assessments. In the traditional ROC metric the FPR is determined by taking the proportion of salient pixels which are above the salience threshold but were not fixated by human observers. In contrast, shuffled ROC metrics, originally developed by Tattler, Baddeley, and Gilchrist (2005), determine the FPR by taking a sample of pixels which were fixated in other images from the dataset but not in the current image and determining the percentage of those pixels which were labeled as salient. By essentially "shuffling" fixation data across all the images in the dataset, shuffled ROC metrics attempt to correct for systematic spatial biases. Examples of evaluations which employ a traditional ROC based assessment can be found in Bruce and Tsotsos (2006), Harel, Koch, and Perona, 2006, and Gao, Mahadevan, and Vasconcelos (2008), while Zhang et al. (2008) and Borji, Shiti, and Itti (2013) use a shuffled ROC score. It is also worth noting that some studies have attempted to solve for central biasing in the data despite the use of a traditional ROC metric by optimizing a Gaussian central prior in an algorithm-dependent manner in order to place all algorithms on common ground (Judd, Durand, & Toralba, 2012).

In the context of ROC evaluation, we also wish to consider similarity among a variety of saliency models that show favorable performance in the benchmark of Borji, Shiti, and Itti (2013). This analysis considers both the Bruce/Toronto (Bruce & Tsotsos, 2006) and Judd/MIT (Judd et al., 2009) datasets individually. Cross-dataset examination is important as one would hope to observe similar topology across datasets assuming the nature of the images is not significantly different. The measure of similarity among models is based on the correlation between algorithms for shuffled area under ROC scores (auROC), corresponding to all of the

Please cite this article in press as: Bruce, N. D. B., et al. On computational modeling of visual saliency: Examining what's right, and what's left. *Vision Research* (2015), http://dx.doi.org/10.1016/j.visres.2015.01.010

images in a given data set (inter-algorithm distances are defined by Pearson's correlation for auROC scores). In computing inter-algorithm distances based on auROC, algorithms are optimized in selection of the blurring factor, and image scale that results in greatest auROC performance. Visualization is achieved via multidimensional scaling (Kruskal, 1964), such that the relative position of algorithms in the 2D plot best matches the distances between algorithms in the higher dimensionality space that exactly captures the inter-algorithm distances. More specifically, dissimilarity between any pair of algorithms is given by $1-\rho$ where ρ corresponds to the correlation between auROC scores for the two algorithms across all images in a dataset. Embedding in two dimensions is achieved in minimizing the squared difference between dissimilarity scores in the original 66-dimensional space required to exactly represent inter-algorithm dissimilarity, and the resulting two dimensional embedding.

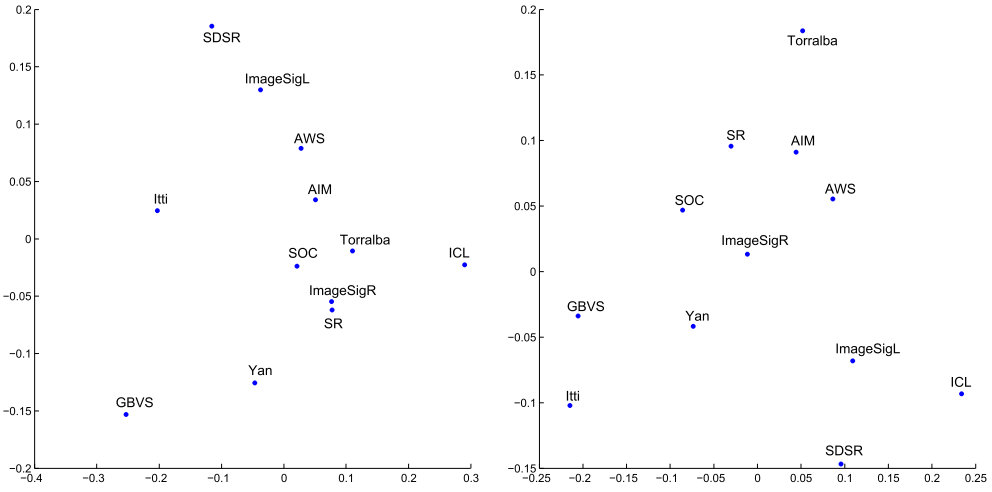
Algorithms considered in this visualization include: The Incremental Coding Length model (ICL) (Hou & Zhang, 2009), the *classic* saliency model of Itti, Koch, and Niebur (1998) (Itti), the Second-Order contrast model (SOC) (Rahman et al., 2014), Graph-Based Visual Salience (GBVS) (Harel, Koch, & Perona, 2006), the Spectral Residual model (SR) (Hou & Zhang, 2007), the Image Signature model (RGB – ImageSigR, Lab – ImageSigL) (Hou, Karel, & Koch, 2012), Adaptive Whitening Saliency (AWS) (Garcia-Diaz et al., 2012), Rank-Sparsity Decomposition (Yan) (Yan et al., 2010), Torralba et al. (2006)'s model based on inverse likelihoods (Torralba), Attention based on Information Maximization (AIM) (Bruce & Tsotsos, 2006; Bruce & Tsotsos, 2009) and Saliency Detection by Self-Resemblance (SDSR) (Seo & Milanfar, 2009) (For a more complete discussion of models, readers may wish to refer to (Frintrop, Rome, & Christensen, 2010; Borji, Shiti, & Itti, 2013)).

Fig. 1 shows the *flattened* manifold in 2D demonstrating the similarity among algorithms with respect to prediction performance for individual images. In Fig. 1, there is similarity in the relative positioning of algorithms that is consistent across the Bruce/Toronto, and Judd/MIT datasets. In particular, a common arc in the center of each 2D plot consists of Torralba, AIM, AWS, ImageSigL, SDSR algorithms (Left: bottom to top. Right: top to bottom). The ICL algorithm appears on one side of this arc, with SOC, ImageSigR, and SR algorithms appearing on the other proximal to the aforementioned arc. An additional grouping of algorithms consists of the Itti, GBVS and Yan algorithms which are more distal to this central arc.

It is also of value in examining the visualization presented in Fig. 1, to consider factors that may explain similarities among models. In particular, this may help to inform upon which factors have a significant influence on model similarity when viewed according to a particular metric. The Torralba model is proximal to the AIM model, and each of these models includes an inverse function of likelihood within its contrast measure. The AIM model has similarities to the AWS model in that decorrelation of features through ICA achieves a similar effect to adaptive whitening, albeit with the latter being a rotation of chromatic, and spectral energy that is image dependent. Decorrelation or whitening of chroma yields a separation of luminance and chromatic channels (akin to YCbCr, CIELab spaces) based on a linear transformation. The outer models of this arc (ImageSigL, SDSR) both employ CIELab color space in determining saliency. While the ImageSigL model shares the property of efficient coding as a motivating factor with AIM and AWS, SDSR employs a non-linear covariance based kernel operator to determine feature contrast. The Itti, GBVS and Yan models all exhibit a greater degree of center bias in their output, with Itti and GBVS sharing similar features, and Yan based on a contrast measure more proximal to the algorithms motivated by efficient coding (including ImageSigR). While this analysis is not exhaustive, it does provide a sense of some of the factors that may determine the landscape of model behavior and similarity between models. It is important to note that some of these factors are associated with basic model mechanics, while others derive from the assumed feature representation, or the impact of spatial bias. This presents one instance of model performance or model relatedness being skewed by the nature of the evaluation protocol which in this instance includes a relatively ad hoc means of normalizing for differences in the spatial distribution associated with model output. As discussed later on in this section, benchmarking of saliency is confounded by a number of latent factors that may be important to consider in further evaluation efforts, and also may have a significant impact on the interpretation of similarity between saliency algorithms. The preceding discussion hints at the nature of some of these considerations, and the importance of features and spatial bias in saliency models and benchmarks is further demonstrated in the remainder of this section.

2.2. Scale

The problem of determining the appropriate scale at which to analyze a signal is an old issue in the field of signal analysis. The challenge goes far beyond simply filtering out fine-scale noise, and instead encompasses the enormous difficulty of teasing apart



(a) Bruce/Toronto Dataset (b) Judd/MIT Dataset

Fig. 1. Two-dimensional multi-dimensional scaling of per image ROC scores based on correlation between shuffled auROC scores.

Please cite this article in press as: Bruce, N. D. B., et al. On computational modeling of visual saliency: Examining what's right, and what's left. *Vision Research* (2015), http://dx.doi.org/10.1016/j.visres.2015.01.010

a signal formed by distinct physical processes, each of which operates at a different spatial scale (Marr, 1982; Witkin, 1984). Wolfe (1998) identified scale as one of the basic features of visual search, albeit a somewhat complex aspect interrelated with stimulus size and spatial frequency. Thus, when saliency is viewed as a fundamental model of overt attentional capture in human visual search, it becomes unavoidable that spatial scale will necessarily impact the results (Culhane & Tsotsos, 1992).

We examine this topic first by looking at the spatial scale of the filters used to generate the saliency signal (Section 2.2.1) followed by a discussion of the spatial scale of the saliency signal itself (Section 2.2.2). For a further discussion of these issues with a stronger emphasis on basic characteristics of human visual processing as it relates to saliency and scale, the reader may refer to Section 3.1.6. When using large convolution kernels or performing frequency domain calculations, the method for handling response values near an image border can become quite important, which is discussed in more detail in Section 2.2.2.

2.2.1. Filter scale and center-bias

Some algorithms are filter-based while others rely on alternative mechanisms, and there also exists differences in the types of features considered. This yields a potentially large parameter space for each algorithm, and for this reason, it is difficult to precisely control the effective spatial scale that any given algorithm operates upon. Many algorithms rely on re-sizing of images to a smaller and common size *a priori* to achieve some degree of invariance to scale, but it remains the case that a particular scale is associated with each algorithm. For example, in Itti, Koch, and Niebur (1998)'s model, filters are represented at multiple spatial scales, however the absolute high and low frequency cut-offs and distribution of spectral coverage varies as a function of image size. For the ICA filters employed in AIM, the low-frequency cut-off at the filter level is a function of the patch size of the filters, and the high frequency cut-off a function of the number of filters retained following PCA (generally those representing higher frequencies capture less variance). We therefore examine the sensitivity to scale by considering ROC scores at 100% scale and 50% scale for the Bruce/Toronto dataset (Bruce & Tsotsos, 2006), and at 50% and 25% scale for the Judd/Torralba dataset (Judd, Durand, & Toralba, 2012). We have also considered both the shuffled ROC metric (Zhang et al., 2008; Borji, Shiti, & Itti, 2013), and the standard ROC metric (Bruce & Tsotsos, 2006) in this evaluation.

Fig. 2 shows a number of ROC evaluations for some of the higher performing algorithms in the benchmark of Borji, Shiti, and Itti (2013), in addition to the relatively simplistic SOC model (Rahman et al., 2014) (used as a reference point for its relation to early cortical computation and simple structure). It is interesting to note that the rank order of algorithms is sensitive to both the shuffled vs. unshuffled conditions, and for select algorithms to the scale of the image. Those that include resizing to a common scale as part of the cascade of operations are relatively unaffected (e.g. all of the spectral domain/DCT approaches). However, many of the models show sensitivity to the scale of the image. In addition, there are also algorithms that are relatively polarized when standard versus shuffled metrics are applied. One would naturally expect an advantage from those algorithms for which center bias is included in considering the standard ROC metric. However, there are additional factors that may contribute to such bias outside of an explicit step of adding central bias to a saliency map. For example, overall post-processing of the saliency map through blurring will induce a center bias in the case that a zero support is assumed outside of the image support. Boundary effects are discussed further in Section 2.2.2. Polarization of scores in the opposite direction (e.g. as in AWS (Garcia-Diaz et al., 2012)), also suggests the possibility of a *periphery bias* effect in considering shuffled AUC scores: Images for which the distribution of fixated points is similar to the population level distribution are guaranteed to receive less weight or be less *predictable* according to a shuffled auROC score. Factors such as post-process blurring are therefore relatively important to consider insofar as they may alter performance subject to any of these metrics.

