ARTICLE IN PRESS

Vision Research xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Vision Research

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

Towards the quantitative evaluation of visual attention models

Z. Bylinskii [a,b,*,1], E.M. DeGennaro [c,1], R. Rajalingham [d,1], H. Ruda [e,1], J. Zhang [f,g,1], J.K. Tsotsos [c,d,h]

[a] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge 02141, USA
[b] Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge 02141, USA
[c] McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge 02141, USA
[d] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge 02141, USA
[e] Computational Vision Laboratory, Department of Communication Sciences and Disorders, Northeastern University, Boston 02115, USA
[f] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[g] Visual Attention Lab, Brigham and Women's Hospital, Cambridge, MA 02139, USA
[h] Electrical Engineering and Computer Science, Centre for Vision Research, York University, Toronto M3J 1P3, Canada

ARTICLE INFO

Article history:
Received 1 August 2014
Received in revised form 15 March 2015
Available online xxxx

Keywords:
Opinion
Visual attention
Computational models
Benchmark datasets
Evaluation
Model taxonomy

ABSTRACT

Scores of visual attention models have been developed over the past several decades of research. Differences in implementation, assumptions, and evaluations have made comparison of these models very difficult. Taxonomies have been constructed in an attempt at the organization and classification of models, but are not sufficient at quantifying which classes of models are most capable of explaining available data. At the same time, a multitude of physiological and behavioral findings have been published, measuring various aspects of human and non-human primate visual attention. All of these elements highlight the need to integrate the computational models with the data by (1) operationalizing the definitions of visual attention tasks and (2) designing benchmark datasets to measure success on specific tasks, under these definitions. In this paper, we provide some examples of operationalizing and benchmarking different visual attention tasks, along with the relevant design considerations.

1. Introduction

Several decades of experimental research have uncovered a variety of neural and behavioral phenomena associated with visual attention. Physiological and brain imaging studies have been useful for exploring neural underpinnings of attention (Kastner & Ungerleider, 2000; Miller & Buschman, 2013), and psychophysical studies have examined various behavioural manifestations of human visual attention (Petersen & Posner, 2012; Simons & Chabris, 1999; Wolfe, 1998, 2007) (see also the 'Course Readings' section of the references). A synthesis of all this data is warranted; however, while it is unclear what it means to truly understand visual attention, these independent data points are likely insufficient. Instead, scientific progress is made by a meaningful compression of data, for example by constructing models that can explain and predict a diverse range of phenomena. In this domain, computational, rather than conceptual (or descriptive) models, have the advantage of providing quantitative explanations of the collected observations as well as making new predictions that are testable and verifiable. The use of computational models has led to progress in our understanding of various phenomena. For instance, developments in bottom-up attention modeling have led to an increased understanding of where people look in different images under varying conditions (Borji et al., 2013; Itti & Baldi, 2009; Judd, 2011; Tatler, 2007), computational models have been able to predict the effects of crowding on visual tasks (Balas, Nakano, & Rosenholtz, 2009; Rosenholtz et al., 2012), and to model top-down scene guidance for visual search tasks (Ehinger et al., 2009; Torralba, Oliva, Castelhano, & Henderson, 2006; Tsotsos, 2011). Taken together, this suggests that constructing computational models to solve specific visual attention tasks could lead to progress in understanding visual attention as a whole.

Nevertheless, we begin in Section 2 by highlighting the difficulties in model evaluation and comparison brought about by the simultaneous abundance of computational models of visual attention and the lack of model overlap across taxonomies. In Section 3 we advocate for quantitative evaluation via (i) operationalizing definitions of individual visual attention tasks and (ii) specifying rigorous protocols for measuring model performance under those tasks, and we provide some implementable examples. Operationalized task definitions are those that include sufficient detail and specificity so that the tasks may be put into practice,

* Corresponding author at: 32-D542, 32 Vassar St., Cambridge, MA 02141, USA.
E-mail address: zoya@mit.edu (Z. Bylinskii).
[1] The first 5 authors, listed alphabetically, contributed equally to this manuscript.

http://dx.doi.org/10.1016/j.visres.2015.04.007
0042-6989/

implemented on a computer and quantitatively evaluated on meaningful input stimuli. We advocate against any abstract and ambiguous constructs that do not lend themselves easily to quantitative evaluation.

Next, in Section 4 we emphasize the need for large, multi-faceted, standardized benchmark datasets, and offer a discussion of the design considerations that surface. Finally, we outline the benefits of competition-style online benchmarks in Section 5 for measuring modeling progress. Altogether, this paper offers a number of suggestions and considerations that have proven successful at bringing structure and standardization to other computational areas (e.g. evaluation methodologies and benchmark datasets in saliency modeling (Borji et al., 2013; Bylinskii et al., 2014; Judd, Durand, & Torralba, 2012), computer vision (Deng et al., 2009; Everingham et al., 2012; Lin et al., 2014; Torralba, Fergus, & Freeman, 2008; Xiao et al., 2010), and natural language processing (NIST, 2013; Voorhees, 2004; Voorhees & Harman, 2005)).