To further evaluate this hypothesis, we have examined the relative spatial bias of the algorithms. To achieve this, the following steps are performed:

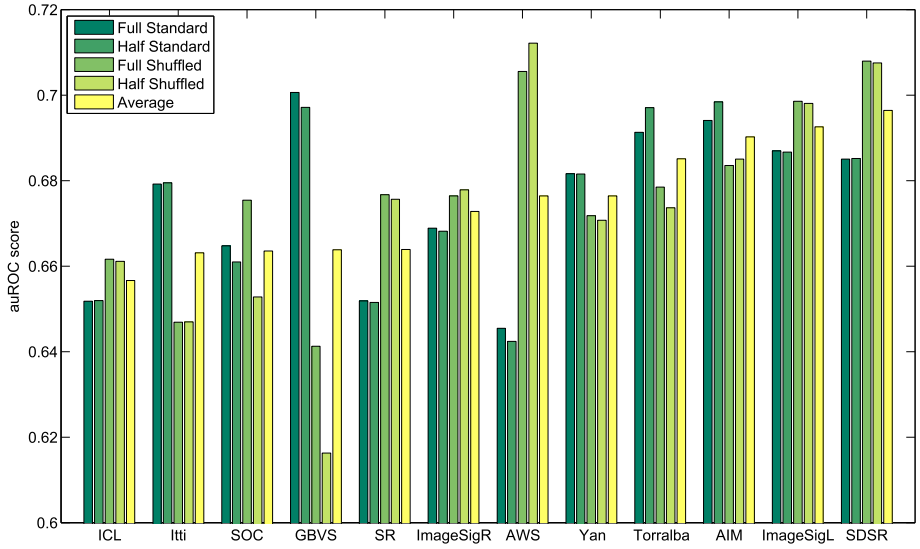
1. The individual saliency maps may have very different raw numeric values. We therefore have first normalized the individual saliency maps for each dataset so that each set of output saliency maps has a mean of 0 and standard deviation of 1. This operation places all of the algorithms on a common numeric scale (with common first and second order statistics).
2. All of the saliency maps corresponding to a particular saliency model are averaged to produce an overall topographical spatial profile of saliency that is algorithm dependent but not image dependent.
3. The same operation is performed across all algorithms to derive a generalized topographical representation of bias agnostic to both image and model.
4. Finally, the generalized topographical map is subtracted from each of the algorithm specific topographical maps to give a sense of relative spatial bias of different algorithms
5. This bias is visualized in raw form, and subject to histogram equalization to show both absolute and relative spatial bias profiles across algorithms

Results from this operation are depicted in Fig. 3. The top case demonstrates the spatial profile based on relative spatial bias. In the bottom image, these scores are histogram equalized to cover the entire contrast range. As can be seen from this line of analysis, an algorithm such as GBVS (Harel, Koch, & Perona, 2006), which fares much better in the traditional ROC based scoring has a relatively strong signature in its spatial bias for the center of the scene. In contrast, AWS (Garcia-Diaz et al., 2012) which demonstrates high shuffled ROC scores with lower traditional ROC scoring shows a relative *peripheral bias* when contrasted against other algorithms. This trend seems to apply to all of the algorithms examined, with those showing the highest differential in favor of shuffled auROC scores having peripheral bias, and those with highest differential towards standard ROC having a stronger center bias. It is also worth noting that model similarity determined via multi-dimensional scaling appears to be influenced to some extent by the spatial bias profiles. In practice, the factors that contribute to spatial bias are more complex and nuanced than a simple spatial weighting is able to capture (see 2.3 for further discussion). While a weighted prior may boost benchmark performance, it also obscures important model differences. It is also conceivable that coupling any representation of saliency with a model that accounts for other factors, including control of eye movements, may change the relative performance of models in a fashion that is quite different than a simple adjustment of spatial weighting. In this view, current strategies that accommodate for spatial bias serve primarily as a means of compensating for the gap between measures of stimulus salience, and the more involved set of processes that give rise to measured fixations. There also exists a more general impact of oculomotor control on spatial bias and scale. This point is revisited in Section 3.1.6.

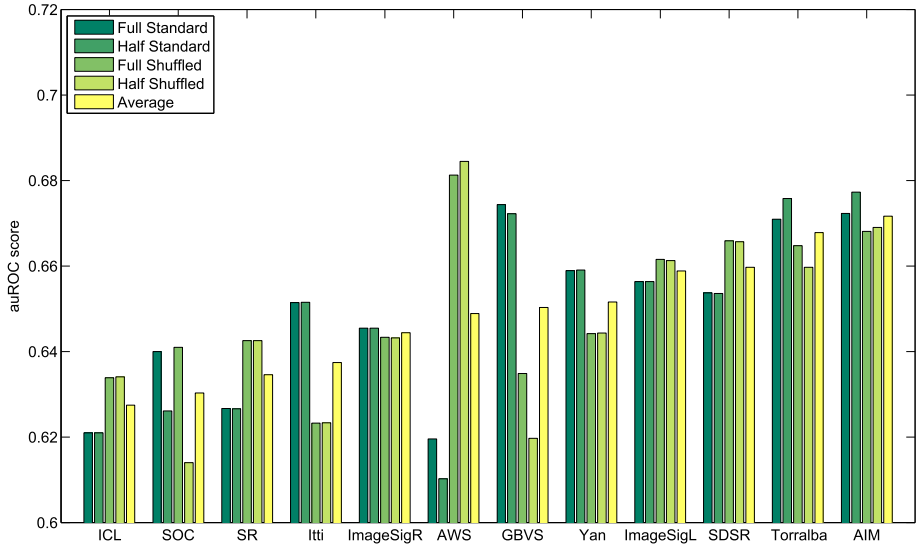
2.2.2. Saliency scale

Once the saliency map has been constructed from the underlying filter, it must still be interpreted (e.g. according to one of the

Please cite this article in press as: Bruce, N. D. B., et al. On computational modeling of visual saliency: Examining what's right, and what's left. *Vision Research* (2015), http://dx.doi.org/10.1016/j.visres.2015.01.010



(a) Bruce/Toronto Dataset



(b) Judd/Torralba Dataset

Fig. 2. Comparison of a variety of saliency algorithms based on auROC scores for both standard and shuffled ROC evaluation, and for both full and half scale images.

metrics listed in Section 2.1), which may be done at multiple scales. Gaussian smoothing of the raw saliency map (sometimes referred to as blurring) is a ubiquitous post-processing step, but the size of the smoothing kernel used to perform this calculation will have obvious implications for the spatial scale at which the saliency signal is ultimately analyzed. Although early applications of post-processing smoothing were motivated by the drop-off in visual acuity as one moves toward the periphery of the retina (Bruce & Tsotsos, 2006), the positive impact of smoothing on AUC scores for predicting human fixations has led it to becoming a standard processing step which is frequently included without justification (beyond its effect on performance) and optimized specifically for the given dataset and algorithm (Hou, Karel, & Koch, 2012; Judd et al., 2012). The size of the smoothing kernel which yields the best on-average performance tends to be quite large (for example, Judd et al., 2012 found the optimal standard deviation for their smoothing kernels ranged from 30–100 pixels depending on the algorithm), suggesting that the smoothing step is doing much more than simply filtering out high frequency spike noise.

The reason for the significant boost in fixation prediction measures from post-processing smoothing is currently unclear. Possible contributing factors include:

- A failure to correctly capture the appropriate stimulus scale with the model filters.
- Provide a grouping or clustering effect to push fixation prediction toward object centers or center of mass style fixations.
- Blur the saliency map to account for human saccadic error.
- Introduce an implicit center bias, particularly when using large smoothing kernels with zero-padding around the image border.

At the moment it is not clear how greatly any of the identified factors contribute to the efficacy of post-processing smoothing, or if there are some as yet unidentified factors at work. Nevertheless, treating the smoothing kernel as simply another parameter to be optimized does little to serve the problem of scientific understanding, and possibly even obscures the quality of the underlying

Please cite this article in press as: Bruce, N. D. B., et al. On computational modeling of visual saliency: Examining what's right, and what's left. *Vision Research* (2015), http://dx.doi.org/10.1016/j.visres.2015.01.010

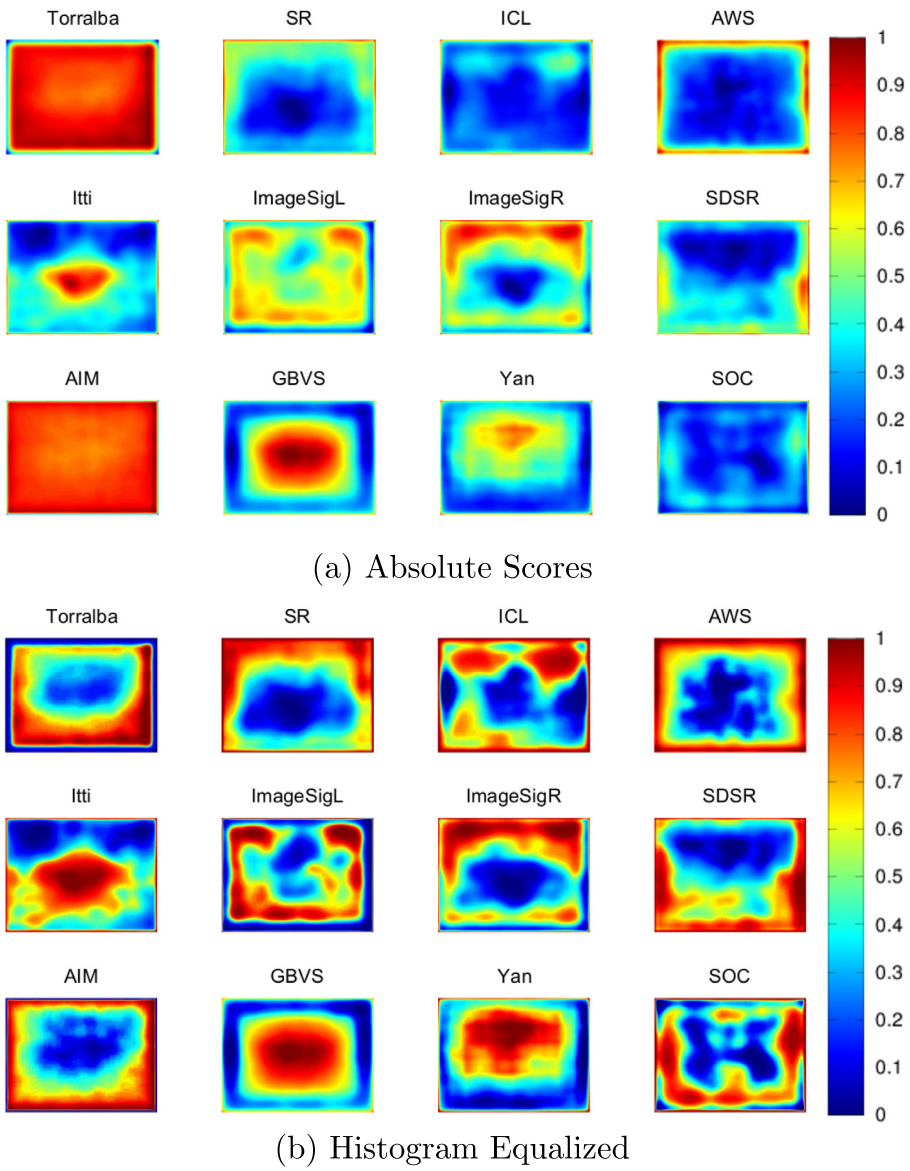


Fig. 3. Spatial bias profile for a number of algorithms relative to the output of all algorithms.

algorithm's ability to predict fixations. Beyond the implications for benchmarking results, the standard practice of blurring as a post-processing operation also raises a number of philosophical concerns. It is apparent that one purpose this operation serves, is to compensate for factors not captured by saliency models, including higher level cognition and visuomotor control. These factors do play a role in determining ground truth fixation data, and the extent to which blurring is capable of improving performance on standard benchmarks is model dependent and variable. However, that this operation may improve benchmark performance for one model more than another, is not related in any obvious way to the validity of a models characterization of stimulus saliency. This again suggests that fixation based evaluation strategies may be ill-suited to measuring the validity of models of visual saliency. Moreover, this type of analysis fails to account for additional important factors such as the variability in spatial extent of competition among stimuli (general or stimulus dependent), and the spatial extent of inter-stimulus interaction according to feature types (e.g. motion vs. orientation contrast). These issues are discussed in further detail in Sections 3.1.5 and 3.1.7 respectively.

2.2.3. Visualization

While much discussion is often devoted to quantitative metrics, qualitative comparison is typically presented with little thought or discussion of the implications of how one might receive a particular visualization of saliency output. Nevertheless, the presentation of a small number of sample images from each data set with saliency map output for each algorithm compared quantitatively is common practice, with performance characteristics of the algorithms concluded from the figures (see, for example, Zhang & Sclaroff, 2013). As mentioned in Section 2.1, in many ways it is the *relative* difference in saliency values that is of primary importance rather than the absolute difference between pixel saliency scores, yet there are many different ways to transform the saliency signal into an output image which drastically alters the qualitative appearance of the saliency signal but preserves the numeric ordering of pixel intensities.

As a demonstration of this phenomenon, the image shown in Fig. 4 was analyzed using AIM (Bruce & Tsotsos, 2006), DVA (Hou & Zhang, 2009), Spectral Salience (Schauerte & Stiefelhagen, 2012), and SUN (Zhang et al., 2008). The top row of Fig. 5 shows

Please cite this article in press as: Bruce, N. D. B., et al. On computational modeling of visual saliency: Examining what's right, and what's left. *Vision Research* (2015), http://dx.doi.org/10.1016/j.visres.2015.01.010

the direct output of each algorithm using the code available for download and smoothed using a common Gaussian kernel with a standard deviation of ten pixels. As can be seen in the images, it appears that Spectral Salience picks up only the bird, while DVA, SUN, and AIM each pick up progressively more of the background in addition to the primary salient target. However, a large part of this apparent disparity comes from the fact that AIM has not been normalized such that its minimum saliency value corresponds to zero. Cubing all the values in the map and then re-normalizing based on the new maximum (shown in the center row) amplifies the most salient pixels and maps the lower saliency values closer to zero, thereby causing the qualitative performance of AIM to shift to be much closer to that seen in the other algorithms in their raw output. The final row of Fig. 5 presents a comparison based on histogram equalization, wherein the spread of saliency values produced by each algorithm is mapped as closely as possible into the same space.

In all of these case, it is evident that the interpretation of output depends heavily on the contrast at which the output is presented. Given that visual comparison of outputs is often used as a means of comparing different algorithms, this point is important. In particular, it is unclear what sort of visualization is amenable to qualitative comparison given that the relationship of output contrast to perceptual similarity may be unpredictable and vary as a function of stimulus characteristics. At a minimum though, any comparisons of this variety should seek to present output for different algorithms based on a similar distribution of gray levels to minimize the potential for misleading conclusions concerning the appearance of saliency maps. Evaluation of this variety is also arguably counterproductive. In light of the discussion throughout the rest of this section, it is evident that the relation of observed gaze patterns to image content is more complex than can be captured by a topological representation of salience. For this reason, methods of the variety suggested in Section 3.1 arguably carry greater value in characterizing model characteristics or model similarity.

2.3. Physiological and motoric biases

Although it is beyond the scope of this paper to cover all possible confounding factors involved in human fixation allocation, it is nevertheless illustrative to point out a number of physiological and motoric factors which may initially seem irrelevant to the task of directing overt shifts of attention but which do, in fact, bias our attentional behavior. The highly heterogeneous nature of the retina means that visual processing varies continuously with increasing eccentricity (Strasburger, Rentschler, & Jnttner, 2011). Not only is there a general drop in acuity as one moves toward the periphery, but additional confounding effects begin to be introduced such as crowding (Bouma, 1970). Crowding refers to the manner in which nearby stimuli impede recognition and processing with respect to the normal performance associated with isolated stimuli at the same spatial location, and this can have a significant impact on the accuracy of saccades (Vlaskamp & Hooge, 2006).

Fig. 4. Sample image from the Judd/Torralba dataset used as input for Fig. 5.

Additionally, the motoric processing involved in preparing and executing saccades can influence the dynamics of visual processing. Rizzolatti et al. (1987) showed that cross-meridian shifts in attention, both horizontal and vertical, incurred an additional cost to stimulus response time over within-quadrant shifts. Although differences across the vertical meridian may perhaps be explained by a cost accrued by cross-hemispheric neural processing, such an explanation fails to account for the cost seen with respect to crossing the horizontal meridian. A more comprehensive theory suggests that the cost rests with the need to recruit a new set of muscles for cross-meridian movements, therefore indicating that purely motor considerations may have subtle but significant effects on the allocation of overt attention which should not be overlooked. Likewise, body and head position have also been shown to introduce attentional biases. Reed, Garza, and Roberts (2007) demonstrate a number of modulating effects on spatial attention, including prioritizing based on hand orientation and location.

While saliency models typically produce a topological representation of visual salience, it is also important to note that regions associated with saliency computation and those that coordinate eye movements and targeting of saccades are not necessarily overlapping. This includes neural activation associated with these processes, and also physical movement of the eyes. While this is also beyond the scope of the current work, it is important to acknowledge this additional source of complexity in assessing models of saliency computation through behavioral output that also necessarily involves regions associated with visuomotor control. More details on some of the specific implications of this in assessing models of visual saliency are discussed in Section 3.1.8.

2.4. Covert attention

In addition to eye movements being potentially confounded by the motoric biases discussed in the previous section, it is also important to note that fixation location does not necessarily represent the locus of visual attention. *Covert attention* refers to attentional allocation without an accompanying eye movement, and was first demonstrated by Helmholtz' experiments with brief flashes of light (von & Southall, 1925). Since that time the nature and role of covert attention in our overall visual experience has been greatly debated, with some claiming that covert attention is independent of overt saccadic shifts (Hunt & Kingstone, 2003) while others have posited a connection between covert attention and microsaccades (Hafed & Clark, 2002; Ko, Poletti, & Rucci, 2010).

Regardless of the true role of covert attention, its existence poses a difficulty for treating human fixation data as the ground truth for the allocation of human attention. Given that saccadic targeting has been demonstrated to sometimes be drawn to a medial location between two salient targets, or even to simply be prone to the occasional inaccuracy (Findlay, 1997), it is unclear how often covert attention is used to correct for a close but inaccurately placed saccade. Perhaps a fruitful avenue for improving saliency benchmarking datasets could be made in better understanding the relationship between microsaccades and covert attention. Until that time, it is important to acknowledge that without a reliable method for detecting covert attentional shifts, overt fixations serve only as an imperfect proxy for attentional targeting.

2.5. Context and Gist

Significant gains from a performance standpoint have been demonstrated in exploiting holistic scene structure (Oliva &

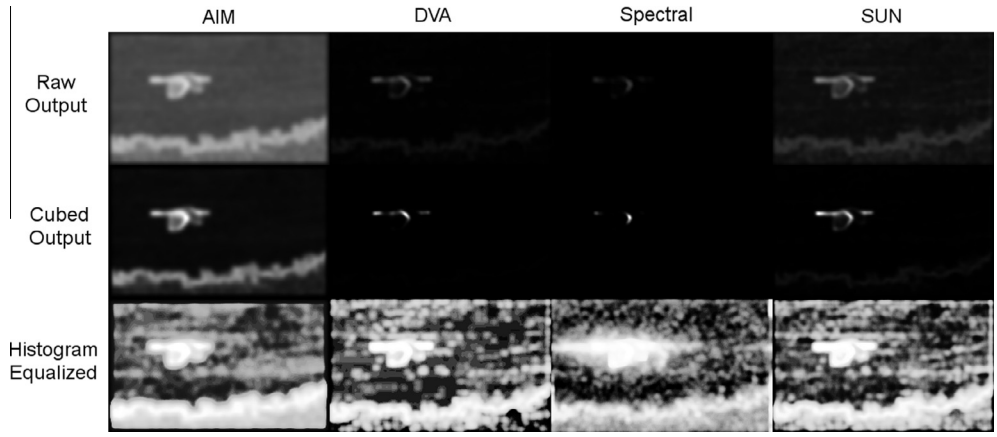


Fig. 5. Saliency maps produced for the image shown in Fig. 4 for the AIM, DVA, Spectral Salience, and SUN models. The top row shows the standard algorithm output after a modest (10 pixel standard deviation) smoothing kernel has been applied. The middle row shows the results of raising the above map to the third power and then re-normalizing. The bottom row shows the histogram equalization of the top row.

Torralba, 2001; Oliva & Torralba, 2006; Torralba et al., 2006; Oliva & Torralba, 2007) to augment prediction of fixation locations. In certain instances, context alone has been shown to predict fixated locations better than low-level measures of visual salience (Torralba et al., 2006). The strategy employed in such models assigns additional weight to a horizontal band within the scene based on coarse scene-level holistic receptive fields. Such a representation has also demonstrated significant capability towards scene classification suggesting that the underlying features present a reasonable compressed representation of the spatial envelope of a scene (Xiao et al., 2010).

Given that a strong center-bias exists within fixation data, one consideration that warrants further investigation is the extent to which conditional likelihoods (spatial position given scene structure) drive improved performance, versus such conditional likelihoods being derived from data where a strong center bias is present. This point is perhaps moot if one considers any spatial bias to be part of the general modeling problem. It is also perhaps reasonable to assume that the horizon of an outdoor scene has a prominent effect, that is distinct from one driven principally by center-bias (Foulsham & Underwood, 2007; Wang et al., 2011). That said, it's unclear whether a photograph where the subject of the photograph tends to be centered should be treated in a significantly different fashion from a set of images of generic outdoor landscapes versus a set of web pages, or video game images. If one is aiming towards a model that is consistent with human behavior in the broadest sense possible, the true test in regards to any spatial bias should lie in the ability of a model to determine the spatial bias that is present at the image level including center-bias, top-left and above the fold bias for a web page (Buscher, Cutrell, & Morris, 2009; Betz et al., 2010), or any other form of bias appearing in a digital representation of the world. This is also an important consideration insofar as humans having an active vision system is concerned. In the same fashion that there is content bias within an image, there may also exist similar bias in an active viewing context.

One might argue that the most natural test for digital media is explicit treatment of a dataset as-is, in the absence of *a priori* knowledge. This is true especially in light of the experimental results presented on the role of spatial, central or peripheral bias in this paper. Moreover, it is also worth noting that further progress may be had in placing a heavier emphasis on actual active vision systems, while also analyzing spatio-temporal and inter-saccade dynamics as part of the evaluation criteria (Herdman & Ryan, 2007; Einhäuser et al., 2007).

2.6. Top–down effects

Although the discussion in this paper is primarily concerned with saliency, stimulus driven behavior and bottom-up attention, it is nevertheless of value to acknowledge the role of top down effects in relation to visual salience. In natural settings, it is unclear how much a role stimulus salience plays in determining guidance of fixations in natural scenarios. There are a variety of experiments (Turano, Geruschat, & Baker, 2003; Jovancevic, Sullivan, & Hayhoe, 2006; Jovancevic-Misic & Hayhoe, 2009) that suggest basic feature contrast may have a relatively small influence in guiding gaze behavior in realistic scenarios. Additionally, the non-specific nature of *free viewing* may allow the observer to independently adopt a specific mode of viewing (Tattler, Baddeley, & Gilchrist, 2005). It is also the case that when a specific task definition is provided, this may have a relatively dominant role in gaze behavior (Land & McLeod, 2000; Land & Hayhoe, 2001; Aivar et al., 2005; Ballard & Hayhoe, 2009). It is therefore important to bear in mind the broader context surrounding gaze behavior, and how some of these task-related observations may factor into observed behavior even for cases where no specific task definition is provided.

3. On the biological plausibility of saliency models

Given the significant concerns with fixation based benchmarking that have been discussed, this section presents analysis grounded more heavily in visual psychophysics. One motivation for this, is to provide an alternative vantage point for assessing saliency models that is more aligned with the scope of computation captured by models of visual saliency. In particular, we aim to establish a base set of behavioral patterns that are pervasive in experimental paradigms in visual search, including search for an oddball stimulus, or pop-out. Some prior efforts have sought to characterize biological plausibility in examining the response of models to a set of patterns reminiscent of those appearing in behavioral studies (e.g. Borji, Shiti, & Itti, 2013). An example of typical patterns (a subset of those considered in Borji, Shiti, & Itti (2013)), are shown in Fig. 6.

One potential shortcoming of this evaluation strategy, which appears across many individual efforts in modeling visual saliency, is that evaluation often takes the form of a *proof by example*. Many demonstrations of model behavior in the computational saliency literature demonstrate output for a subset of such patterns to justify general statements on biological plausibility or inspiration. Importantly, failure on any of these examples from a theoretical

Please cite this article in press as: Bruce, N. D. B., et al. On computational modeling of visual saliency: Examining what's right, and what's left. *Vision Research* (2015), http://dx.doi.org/10.1016/j.visres.2015.01.010

or mechanistic standpoint implies the need to revisit, or justify the claim of a functional or neurally plausible characterization of visual saliency computation. Quantitative metrics that consider ensemble performance across a broad set of examples, also lacks sufficient detail to distinguish between implementation level differences, or failure in agreement with underlying theoretical substrates or expected patterns in model behavior. We therefore consider an alternative approach in this section in identifying basic behavioral observations that appear in the literature, and the associated implications for model structure. We have not measured specific models against these guidelines other than evaluation our own modeling work as an example. This is in part to avoid making assumptions about the intentions of authors of alternative models, but also due to the scale of evaluation that such an assessment would require across a broad set of models. It is our hope nevertheless, that this serves as a useful framework for characterizing model behavior with consideration to the behavioral psychophysics literature.