2. Moving beyond taxonomies

Many computational models of visual attention have been built during the past three decades. However, the sheer diversity of models makes comparison and evaluation of progress in the field of visual attention particularly difficult. In an attempt to understand the relationships between different models, various taxonomies and other categorizations have been introduced, some of which attempt to cover multiple types of computational models, and others that focus on specific subareas of visual attention or specific model structures. For instance, Frintrop, Rome, and Christensen (2010) classify models according to their structure, labeling them either as filter models, those that parse image features via image mapping, or connectionist models, those that employ neural network computations to process images. Tsotsos and Rothenstein (2011) divide computational models (themselves branching off from both computer and biological vision categories) into four types: selective routing models, saliency map models, temporal tagging models, and emergent attention models. Kimura, Yonetani, and Hirayama (2013) classify models as either bottom-up or top-down, each composed of several subcategories determined by the models' algorithmic approach. Borji and Itti (2013) present a categorization of bottom-up and top-down models, qualitatively comparing 13 criteria.

In Fig. 1 we visualize the number of models that are considered by each of 4 categorizations (Borji & Itti, 2013; Frintrop et al., 2010; Kimura et al., 2013; Tsotsos & Rothenstein, 2011). We can see that relatively few models occur in more than one taxonomy/categorization, making comparisons very difficult. Each categorization covers only a subset of models and proceeds by carving up these models according to some author-defined set of characteristics. Another observation is that the sheer number of visual attention models that have been developed over the past few decades is staggering, and continues to grow.

Let us consider a single model categorization in greater detail. According to Borji and Itti (2013), there are a total of 13 criteria by which many of these models may be compared: bottom-up, top-down, spatial/spatiotemporal, task-type, space-based/object-based, features, model type, static, dynamic, synthetic, natural, measures, and dataset used. The first 7 criteria correspond to the models themselves, and the latter 6 are specific to task completion and evaluation. As Borji and Itti note, these criteria help establish the scope of applicability of these different models. In Fig. 2a, we visually represent this taxonomy by projecting down the model characteristics onto 3 dimensions. Gaussian noise was added to the projections to visualize models with

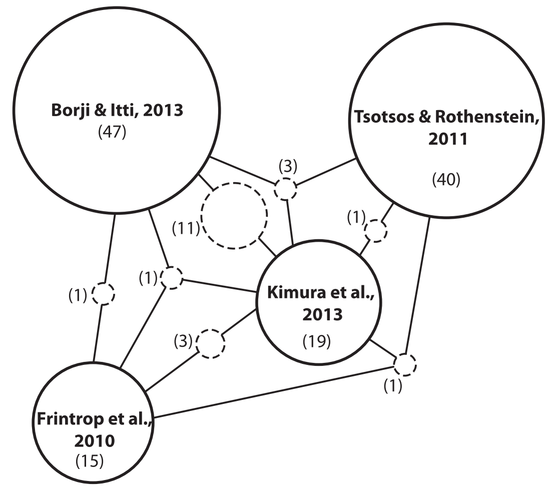

Fig. 1. There are many logical ways of carving up the space of models in the visual attention literature, and different taxonomies/categorizations consider different subsets of models. Here we include four categorizations that cover a total of 142 models (listed in the appendix). Many, but not all, of the models included in each categorization are accounted for here (45 from Tsotsos and Rothenstein (2011), 21 from Frintrop et al. (2010), 39 from Kimura et al. (2013), and 63 from Borji and Itti (2013)). This figure shows the overlap in models across these 4 categorizations. Each solid circle denotes a categorization, and each dashed circle is a node connecting categorizations, denoting their intersection. The parenthesized numbers are model counts. For instance, the model categorizations of Borji and Itti (2013) and Kimura et al. (2013) only have 11 models in common. It is clear that there is little overlap in models between the four categorizations.

identical 3-dimensional projections. The resulting representation accurately captures the factor similarity of models, i.e. models that are spatially clustered together share many taxonomical attributes. The dimensions of this representation are principal components[2] that represent a linear combination of factors, although they do align fairly well with the factors: bottom-up/top-down, dynamic/static, and synthetic/natural. In 2b, we hold this spatial layout of models fixed, and overlay on top of it multiple model characteristics (represented by the coloring of models). From such a visualization we can see that models are clustered together in model space, with many overlapping and correlated characteristics. For example, bottom-up and top-down models are segregated along the first dimension of this representation, while models with synthetic versus natural stimuli are segregated along the third dimension. Thus, although the quantity of models is large, many reuse the same principles and computational approaches, and thus have similar application areas (use cases).