3.1. Towards an axiomatic set of model constraints

In the domain of saliency models, the emphasis in evaluation tends to be towards quantitative benchmarks that rely on fixation data. In the previous section, we establish that there are many nuances of a fixation based evaluation strategy that are problematic. Passing references to the biological plausibility of models are made, but this correspondence is often weak at best. One of the primary goals of this section is towards testing the alignment of model behavior with some of the fundamental principles that are implied by observations from visual psychophysics. It is worth noting that the claim of biological plausibility may vary in the specificity of correspondence between model properties and neural computation. In its weakest form, this might correspond to coarse-grained similarity with functional observations derived from visual psychophysics. In a stronger form, agreement with detailed properties of neural circuitry might be assumed. For this reason, this section also distinguishes between different degrees of assumed agreement with biology in examining conditions on models within the discussion. The base level that is assumed for much of this section, is agreement at a functional level. It is also established that even subject to the relatively weak conditions on biological plausibility that reside at a functional level, the majority of existing models fall short in some respects.

As discussed prior, the goal of the following discussion is distilling out important behavioral observations from the psychophysics literature towards determining implications for computational models. More specifically, the goal of this lies in translating patterns in behavioral psychology associated with visual search into a form closer to an *axiomatic* set of model constraints consistent with behavioral psychophysics. Given that the concept of saliency is akin to the case of an oddball search, we use the concept of *guidance* (Wolfe, 1994; Wolfe & Horowitz, 2004) in considering the broader visual search literature, given the similarity of this notion to what is typically sought from models of visual saliency, namely the degree of rapid and automatic guidance of attention to certain content in a stimulus array or image. In what follows, a number of factors important to visual saliency are discussed. For each of these factors, a trailing statement that summarizes their implications for computational modeling appears under the heading of *Modeling Implications*. Key points supporting these implications are emphasized in italic lettering. It is worth noting for completeness, that oddball paradigms in visual search comprise only a small subset of the broader pool of research relating to visual search. Although such experiments emphasize the stimulus driven characteristics of visual search, the relationship between stimulus and task remains a predominant influence on orienting of visual attention, especially in real-world scenarios. For this reason, a more comprehensive understanding of the problem requires similar analysis that includes top-down influences, and the role of task. This is beyond the scope of this manuscript, but presents an important area of emphasis for future research.

3.1.1. Conjunctions and search efficiency

Many early efforts in modeling attention and visual search include within their motivation the observation that some search tasks appear to be independent of the number of items within the display, while others seem to require a serial search of individual items (Treisman & Gelade, 1980; Nakayama et al., 1986; Luck & Hillyard, 1990; Zelinsky & Sheinberg, 1997). This separation has often been described in terms of stimuli defined by individual features, or those combined by conjunctions (i.e. a combination of *basic* features). Since that time, this view has broadened such that visual search typically described as being characterized by a continuum of search slopes, with search tasks ranging from very efficient to very inefficient (Wolfe, 1998; Duncan & Humphreys, 1989; Eckstein, 1998; Wolfe et al., 2011). In relating this point to saliency, features within an oddball search, or those described as eliciting pop-out might be clearly categorized as patterns that are salient. This also includes contrast between the features and their surround (Nothdurft, 1993; Nothdurft, 2006) as a critical component. However, in assuming this relationship, it also follows that saliency may be related to general visual search tasks such that *very efficient* searches are at one continuum of visual saliency, while *very inefficient* search behavior might be characterized as falling at the other

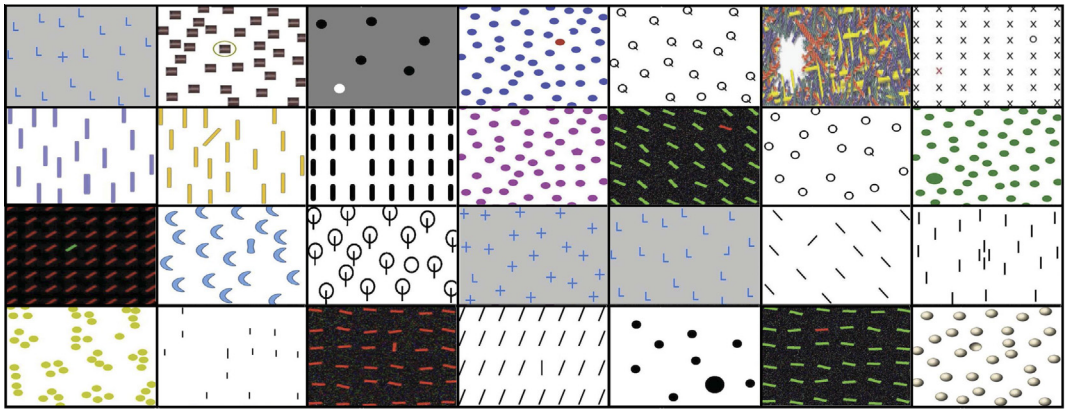


Fig. 6. Examples of common psychophysics-style patterns used in analyzing and assessing models of visual saliency for oddball search tasks.

Please cite this article in press as: Bruce, N. D. B., et al. On computational modeling of visual saliency: Examining what's right, and what's left. *Vision Research* (2015), http://dx.doi.org/10.1016/j.visres.2015.01.010

end of the spectrum, with the difference lying in the degree of *guidance*.

An alternative definition might lie in characterizing salient content as only those patterns that elicit overt (or covert) attention subject to a pre-attentive process that does not involve binding (e.g. not requiring recurrence binding (Tsotsos, 2011; Rothenstein & Tsotsos, 2014)). This does not necessarily preclude different degrees of saliency, but does imply additional conditions for model behavior.

One challenge in defining expected model behavior therefore lies in the defined scope of neural computation that falls under the umbrella of the term *saliency*. Given that this ambiguity lies at the more abstract level of where saliency fits within the overall scope of attention for which there are differing views (Li, 2002; Fecteau & Munoz, 2006), we instead focus on some basic behaviors that are related to feature conjunction and efficiency in considering implied model constraints. Perhaps the most important facet of this point, is the role of the base representational units that define the putative representation of stimuli in visual cortex. The importance of this point is discussed further under Modeling Implications. *Representational units* in the context of the discussion in this section therefore refer to the neurons, image filters, or feature maps that define the distributed representation of the input array.

Two of the central tenets in considering feature conjunctions from a modeling perspective are constancy of features, and involvement of surrounding features in defining feature contrast. If one assumes a base set of representational units (or feature maps), this *set of features remains fixed* in determining feature contrast. This raises questions for models that derive contrast from feature covariance, joint likelihoods (e.g. Avraham & Lindenbaum, 2010; Erdem & Erdem, 2013) or any other transformation that assumes pairwise or multi-feature conjunctions. The *central issue with such a contrast measure, is that this allows for contrast derived from arbitrary conjunctions that are not constrained by representational units and their connectivity*. For example, an "L" among "Ts" should not elicit a strong degree of contrast, but does so in some systems which employ an intermediate representation based on covariance or joint-likelihood methods. A second instance where the underlying feature set is variable is in the presence of a linear rotation of the coordinate space, e.g. with PCA applied at the level of pixels, or local features (e.g. Garcia-Diaz et al., 2012). In this instance, contrast is derived from a rotation of the coordinate space corresponding to base representational units, also allowing for arbitrary conjunctions to generate contrast.

The notion of local contrast is less nuanced, but it is worth stating for completeness that the relationship between a local pattern and its context should factor into the degree of saliency assigned to the local pattern. For example, a red singleton amongst green distractors should give a similar measure of saliency to a green singleton amongst red distractors. This is true of most but not all (e.g. Zhang et al., 2008) models.

Modeling Implications: (i) Saliency should not result from conjunctions of some set of basic features either by virtue of a direct mechanism for defining feature contrast, or implicitly through a transformation that alters the feature space in an unconstrained manner. (ii) Local feature contrast should be a determining factor in visual saliency. (iii) Cross-feature interaction might take place according to *a priori* statistical dependencies, but not arbitrarily.

3.1.2. Stimulus similarity

Saliency is inherently tied to contrast in feature space. Similarity between elements in a display within a visual search task (e.g. oddball search) therefore figures heavily into the determination of the relative salience of items in the display. The precise details of this variation may be relatively stimulus dependent, and this consideration requires strong assumptions of how content is represented neurally in expressing any specific constraints. With that said, there are certain behavioral patterns that are common across different types of stimuli that are important to explore. With respect to search for a target among distractors, the role of target-distractor similarity is well explored in some of the earlier studies targeting this interaction (Duncan & Humphreys, 1989; Pashler, 1987): *As target distractor distance increases, targets become easier to find. This trend occurs only up to a certain limit, at which point search efficiency is uniformly high and does not increase any further.* This has important implications for model structure in considering biological plausibility.

Also important to this discussion, is the nature of pooling of saliency across different stimulus dimensions. For example, an item in an array of stimuli that is unique in both orientation and luminance may carry a greater degree of salience than if this item is unique in only one of these dimensions (Nothdurft, 2000). This combination is typically additive in a non-linear fashion such that the salience produced by uniqueness across 2 feature dimensions is less than the sum of the increase in salience that one would predict from the 2 feature dimensions independently. This effect may also scale with the overlap shared by features across the 2 perceptual dimensions. Luminance and orientation contrast show greater independence than orientation and chromatic contrast in this respect.

An additional important point is that all content in the display (target, distractors, background) interacts to determine preferential guidance to certain stimuli. In visual search tasks, *increasing distractor heterogeneity tends to result in a more difficult search task* (Duncan & Humphreys, 1989; Nagy & Thomas, 2003; Rosenholtz, Nagy, & Bell, 2004). This is also a very important effect to consider as a test of biological plausibility, and appeals to the importance of guidance to stimuli being a function of interaction (e.g. mutual suppression) among all elements in the display, which might be weakened for an oddball target if distractor-distractor suppression is diminished by increased distractor heterogeneity.

Another factor of importance related to stimulus similarity is the spatial layout of stimuli within the visual field. Both local feature contrast, and more global feature contrast (or heterogeneity) may have an impact on stimulus salience. A relatively small difference in orientation in a relatively uniform field of stimuli leads to pop out, but may fail to do so in the presence of greater orientation heterogeneity in the extended surround (Nothdurft, 1993; Nothdurft, 1993). This has also been observed for direction of motion wherein an oddball in changes to direction of motion may elicit pop-out, but only in the case that the magnitude of direction shift among neighboring stimuli is relatively small. Salience as a function of spatial proximity is considered by Nothdurft, Gallant, and van Essen (1999) demonstrating the basic constraint that contextual modulation among stimuli depends critically on spatial proximity. This is examined in Nothdurft (2000), who demonstrates maximal salience of oriented lines with a line spacing slightly larger than the length of lines, and which diminishes to a nearly zero effect at a range of 2–4 degrees.

One challenge in examining target salience is that salience is often associated with highly super-threshold stimuli, making the measurement of salience a challenging task. Nothdurft (1993) presents a useful paradigm in which observers make a forced choice on the salience of two possible targets. In this fashion, the relative salience as a function of variation along different feature dimensions (e.g. luminance and orientation) may be related. This provides greater specificity in the tests that may be applied to models of saliency. In particular, the nature of this interaction, and the point of equivalence for different feature combinations present very specific targets for model assessment. An example of this type of assessment is presented by Gao (2008) (Fig. IV.5) in comparing predicted salience with human data on target

Please cite this article in press as: Bruce, N. D. B., et al. On computational modeling of visual saliency: Examining what's right, and what's left. *Vision Research* (2015), http://dx.doi.org/10.1016/j.visres.2015.01.010

salience as a function of varying target and background orientation, and luminance.

Modeling Implications: (i) Target-Distractor distance is correlated with saliency, with saturation beyond some target-distractor distance. (ii) For sufficient proximity in feature contrast, pop-out disappears (i.e. saliency may be negligible). (iii) Salience combines in an additive manner across feature dimensions in a non-linear manner. (iv) Increasing distractor heterogeneity should result in relatively diminished stimulus saliency. (v) Saliency depends critically on proximity in both feature space, and spatial position. Increased feature contrast, and increased variation of stimuli in the surround are opposing factors in driving stimulus salience. (vi) More specific quantitative analysis might be achieved by considering relative salience of two targets along differing feature dimensions, and their point of equivalence. This consideration requires some care in considering both properties of representational units in the model, and model mechanisms for determining salience.

3.1.3. Presence versus absence of features

An additional important effect lies in *the asymmetry that results from the absence of a feature* (e.g. a "−" among many "+" shaped elements) (Treisman & Gormican, 1988) *compared with the presence of a basic feature* (e.g. a "+" among many "−" shaped stimuli). If one views this situation at the level of neural representation, where all of the stimulus elements are divided into more primitive features, then activation of features with a bias for horizontal edge structure is relatively homogenous across the display, as is the case for vertical structure in the absent case. For the alternative (present) case, the vertical structure in the "+" elicits activation among units that is unique to a singular specific region of the stimulus array. While this is a rather simplistic case, any instance where swapping of target and distractor elements results in an asymmetric shift within the neural representation such that there is no strong competition on particular feature maps, might result in stronger guidance to the unique response defined by the feature level representation. Additional unintended asymmetries of this form have been demonstrated by Rosenholtz (2001), owing to the impact of background content on stimulus/distractor resulting in an asymmetric behavioral pattern.

Search asymmetries, including the presence/absence asymmetry also bear an interesting relationship to Weber's law (Treisman & Gormican, 1988). Treisman and Gormican (1988) posited a mechanism based on grouping of stimuli producing a pooled response. In instances where the pooled response for relevant target features is greater over the group containing the search target than the pooled response over distractor stimuli, the target is detected. This hypothesis also carries the more specific claim that the degree of background activation determines the just noticeable difference threshold as a function of a constant proportion of the background activation. This provides a more specific quantitative test of expected behavior for psychophysical patterns. An example of this form of analysis is provided by Gao and Vasconcelos (2009) based on the experiments of Treisman and Gormican (1988) wherein a target appears in an array of lines of common orientation with the target pattern differing in length. Swapping the target and distractor patterns results in a performance asymmetry. However, for a fixed set of distractors, performance is symmetric for targets that are longer or shorter than the distractors by the same magnitude.

Modeling Implications: i. The presence of a feature may result in a salient target, but absence of a feature should not except in the case that there is an explicit representational unit, or feature level representation tuned to such a pattern. ii. The background is part of the stimulus pattern. This is true of most implemented computational models, but not those that involve segmentation of the scene prior to determining salience. iii. The threshold for saliency is related to the background activation of the stimulus array. Contrast for saliency should scale with background activation according to Weber's law.

3.1.4. "Basic" asymmetries

An additional form of asymmetry that carries some similarity to the presence/absence case, is the class of asymmetries often referred to as *Basic Asymmetries* (Treisman & Gormican, 1988; Wolfe, 2001). Such asymmetries occur for stimulus configurations that are apparently symmetric in terms of how the properties of the stimuli are defined, but demonstrate a pronounced increase (or decrease) in ease of locating the target when target and distractors are swapped. A relatively simplistic example of this is the case of a 15 degree tilted bars amongst vertical bar distractors, which is much easier to spot than the converse case in which the vertical bar is the oddball. A similar pattern is also seen for stimuli such as reversed letters, or upside-down silhouettes of animals (Wolfe, 2001). One possibly relevant observation, is that asymmetries require an associated frame of reference, or coordinate system in their definition. Often this comes from a linguistic description of the conditions such as vertical vs. tilted, or forwards vs. backwards. While some explanations for such effects also reside at this relatively high level of abstraction (Wolfe et al., 1992), there is also a reasonable basis for a low-level explanation for some of these effects that is a consequence of how visual patterns are encoded in the brain. *Asymmetries in how different patterns may be represented, contrast thresholds and discriminative power of the corresponding neural representation may vary subject to two conditions that are symmetric at a descriptive level.* While we stop short of making any strong statements concerning the impetus for this class of observed asymmetries, we also note that a feature level representation that is biased towards the statistical structure of patterns provides sufficient conditions for observed asymmetric behavior (Bruce & Tsotsos, 2011).

Modeling Implications: A model should elicit asymmetric assignments of saliency in swapping some combinations of target/ distractors. This *may* be predicted by the specificity of the representation of each of the target and distractors across representational units.

3.1.5. Specific and non-specific suppression

While computational models often determine salience on the basis of local contrast for independent feature channels (e.g. a vertical bar surrounded by horizontal bars), there is a distinction made in the literature between local suppressive effects that are stimulus dependent (Type 2, which only suppress within feature channels), and suppression that does not depend on specific stimulus properties (Type 1, which can suppress across feature channels) (Nothdurft, Gallant, & van Essen, 1999). In populations of V1 cells, both types of suppression are observed. What is important in this observation, is that there appear to be distinct mechanisms that drive competition among neurons responding to different stimuli, and a need to recognize this within the context of computational models.

Modeling Implications: Models should consider both stimulus dependent, and stimulus independent mechanisms of suppression. Ideally, this should also be attached to specific hypotheses concerning lateral or recurrent mechanisms for suppressive interactions among neurons.

3.1.6. Visual field effects

Performance for a wide array of visual search tasks is not constant across the visual field, but instead may vary substantially according to where the stimulus appears. There are at least four different sources that are likely to contribute to observed behavior:

Please cite this article in press as: Bruce, N. D. B., et al. On computational modeling of visual saliency: Examining what's right, and what's left. *Vision Research* (2015), http://dx.doi.org/10.1016/j.visres.2015.01.010

Anisotropic statistics: Given that the structure of the world (both natural, and manmade), and our typical position within the world (on the ground, with sky or ceiling above, and ground/floor below), *there is inherent bias in the statistics of visual input across the visual field* (Bruce & Tsotsos, 2006; Nandy & Tjan, 2012). In a more general context, *this has implications for the sensitivity to basic features such as edge orientation* (van Essen, Newsome, & Maunsell, 1984), or the relative contribution of magnocellular vs. parvocellular pathways (Previc, 1990; McAnany & Levine, 2007) implying both spatial and temporal differences that vary according to position in the visual field. It is important at a low-level (i.e. ignoring context, or scene understanding) to consider the potential role of variation of representational units in the reference frame of the captured image in determining target visibility/ contrast, or visual salience.

Anisotropic sampling: *Visual acuity varies according to position within the visual field*, dropping off steeply as one moves outside of the fovea. It is evident that this will play a significant role in how visual content is sampled, and warrants consideration in the neural determination of saliency in addition to the interaction between said representation, and targeting mechanisms for fixations.

Anisotropic compression: While uneven representation of content within the visual field results from the drop-off in visual acuity due to foveation, there are instances where degradation of performance is beyond what might be explained purely due to this loss in acuity (Rosenholtz, Huang, & Ehinger, 2012). One possible explanation for the nature of this effect might be found in bias that results from an active vision system. To examine this point further, we begin by looking at bias across a large number of photographs drawn from the SUN image database (Xiao et al., 2010), all with a 4:3 aspect ratio and with a common size of 1024x768 (approx 14,000 images). For 25 (5 vertical $\times$ 5 horizontal) positions in the image, a 31 $\times$ 31 patch is extracted from each combination of the leftmost position to the rightmost (left, 25 percent, middle, 75 percent, right) position, and top to bottom of the image (left, 25 percent, middle, 75 percent, right). Subsequently, PCA is applied to each stack of patches (with the mean subtracted) to give a sense of the drop-off in variance across principal components in a position dependent fashion. In Fig. 7, the relative proportion of the total variance captured in the first 16 principal components is shown. It is clear that for photographs, *peripheral content may be compressed to a greater extent while retaining equal variance to the center. Given a biological vision system that includes fixations, and head movements it is plausible that one might also observe positional bias* in content within and outside of the fovea (Schumann et al., 2008). If the fidelity of the representation of content (in terms of variance retained) is uniform across the visual field, this would imply conditions for greater compression peripherally in addition to acuity effects.

Sensing Geometry: For biological vision systems, the visual stimulus is not the result of projection of the world onto a flat sensory array, but may be better characterized by a spherical approximation of the eye (or position of retinal cells). The implications of this point from the perspective of computational modeling are somewhat fuzzy. It is however important to note that some *visual field effects may be explained by a combination of both the statistically biased nature of the input, and also the geometry of the input system* (Pamplona, Triesch, & Rothkopf, 2013). This is important to bear in mind in translating behavioral observations to computational mechanisms.

Clutter, Grouping and Crowding: The spatial distribution of salient content may have some bearing on how content is sampled as reflected in visual fixations. Categorization of fixations as ambient vs. focal (Velichkovsky et al., 2005; Follet, Le Meur, & Baccino, 2011), or corresponding to a center-of-mass/gravity (Shuren, Jacobs, & Heilman, 1997; Zhou et al., 2066; Zelinsky, 2008) are

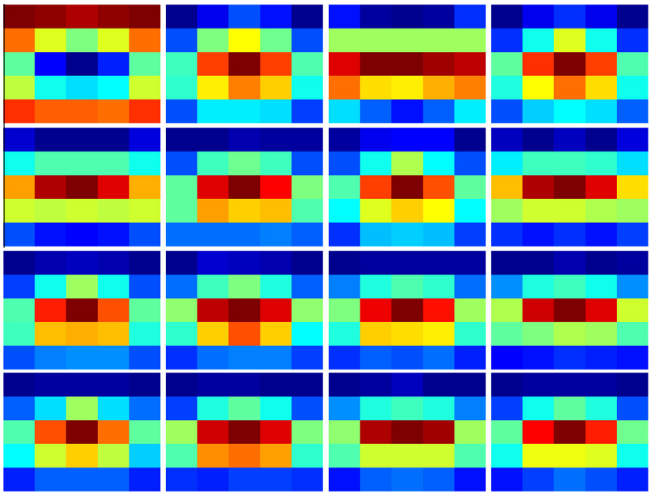


Fig. 7. Relative proportion of total variance for each window position, across the first 16 principal components. Each instance in the 4 $\times$ 4 grid shows a heatmap capturing the proportion of the total variance across all principal components for 25 evenly distributed positions within the image. Each sub-image is scaled independently, and demonstrates the relative proportions. Top row: 1st four principal components. 2nd row: PCs 5 to 8. 3rd row: PCs 9 to 12. 4th row: PCs 13 to 16.

motivated by behavioral observations that support this notion. Two additional concepts that are space-dependent are crowding (van den Berg, Cornelissen, & Roerdink, 2009) and clutter (Rosenholtz, Li, & Nakano, 2007; Henderson, Chanceaux, & Smith, 2009) respectively. *Close proximity of stimuli, or patterns in a psychophysical experiment can produce deficits in guidance towards such targets*. Evidence also suggests that *summary statistics may play an important role beyond more local measures of feature contrast* (Rosenholtz et al., 2012). In examining visual saliency from the perspective of fixations, it is again important to acknowledge that this is the product of both the inherent pattern salience, but also saccade control which may factor into the spatial distribution of fixation patterns. The demonstration of spatial drop-off in variation of image content moving to the periphery of an image shown in Fig. 7 as well as drop-off in visual acuity are also relevant to this interaction. While motor control in overt attention is of considerable importance in addressing visual saliency, a more detailed treatment of this subject matter is beyond the scope of the current work but remains a crucial consideration for future efforts to address.

Modeling Implications: Model behavior should vary according to position within the visual field. A fully comprehensive model might account for all of foveation, spatially varying feature specificity, and varying compression at the level of representational features. In addition, care should be taken in evaluation in recognition of fixation data being the product of mechanisms that may extend outside of the base determination of visual saliency including eye movements.

3.1.7. Complex features

While many efforts in visual saliency modeling are based primarily on relatively simple features that capture local contrast, edge content, color opponency, or motion, *there is some evidence to suggest that more complex features* such as shading (Kleffner & Ramachandran, 1992), luster (Wolfe & Franzel, 1988), curvature (Treisman & Gormican, 1988; Fahle, 1991; Wolfe, Yee, & Friedman-Hill, 1992), closure (Treisman & Souther, 1985; Elder & Zucker, 1994; Kovacs & Julesz, 1993; Williams & Julesz, 1992), or topology (Chen, 1982; Chen, 2005; Rubin & Kanwisher, 1985) *can provide significant guidance towards certain stimuli*. For example, a

Please cite this article in press as: Bruce, N. D. B., et al. On computational modeling of visual saliency: Examining what's right, and what's left. *Vision Research* (2015), http://dx.doi.org/10.1016/j.visres.2015.01.010

convex bump among concave *un-bumps* defined by shading can result in strong guidance. For many of these effects, it's unclear whether these are indeed strong evidence for *high-level* or complex feature guidance, or can be explained by differences in simple *basic* features. There also appears to be a strong preference for certain types of stimuli such as faces and text (Cerf, Frady, & Koch, 2009) that tend to draw gaze subject to a relatively rapid time course. Asymmetric effects in stimuli such as letters, or animal forms also suggests some support for guidance driven by more complex patterns. A model in which saliency, or feature contrast is expressed more broadly across layers, with guidance facilitated through recurrent modulation of representational units (Lee & Mumford, 2003) provides one plausible representation for the involvement of such complex features. In considering high-level features, it is important to bear in mind the interplay between attention and recognition. The role of attention in recognition and localization of objects largely precludes the possibility of complex feature detectors having rapid detection and localization as a primary function. It is therefore important to exercise caution in making strong assumptions concerning how such features fit within models of visual saliency. Moreover, more study is needed to better understand the impetus for the apparent rapid guidance elicited by some categories of complex patterns for inclusion of such behavior within models to be prudent.

Relevant to the possibility of complex features in saliency computation, are observations relating to timing associated with target salience, and extent of suppressive effects related to target salience. Salience from motion contrast may occur over a greater spatial distance than salience from orientation contrast (Nothdurft, 2000). There is some suggestion that this may be associated with differences in spatial properties associated with stimuli in MT vs. V1. Differences in time course are also observed that correlate with target salience. Salience driven by luminance contrast occurs subject to a more rapid time course than salience effects due to motion or orientation contrast (Nothdurft, 2000). Context dependent suppression from similar surrounding features is also delayed relative to target onset (Nothdurft, Gallant, & van Essen, 1999), suggesting feedback is involved in such suppression.

Modeling Implications: Saliency models might account for the apparent saliency of high-level features, although this depends on more general assumptions tied to the interplay among systems involved in coordinating visual attention. The notions of *basic feature* and *basic asymmetry* are also important here given the identification of some more complex features as *basic features*. One possible connection between these notions lies in the efficiency of representation as discussed in Bruce and Tsotsos (2011) which may parallel the efficiency in guidance in visual search.

3.1.8. Physiological considerations

There are additional considerations that are motivated by observed properties of neural computation in visual cortex, as well as cortical and subcortical regions involved in motor control of the eyes. With respect to properties of neurons, one important consideration is the receptive field size of units expressed across different layers of the visual cortex. While many models rely on representational units that have similarities to V1 in the patterns that are effective stimuli, *units in V1 are characterized by a relatively small receptive field size* (van Essen & Maunsell, 1983; Kastner et al., 2001; Bullier, 2001). This implies *limitations on the frequency bands that may be expressed by these units alone*. In contrast, many models employ filters, feature extraction or representational units that are defined in a manner that covers a large portion of a scene (and certainly more than V1 receptive fields). This calls into question the precise relationship between models and neuroanatomy, since it assumes either that a model is a functional characterization of behavior, or that higher visual areas need be involved due to receptive field size alone.

Given the important role of motor control in saccade targeting, there are additional points that deserve mentioning. One point of importance, lies in the *ballistic nature of saccades and their imperfect targeting* (Grossberg & Kuperstein, 2011), and the *involvement of corrective saccades* (Henson, 1978; Deubel, Wolf, & Hauske, 1982). These present additional noise sources in using fixation data to assess saliency models. In addition, dependencies in spatial position across saccades are also important to observed gaze patterns. This is further reason for having mechanisms for model assessment attached to a dynamic process while also considering dependencies across fixations in space, time and features.

Modeling Implications: The relative size of features versus visual input has important implications for the areas of visual cortex which are implicitly assumed to be involved in a saliency model. This also suggests further benefit from a stronger statement of neural correlates of a biologically motivated model grounded in anatomy, as well as involvement of higher visual areas and recurrence in the overall characterization. Fixation points carry additional noise from the targeting mechanism, which might be further quantified by the properties of the saccade (e.g. velocity, distance, latency, eccentricity) or its relation to the content being viewed (Bruce, 2014).

For an example of these principles measured against a specific model, the reader may refer to the Supplementary material. These observations provide a number of specific testable constraints that may be applied to saliency models. Of even greater value perhaps, are those observations that are quantitative in nature, and adhere to strict conditions confirmed by experimentation in psychophysics. A goal for further modeling efforts in this area, is the development of a more comprehensive benchmark of this variety, that focuses on specific quantitative measurements of model behavior subject to controlled sets of stimuli. A challenge in producing such a benchmark, are the simplifications inherent in producing an implemented model of visual salience. Gabor-like cells present an approximation of V1 type interpretive units, but fail to capture many of the nuances that are present in real cortical cells. This includes at least the distribution of scale space across cells, anisotropies in sensitivity to orientation and scale, end-stopping, curvature and complex cells, and general heterogeneity expressed in neurons within visual areas. A failure to match more specific quantitative results may arise due to a failure in assumptions related to the computational model of visual salience, or to concessions and imperfections associated with an assumed putative representation of cortical cells. For this reason, a grand challenge to ground model behavior in testable quantitative measures requires careful consideration of both saliency mechanisms and the properties of cells in visual cortex expressed in the models.

4. Discussion

It is clear that significant progress has been made in recent years towards the prediction of fixation patterns, and in better understanding the basis for neural computation that defines visual saliency. Many interesting facets of the problem have emerged through these efforts, and progress has also been made towards appropriate methods for evaluation. That said, there are also a number of remaining challenges in both benchmarking, and behavioral modeling that have been outlined. The introduction of a greater number and variety of datasets has helped to guide efforts in assessing models, but a relatively heavy focus on fixation based benchmarking has also conceivably steered some of the analysis away from the base underpinnings of models and their alignment with human vision. This is especially problematic given the variety

Please cite this article in press as: Bruce, N. D. B., et al. On computational modeling of visual saliency: Examining what's right, and what's left. *Vision Research* (2015), http://dx.doi.org/10.1016/j.visres.2015.01.010

of issues that imply that fixation based quantitative benchmarks provide a relatively poor test of the validity of underlying computational principles. We have made a preliminary step towards a more measured evaluation of properties tied to saliency computation in human vision, in focusing on the rudiments of careful analysis of model behavior from a systematic, and theoretical perspective. A goal for the future may therefore be consideration of a broader set of tools for examining the relationship between models and fixation data. This might include careful consideration of spatial bias, and heavier emphasis on dynamics or mechanisms that are involved in overt attention (e.g. in shifting from fixed spatial representations to sequential and dynamic metrics). This may also be bolstered by a careful and measured assessment of model behavior that is not purely quantitative. With that said, there are also a number of areas for extending beyond the scope of current efforts, and potential targets are identified and discussed in what follows.

4.1. Fruitful directions forward

Scale space and the spatial envelope: It is evident that scale plays an important role in determining how content within a scene is selected for viewing or attention. While we have discussed the importance of scale, there are evidently benefits to be had through further investigation of scale, including its interaction with scene composition, part vs. whole relationships and rapid scene understanding in general. While some efforts have addressed the role of context, or considered holistic representations to refine the saliency map, there is further benefit to be had in modeling targeted at mechanisms influencing spatial bias, or the scale of content of interest within a scene. Moreover, the simplistic nature of current methods for treating these problems within the saliency literature are likely to be counterproductive to advancing progress in this area.

The bigger picture: There is also evidently a need to move beyond the characterization of saliency as an isolated phenomenon towards incorporating its role within a larger system that might include any subset of active vision, spatio-temporal dynamics, inter saccadic dependencies in location/content, and the statistical structure of scanpaths. This also includes the roles of local and distal connections in modulating neural activity through normalization (Reynolds & Heeger, 2009; Carandini & Heeger, 2011; Coen-Cagli, Dayan, & Schwartz, 2012) or other mechanisms (Rothenstein & Tsotsos, 2014).

Reasoning about fixations: While some efforts have focused on characterizations such as ambient vs. focal saccades, or center-of-mass saccades, there is a great potential to apply reasoning towards determining factors that may contribute to a given fixation including the relative contribution of local contrast, scene structure, implicit behavioral biases or gestalt principles. One preliminary effort addressing this is directed at separating out the contribution that one might expect from scene gist from the overall distribution of fixation density (Bruce, 2014). A more comprehensive explanation of driving factors behind observed saccades may be had in modeling, especially in the presence of more data and more varied data.

Active vision and dynamics: While several efforts have examined saliency in a dynamic context (Najemnik & Geisler, 2005; Schumann et al., 2008; Einhäuser et al., 2009; Schneider et al., 2009; Foerster et al., 2012; Borji & Itti, 2013), there has been a heavy emphasis on modeling directed at predicting fixated locations for images and video presented on a static display. As we have discussed, there are many factors in overt attention that deny explanation without assuming a system that is non-stationary in its position and viewpoint. This includes explicit mechanisms for saccade targeting, interplay between overt and covert attention which might be understood in analyzing saccadic latency, and implications of time-varying visual input on fixation targets. Systems with an active vision component, e.g. (Tsotsos, 2011; Tsotsos and Kruijne, 2014), therefore present potentially significant gains in the computational understanding of saliency and attention. With respect to dynamical aspects of overt attention, there is also a dire need for benchmarking efforts to address the sequential and time-varying aspects of this process.

Localized versus distributed representation: While many efforts characterize saliency on the basis of distinct feature maps, contrast or conspicuity maps, and finally a *master* saliency map, an alternative lies in the feature contrast that defines saliency being expressed in a distributed fashion throughout a hierarchical neural representation (Lee & Mumford, 2003; Tsotsos et al., 1995). Such a characterization does well to capture the role of *complex* features or high-level factors that appear to elicit strong affinity as gaze targets in a rapid and automatic fashion. An alternative (e.g. if one assumes V1 as the locus of the saliency map (Li, 2002)) lies in the rapid recurrent feedback from higher visual layers to lower visual layers, including the interplay between faster magnocellular pathways and the slower feed-forward parvocellular activity (Bullier, 2001). This has also been examined in preliminary form (Shi, Bruce, & Tsotsos, 2011), however there are additional gains to be had towards a comprehensive understanding in considering the distinct regions of visual cortex and their role in driving saliency mechanisms.

A common theme of these targets, is extending beyond the composition of different feature representations combined with different measures of feature contrast towards a more comprehensive model that includes areas in visual cognition and computation that interact with saliency, including more general attention mechanisms and visual search behavior, control of eye movements and dynamics, and the role of scene understanding and higher visual areas including the distinct role of ventral and dorsal streams and their interaction.

5. Conclusions

In this paper, we have demonstrated some of the challenges faced in fixation based benchmarking for assessing models of visual saliency. This includes obstacles presented by spatial bias, edge and boundary effects, fragmentation in model abstraction and intent and varied benchmarking metrics. Benchmarking efforts have been useful, but there remains room for further efforts addressing some very specific issues for static positional fixation data. In particular, spatial bias continues to present a confound to empirical studies as does post-processing blur, and the role of active vision in models of overt attention has been underemphasized. A fundamental problem with fixation based benchmarking, is that the scope of computation characterized by models differs from mechanisms in neural computation that give rise to the ground truth fixation data. In moving forward, this may be remedied in part by considering a dynamic context that includes the sequential nature of fixations and mechanisms for visuomotor control. While the distinction between many existing models of visual saliency is focused on the measure of statistical distance that defines saliency, greater emphasis on the underlying features or representational units stands to provide further benefit to understanding from a biological standpoint. Finally, given variety in the purpose of saliency models, stronger statements on modeling goals and level of abstraction may help to clarify model intent, and also dictate the type of evaluation that is appropriate.

We have also examined a number of points motivated by the role of scale in characterizing visual saliency. Early efforts in the computer vision literature that sought *interest points* (Kadir & Brady, 2001; Mikolajczyk & Schmid, 2001) focused very heavily on the representation of scale space (Florack et al., 1992; Lindeberg, 1994). The results presented, along with the broader

Please cite this article in press as: Bruce, N. D. B., et al. On computational modeling of visual saliency: Examining what's right, and what's left. *Vision Research* (2015), http://dx.doi.org/10.1016/j.visres.2015.01.010

discussion in this paper also call for a greater understanding of the role of scale, and its interaction with scene composition, representation among higher visual areas, and the role of active vision and eye movements in the loop.

We have also identified a number of potentially fruitful directions forward which include further analytics at the level of raw fixation data e.g. (Bruce, 2014). Efforts that modify the task definition from a free viewing paradigm towards one that is more saliency driven (Koehler et al., 2014) are of significant utility in addressing this problem. This includes e.g. manual labeling of salient regions (Borji, Sihite, & Itti, 2013), or explicit manual selection of salient points (Koehler et al., 2014). Given the multi-factorial nature of gaze data, variety in data sources presents a means for evaluating saliency models subject to higher level cognition and also to tease apart the underlying processes that give rise to observed fixation patterns.

Acknowledgments

NB and SR acknowledge support from the Natural Sciences and Engineering Research Council of Canada and the University of Manitoba. JKT acknowledges support from the Canada Research Chairs program and the Natural Sciences and Engineering Research Council of Canada. CW and NF are supported by the Natural Sciences and Engineering Research Council of Canada.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.visres.2015.01.010.

References

Achanta, R., & Susstrunk, S. (2009). Saliency detection for content-aware image resizing. In: Image Processing (ICIP), 2009 16th IEEE International Conference on. IEEE (pp. 1005–1008).

Achanta, R., Estrada, F., Wils, P., & Süsstrunk, S. (2008). Salient region detection and segmentation. In *Computer Vision Systems* (pp. 66–75). Springer.

Aivar, M. P., Hayhoe, M. M., Chizk, C. L., & Mruczek, R. E. (2005). Spatial memory and saccadic targeting in a natural task. *Journal of Vision, 5*, 3.

Andreopoulos, A., & Tsotsos, J. K. (2012). On sensor bias in experimental methods for comparing interest-point, saliency, and recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*, 110–126.

Avidan, S., & Shamir, A. (2007). Seam carving for content-aware image resizing. In *ACM Transactions on Graphics (TOG)* (vol. 26, pp. 10). ACM.

Avraham, T., & Lindenbaum, M. (2010). Esaliency (extended saliency): Meaningful attention using stochastic image modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*, 693–708.

Ballard, D. H., & Hayhoe, M. M. (2009). Modelling the role of task in the control of gaze. *Visual Cognition, 17*, 1185–1204.

Betz, T., Kietzmann, T. C., Wilming, N., & König, P. (2010). Investigating task-dependent top-down effects on overt visual attention. *Journal of Vision, 10*, 15.

Borji, A., & Itti, L. (2013). Bayesian optimization explains human active search. In: Advances in neural information processing systems (pp. 55–63).

Borji, A., Sihite, D.N., & Itti, L. (2012). Salient object detection: A benchmark. In: Proc. European Conference on Computer Vision (ECCV), Florence, Italy.

Borji, A., Tavakoli, H.R., Sihite, D.N., & Itti, L. (2013c). Analysis of scores, datasets, and models in visual saliency prediction. In: 2013 IEEE International Conference on Computer Vision (ICCV), IEEE (pp. 921–928).

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

Borji, A., Sihite, D. N., & Itti, L. (2013). What stands out in a scene? A study of human explicit saliency judgment. *Vision Research, 91*, 62–77.

Bouma, H. (1970). Interaction effects in parafoveal letter recognition. *Nature, 226*, 177–178.

Bruce, N. D. (2014). Towards fine-grained fixation analysis: distilling out context dependence. In *Proceedings of the symposium on eye tracking research and applications* (pp. 99–102). ACM.

Bruce, N. D., & Tsotsos, J. K. (2006). Saliency based on information maximization. *Advances in Neural Information Processing Systems, 18*, 155–162.

Bruce, N. D., & Tsotsos, J. K. (2006). A statistical basis for visual field anisotropies. *Neurocomputing, 69*, 1301–1304.

Bruce, N. D., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision, 9*, 5.

Bruce, N. D., & Tsotsos, J. K. (2011). Visual representation determines search difficulty: Explaining visual search asymmetries. *Frontiers in Computational Neuroscience, 5*.

Bullier, J. (2001). Integrated model of visual processing. *Brain Research Reviews, 36*, 96–107.

Buscher, G., Cutrell, E., & Morris, M. R. (2009). What do you see when you're surfing? Using eye tracking to predict salient regions of web pages. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 21–30). ACM.

Butko, N.J., Zhang, L., Cottrell, G.W., & Movellan, J.R. (2008). Visual saliency model for robot cameras. In: Robotics and automation, 2008. ICRA 2008. IEEE International conference on IEEE (pp. 2398–2403).

Carandini, M., & Heeger, D. J. (2011). Normalization as a canonical neural computation. *Nature Reviews Neuroscience, 13*, 51–62.

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research, 51*, 1484–1525.

Cerf, M., Frady, E. P., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision, 9*, 10.

Chang, C.-K., Siagian, C., & Itti, L. (2010). Mobile robot vision navigation & localization using gist and saliency. In: Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ international conference on, IEEE (pp. 4147–4154).

Chen, L. (1982). Topological structure in visual perception. *Science, 218*, 699–700.

Chen, L. (2005). The topological approach to perceptual organization. *Visual Cognition, 12*, 553–637.

Chen, X., & Zelinsky, G. J. (2006). Real-world visual search is dominated by top–down guidance. *Vision Research, 46*, 4118–4133.

Coen-Cagli, R., Dayan, P., & Schwartz, O. (2012). Cortical surround interactions and perceptual salience via natural scene statistics. *PLoS Computational Biology, 8*, e1002405.

Culhane, S. M., & Tsotsos, J. K. (1992). An attentional prototype for early vision. In *Computer Vision ECCV'92* (pp. 551–560). Springer.

Deubel, H., Wolf, W., & Hauske, G. (1982). Corrective saccades: Effect of shifting the saccade goal. *Vision Research, 22*, 353–364.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review, 96*, 433.

Eckstein, M. P. (1998). The lower visual search efficiency for conjunctions is due to noise and not serial attentional processing. *Psychological Science, 9*, 111–118.

Einhäuser, W., Moeller, G. U., Schumann, F., Conradt, J., Vockeroth, J., Bartl, K., Schneider, E., & König, P. (2009). Eye–head coordination during free exploration in human and cat. *Annals of the New York Academy of Sciences, 1164*, 353–366.

Einhäuser, W., Schumann, F., Bardins, S., Bartl, K., Böning, G., Schneider, E., & König, P. (2007). Human eye–head co-ordination in natural exploration. *Network: Computation in Neural Systems, 18*, 267–297.

Elder, J., & Zucker, S. (1994). A measure of closure. *Vision Research, 34*, 3361–3369.

Erdem, E., & Erdem, A. (2013). Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of Vision, 13*, 11.

Fahle, M. (1991). Parallel perception of vernier offsets, curvature, and chevrons in humans. *Vision Research, 31*, 2149–2184.

Fecteau, J. H., & Munoz, D. P. (2006). Salience, relevance, and firing: A priority map for target selection. *Trends in Cognitive Sciences, 10*, 382–390.

Findlay, J. M. (1997). Saccade target selection during visual search. *Vision Research, 37*, 617–631.

Florack, L. M., ter Haar Romeny, B. M., Koenderink, J. J., & Viergever, M. A. (1992). Scale and the differential structure of images. *Image and Vision Computing, 10*, 376–388.

Foerster, R. M., Carbone, E., Koesling, H., & Schneider, W. X. (2012). Saccadic eye movements in the dark while performing an automatized sequential high-speed sensorimotor task. *Journal of Vision, 12*, 8.

Follet, B., Le Meur, O., & Baccino, T. (2011). New insights into ambient and focal visual fixations using an automatic classification algorithm. *i-Perception, 2*, 592.

Foulsham, T., & Underwood, G. (2007). How does the purpose of inspection influence the potency of visual salience in scene perception? *Perception-London, 36*, 1123.

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision, 8*, 6.

Frintrop, S., Backer, G., & Rome, E. (2005). Goal-directed search with a top–down modulated computational attention system. *Pattern Recognition* (pp. 117–124). Springer.

Frintrop, S., Jensfelt, P., & Christensen, H. (2007). Simultaneous robot localization and mapping based on a visual attention system. In *Attention in cognitive systems. Theories and systems from an interdisciplinary viewpoint* (pp. 417–430). Springer.

Frintrop, S., Rome, E., & Christensen, H. I. (2010). Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception (TAP), 7*, 6.

Gao, D. (2008). A discriminant hypothesis for visual saliency: computational principles, biological plausibility and applications in computer vision, ProQuest.

Gao, D., Mahadevan, V., & Vasconcelos, N. (2008). The discriminant center-surround hypothesis for bottom-up saliency. In: Advances in neural information processing systems (pp. 497–504).

Gao, D., & Vasconcelos, N. (2009). Decision-theoretic saliency: Computational principles, biological plausibility, and implications for neurophysiology and psychophysics. *Neural Computation, 21*, 239–271.

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

Green, D. M., Swets, J. A., et al. (1966). *Signal detection theory and psychophysics* (vol. 1). New York: Wiley.

Grossberg, S., & Kuperstein, M. (2011). *Neural dynamics of adaptive sensory-motor control: Ballistic eye movements volume 30*. Elsevier.

Hafed, Z. M., & Clark, J. J. (2002). Microsaccades as an overt measure of covert attention shifts. *Vision Research, 42*, 2533–2545.

Halverson, T., & Hornof, A. J. (2007). A minimal model for predicting visual search in human–computer interaction. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 431–434). ACM.

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. In: Advances in neural information processing systems (pp. 545–552).

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences, 9*, 188–194.

Henderson, J. M., Chanceaux, M., & Smith, T. J. (2009). The influence of clutter on real-world scene search: Evidence from search efficiency and eye movements. *Journal of Vision, 9*, 32.

Henderson, J.M., Brockmole, J.R., Castelhano, M.S., & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. Eye movements: A window on mind and brain (pp. 537–562).

Henson, D. (1978). Corrective saccades: Effects of altering visual feedback. *Vision Research, 18*, 63–67.

Herdman, A. T., & Ryan, J. D. (2007). Spatio-temporal brain dynamics underlying saccade execution, suppression, and error-related feedback. *Journal of Cognitive Neuroscience, 19*, 420–432.

Hopfinger, J., Buonocore, M., & Mangun, G. (2000). The neural mechanisms of top-down attentional control. *Nature Neuroscience, 3*, 284–291.

Hou, X., & Zhang, L. (2007). Saliency detection: A spectral residual approach. In: IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07, IEEE (pp. 1–8).

Hou, X., & Zhang, L. (2009). Dynamic visual attention: Searching for coding length increments. In: Advances in neural information processing systems (pp. 681–688).

Hou, X., Harel, J., & Koch, C. (2012). Image signature: Highlighting sparse salient regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*, 194–201.

Hunt, A. R., & Kingstone, A. (2003). Covert and overt voluntary attention: linked or independent? *Cognitive Brain Research, 18*, 102–105.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience, 2*, 194–203.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*, 1254–1259.

Jovancevic-Misic, J., & Hayhoe, M. (2009). Adaptive gaze control in natural environments. *The Journal of Neuroscience, 29*, 6234–6238.

Jovancevic, J., Sullivan, B., & Hayhoe, M. (2006). Control of attention and gaze in complex environments. *Journal of Vision, 6*, 9.

Judd, T. (2011). Understanding and predicting where people look in images. Ph.D. thesis, Massachusetts Institute of Technology.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In: Computer vision, 2009 IEEE 12th international conference on, IEEE (pp. 2106–2113).

Judd, T., Durand, F., & Torralba, A. (2012a). A benchmark of computational models of saliency to predict human fixations, technical report, Massachusetts Institute of Technology.

Judd, T., Durand, F., & Torralba, A. (2012b). A benchmark of computational models of saliency to predict human fixations. In: MIT, technical report.

Kadir, T., & Brady, M. (2001). Saliency, scale and image description. *International Journal of Computer Vision, 45*, 83–105.

Kastner, S., De Weerd, P., Pinsk, M. A., Elizondo, M. I., Desimone, R., & Ungerleider, L. G. (2001). Modulation of sensory suppression: Implications for receptive field sizes in the human visual cortex. *Journal of Neurophysiology, 86*, 1398–1411.

Kendall, M.G. et al. (1946). The advanced theory of statistics. The advanced theory of statistics.

Kim, Y., & Varshney, A. (2006). Saliency-guided enhancement for volume visualization. *IEEE Transactions on Visualization and Computer Graphics, 12*, 925–932.

Kleffner, D. A., & Ramachandran, V. S. (1992). On the perception of shape from shading. *Perception and Psychophysics, 52*, 18–36.

Klein, D.A., & Frintrop, S. (2011). Center-surround divergence of feature statistics for salient object detection. In: 2011 IEEE International Conference on Computer Vision (ICCV). IEEE (pp. 2214–2219).

Koch, C., & Ullman, S. (1987). Shifts in selective visual attention: Towards the underlying neural circuitry. In *Matters of Intelligence* (pp. 115–141). Springer.

Koehler, K., Guo, F., Zhang, S., & Eckstein, M. P. (2014). What do saliency models predict? *Journal of Vision, 14*, 14.

Ko, H., Poletti, M., & Rucci, M. (2010). Microsaccades precisely relocate gaze in a high visual acuity task. *Nature Neuroscience, 13*, 1549–1553.

Kovacs, I., & Julesz, B. (1993). A closed curve is much more than an incomplete one: Effect of closure in figure-ground segmentation. *Proceedings of the National Academy of Sciences, 90*, 7495–7497.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29*, 1–27.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 79–86.

Land, M. F., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research, 41*, 3559–3565.

Land, M. F., & McLeod, P. (2000). From eye movements to actions: How batsmen hit the ball. *Nature Neuroscience, 3*, 1340–1345.

Lee, D. K., Itti, L., Koch, C., & Braun, J. (1999). Attention activates winner-take-all competition among visual filters. *Nature Neuroscience, 2*, 375–381.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *JOSA A, 20*, 1434–1448.

Lee, C. H., Varshney, A., & Jacobs, D. W. (2005). Mesh saliency. In *ACM transactions on graphics (TOG)* (vol. 24, pp. 659–666). ACM.

Le Meur, O., & Baccino, T. (2013). Methods for comparing scanpaths and saliency maps: Strengths and weaknesses. *Behavior Research Methods, 45*, 251–266.

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences, 6*, 9–16.

Lindeberg, T. (1994). Scale-space theory: A basic tool for analyzing structures at different scales. *Journal of Applied Statistics, 21*, 225–270.

Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., & Shum, H.-Y. (2011). Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*, 353–367.

Longhurst, P., Debattista, K., & Chalmers, A. (2006). A gpu based saliency map for high-fidelity selective rendering. In *Proceedings of the 4th international conference on computer graphics, virtual reality, visualisation and interaction in Africa* (pp. 21–29). ACM.

Luck, S. J., & Hillyard, S. A. (1990). Electrophysiological evidence for parallel and serial processing during visual search. *Perception and Psychophysics, 48*, 603–617.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Henry Hold and Co., Inc..

McAnany, J. J., & Levine, M. W. (2007). Magnocellular and parvocellular visual pathway contributions to visual field anisotropies. *Vision Research, 47*, 2327–2336.

Mikolajczyk, K., & Schmid, C. (2001). Indexing based on scale invariant interest points. In: Proceedings eighth IEEE international conference on computer vision, 2001. ICCV 2001, IEEE, vol. 1.

Nagy, A. L., & Thomas, G. (2003). Distractor heterogeneity, attention, and color in visual search. *Vision Research, 43*, 1541–1552.

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature, 434*, 387–391.

Nakayama, K., Silverman, G. H., et al. (1986). Serial and parallel processing of visual feature conjunctions. *Nature, 320*, 264–265.

Nandy, A. S., & Tjan, B. S. (2012). Saccade-confounded image statistics explain visual crowding. *Nature Neuroscience, 15*, 463–469.

Nothdurft, H.-C. (1993). The conspicuousness of orientation and motion contrast. *Spatial Vision, 7*, 341–366.

Nothdurft, H.-C. (1993). Saliency effects across dimensions in visual search. *Vision Research, 33*, 839–844.

Nothdurft, H.-C. (1993). Saliency effects across dimensions in visual search. *Vision Research, 33*, 839–844.

Nothdurft, H.-C. (2000). Salience from feature contrast: Additivity across dimensions. *Vision Research, 40*, 1183–1201.

Nothdurft, H.-C. (2000). Salience from feature contrast: Temporal properties of saliency mechanisms. *Vision Research, 40*, 2421–2435.

Nothdurft, H.-C. (2000). Salience from feature contrast: Variations with texture density. *Vision Research, 40*, 3181–3200.

Nothdurft, H.-C. (2006). Salience and target selection in visual search. *Visual Cognition, 14*, 514–542.

Nothdurft, H.-C., Gallant, J. L., & van Essen, D. C. (1999). Response modulation by texture surround in primate area v1: Correlates of popout under anesthesia. *Visual Neuroscience, 16*, 15–34.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42*, 145–175.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research, 155*, 23–36.

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences, 11*, 520–527.

Pamplona, D., Triesch, J., & Rothkopf, C. (2013). Power spectra of the natural input to the visual system. *Vision Research, 83*, 66–75.

Pashler, H. (1987). Detecting conjunctions of color and form: Reassessing the serial search hypothesis. *Perception and Psychophysics, 41*, 191–201.

Peters, R.J., & Itti, L. (2007). Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In: IEEE conference on computer vision and pattern recognition, 2007. CVPR'07, IEEE (pp. 1–8).

Previc, F. H. (1990). Functional specialization in the lower and upper visual fields in humans: Its ecological origins and neurophysiological implications. *Behavioral and Brain Sciences, 13*, 519–542.

Rahman, S., Rochan, M., Wang, Y., & Bruce, N.D. (2014). Examining visual saliency prediction in naturalistic scenes. In: ICIP 2014. Proceedings IEEE international conference on image processing (pp. 0–0). IEEE, vol. 0.

Reed, C. L., Garza, J. P., & Roberts, R. J. (2007). The influence of the body and action on spatial attention. *Attention in Cognitive Systems, 4840*, 42–58.

Reynolds, J. H., & Heeger, D. J. (2009). The normalization model of attention. *Neuron, 61*, 168–185.

Please cite this article in press as: Bruce, N. D. B., et al. On computational modeling of visual saliency: Examining what's right, and what's left. *Vision Research* (2015), http://dx.doi.org/10.1016/j.visres.2015.01.010

Riche, N., Duvinage, M., Mancas, M., Gosselin, B., & Dutoit, T. (2013). Saliency and human fixations: State-of-the-art and study of comparison metrics. In: 2013 IEEE International Conference on Computer Vision (ICCV), IEEE (pp. 1153–1160).

Rizzolatti, G., Riggio, L., Dascola, I., & Umilta, C. (1987). Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention. *Neuropsychologia, 25*, 31–40.

Rosenholtz, R. (2001). Search asymmetries? What search asymmetries? *Perception and Psychophysics, 63*, 476–489.

Rosenholtz, R., Huang, J., & Ehinger, K. A. (2012). Rethinking the role of top-down attention in vision: Effects attributable to a lossy representation in peripheral vision. *Frontiers in Psychology, 3*.

Rosenholtz, R., Huang, J., Raj, A., Balas, B. J., & Ilie, L. (2012). A summary statistic representation in peripheral vision explains visual search. *Journal of Vision, 12*, 14.

Rosenholtz, R., Li, Y., & Nakano, L. (2007). Measuring visual clutter. *Journal of Vision, 7*, 17.

Rosenholtz, R., Nagy, A. L., & Bell, N. R. (2004). The effect of background color on asymmetries in color search. *Journal of Vision, 4*, 9.

Rothenstein, A. L., & Tsotsos, J. K. (2014). Attentional modulation and selection – An integrated approach. *PloS One, 9*, e99681.

Rubin, J. M., & Kanwisher, N. (1985). Topological perception: Holes in an experiment. *Attention, Perception, and Psychophysics, 37*, 179–180.

Rubner, Y., Tomasi, C., & Guibas, L.J. (1998). A metric for distributions with applications to image databases. In: Sixth international conference on computer vision, 1998, IEEE (pp. 59–66).

Schauerte, B., & Stiefelhagen, R. (2012). Quaternion-based spectral saliency detection for eye fixation prediction. *Proc. of European conference on computer vision*. Springer.

Schneider, E., Villgrattner, T., Vockeroth, J., Bartl, K., Kohlbecher, S., Bardins, S., Ulbrich, H., & Brandt, T. (2009). EyeSeeCam: An eye movement–driven head camera for the examination of natural visual exploration. *Annals of the New York Academy of Sciences, 1164*, 461–467.

Schumann, F., Einhäuser, W., Vockeroth, J., Bartl, K., Schneider, E., & König, P. (2008). Salient features in gaze-aligned recordings of human visual input during free exploration of natural environments. *Journal of Vision, 8*, 12.

Schütz, A. C., Trommershäuser, J., & Gegenfurtner, K. R. (2012). Dynamic integration of information about salience and value for saccadic eye movements. *Proceedings of the National Academy of Sciences, 109*, 7547–7552.

Seo, H. J., & Milanfar, P. (2009). Static and space–time visual saliency detection by self-resemblance. *Journal of Vision, 9*, 15.

Shi, X., Bruce, N.D., & Tsotsos, J.K. (2011). Fast, recurrent, attentional modulation improves saliency representation and scene recognition. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE computer society conference on, IEEE (pp. 1–8).

Shuren, J. E., Jacobs, D. H., & Heilman, K. M. (1997). The influence of center of mass effect on the distribution of spatial attention in the vertical and horizontal dimensions. *Brain and Cognition, 34*, 293–300.

Strasburger, H., Rentschler, I., & Jnttner, M. (2011). Peripheral vision and pattern recognition: A review. *Journal of Vision, 11*, 1–82.

Suh, B., Ling, H., Bederson, B. B., & Jacobs, D. W. (2003). Automatic thumbnail cropping and its effectiveness. In *Proceedings of the 16th annual ACM symposium on user interface software and technology* (pp. 95–104). ACM.

Tatler, B. W. (2009). Current understanding of eye guidance. *Visual Cognition, 17*, 777–789.

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research, 45*, 643–659.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 113*, 766.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*, 97–136.

Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review, 95*, 15.

Treisman, A., & Souther, J. (1985). Search asymmetry: A diagnostic for preattentive processing of separable features. *Journal of Experimental Psychology: General, 114*, 285.

Tseng, P.-H., Carmi, R., Cameron, I. G., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision, 9*, 4.

Tsotsos, J. K. (2011). *A computational perspective on visual attention*. MIT Press.

Tsotsos, J. K., & Kruijne, W. (2014). Cognitive programs: software for attention's executive. *Frontiers in psychology, 5*, 1260. http://dx.doi.org/10.3389/fpsyg.2014.01260.

Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y., Davis, N., & Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial Intelligence, 78*, 507–545 (Special Volume on Computer Vision).

Turano, K. A., Geruschat, D. R., & Baker, F. H. (2003). Oculomotor strategies for the direction of gaze tested with a real-world activity. *Vision Research, 43*, 333–346.

van den Berg, R., Cornelissen, F. W., & Roerdink, J. B. (2009). A crowding model of visual clutter. *Journal of Vision, 9*, 24.

van Essen, D. C., & Maunsell, J. H. (1983). Hierarchical organization and functional streams in the visual cortex. *Trends in Neurosciences, 6*, 370–375.

van Essen, D. C., Newsome, W. T., & Maunsell, J. H. (1984). The visual field representation in striate cortex of the macaque monkey: Asymmetries, anisotropies, and individual variability. *Vision Research, 24*, 429–448.

Velichkovsky, B. M., Joos, M., Helmert, J. R., & Pannasch, S. (2005). Two visual systems and their eye movements: Evidence from static and dynamic scene perception. In *Proceedings of the XXVII conference of the cognitive science society* (pp. 2283–2288). NJ: Lawrence Erlbaum Mahwah.

Vlaskamp, B. N., & Hooge, I. T. (2006). Crowding degrades saccadic search performance. *Vision Research, 46*, 417–425.

von Helmholtz, H., & translated by James P.C. Southall (1925). Treatise on physiological optics. III. The perceptions of vision. (Translated from the 3rd German ed.), Dover.

Wang, M., Konrad, J., Ishwar, P., Jing, K., & Rowley, H. (2011). Image saliency: From intrinsic to extrinsic context. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE conference on, IEEE (pp. 417–424).

Williams, D., & Julesz, B. (1992). Perceptual asymmetry in texture perception. *Proceedings of the National Academy of Sciences, 89*, 6531–6534.

Winkler, S., & Ramanathan, S. (2013). Overview of eye tracking datasets. In: QoMEX (pp. 212–217).

Witkin, A.P. (1984). Scale-space filtering: A new approach to multi-scale description. In: Acoustics, speech, and signal processing, IEEE international conference on ICASSP, vol. 9 (pp. 150–153).

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic Bulletin and Review, 1*, 202–238.

Wolfe, J. M. (1998). Visual search. In H. Pashler (Ed.), *Attention*. Psychology Press.

Wolfe, J. M. (1998). What can 1 million trials tell us about visual search? *Psychological Science, 9*, 33–39.

Wolfe, J. M. (2001). Asymmetries in visual search: An introduction. *Perception and Psychophysics, 63*, 381–389.

Wolfe, J. M., & Franzel, S. L. (1988). Binocularity and visual search. *Perception and Psychophysics, 44*, 81–93.

Wolfe, J. M., Friedman-Hill, S. R., Stewart, M. I., & O'Connell, K. M. (1992). The role of categorization in visual search for orientation. *Journal of Experimental Psychology: Human Perception and Performance, 18*, 34.

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience, 5*, 495–501.

Wolfe, J. M., Võ, M. L.-H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Sciences, 15*, 77–84.

Wolfe, J. M., Yee, A., & Friedman-Hill, S. R. (1992). Curvature is a basic feature for visual search tasks. *Perception-London-, 21*. 465–465.

Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., & Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In: Computer vision and pattern recognition (CVPR), 2010 IEEE conference on, IEEE (pp. 3485–3492).

Yan, J., Liu, J., Li, Y., Niu, Z., & Liu, Y. (2010). Visual saliency detection via rank-sparsity decomposition. In: Image processing (ICIP), 2010 17th IEEE international conference on, IEEE (pp. 1089–1092).

Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review, 115*, 787.

Zelinsky, G. J., & Sheinberg, D. L. (1997). Eye movements during parallel–serial visual search. *Journal of Experimental Psychology: Human Perception and Performance, 23*, 244.

Zhang, J., & Sclaroff, S. (2013). Saliency detection: A boolean map approach. In: Proc. of the IEEE international conference on computer vision.

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision, 8*, 32.

Zhou, X., Chu, H., Li, X., & Zhan, Y. (2006). Center of mass attracts attention. *Neuroreport, 17*, 85–88.

Ziman, J. M. (1969). Information, communication, knowledge. *Nature, 224*, 318–324.

Please cite this article in press as: Bruce, N. D. B., et al. On computational modeling of visual saliency: Examining what's right, and what's left. *Vision Research* (2015), http://dx.doi.org/10.1016/j.visres.2015.01.010