Taxonomies thus provide a way to describe models, but not with a method of sorting through them to discover the most accurate representation of human visual attention. We can use taxonomies to describe the characteristics of different models, or to identify models which may be sensibly compared, because they solve similar tasks or use comparable computational approaches. However, if a quantitative evaluation is sought, these descriptions need to be supplemented with a methodology of comparison. Quantitative evaluation can help us isolate the model characteristics that are essential to performance on different visual attention tasks.

Even though some attempts have been made to quantitatively evaluate a wide varieties of models according to some predefined criteria (Borji, Sihite, & Itti, 2012; Filipe & Alexandre, 2013; Heinke & Humphreys, 2005; Judd et al., 2012; Koehler et al., 2014), these endeavors only provide a comparison of a relatively

[2] Corresponding to combinations of factors with highest variance, as computed via Principal Components Analysis (PCA).

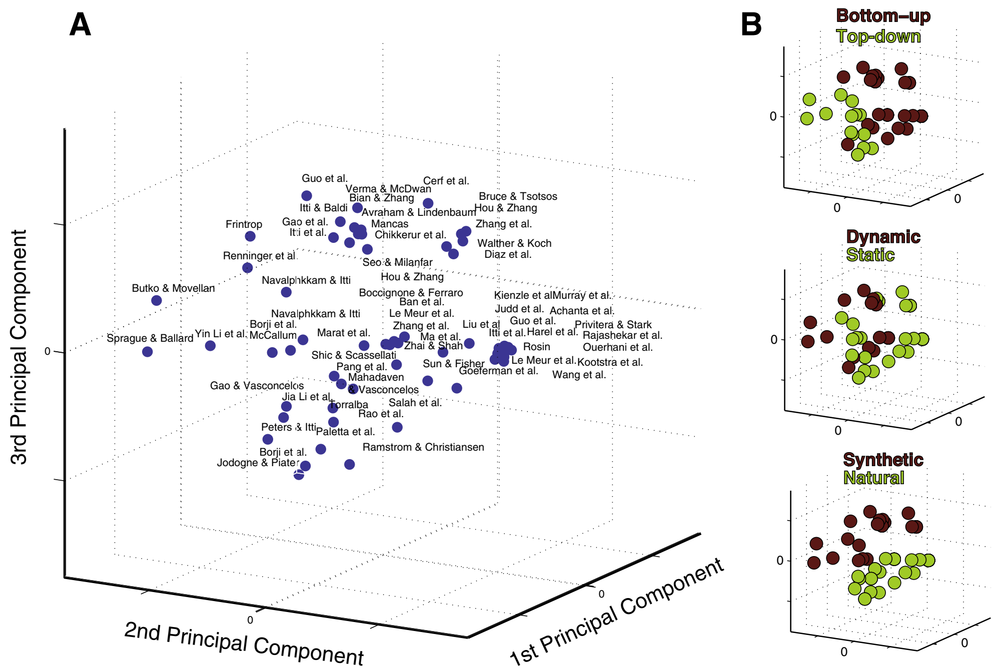

Fig. 2. (A) Spatial layout of 65 models, projecting the model factors from Borji and Itti (2013) onto 3 dimensions of maximal variance (with a small random perturbation to reduce visual overlap). These dimensions roughly correspond to the factors: bottom-up/top-down, dynamic/static, and synthetic/natural. The factors are indicative of the scope of applicability of these models. Note that spatial proximity in this representation tightly corresponds to similarity in factors; models that are clustered together share many taxonomical attributes. For example, many top-down models are clustered in the same region of space, as is clear in the first panel of (B). (B) Classifying models according to different factors (here 3 of the 11 factors from Borji and Itti (2013)) provides us with a way to describe models and to discover which models can be sensibly compared. We must then turn to other means for quantitative model comparison. The points in these 3 plots occur in the same spatial layout as in (A).

small subset of computational models, which highlight further the need for more coherence within the field. Thus it is evident that there is much work left to be done in making comparison and evaluation more quantitative. We suggest that this can be accomplished via developments in: (1) operationalizing the definitions of visual attention tasks (Section 3), and (2) defining success on these tasks via benchmarking datasets (Section 4). In this paper, we focus on these two particular issues with respect to the computational modeling of visual attention.

3. Operationalizing definitions of visual attention tasks (for model evaluation)

William James famously said "Everyone knows what attention is." Nonetheless, attention has been variously described by those attempting to study it as an emergent property (Desimone & Duncan, 1995), a controlling factor (Rensink, 2000), a mental ability that allows for the selection of behaviorally relevant stimuli (Corbetta, 1998) and a reallocation of visual processing resources while preserving reactivity to rapid changes in the environment (Foley, Grossberg, & Mingolla, 2012). Here, we propose that instead of attempting to reach agreement on a particular semantic definition of attention, more progress can be achieved by moving towards **operational definitions**[3] of different visual attention tasks, including free viewing and visual search (Table 1). These operational definitions should provide specifications of the phenomena (behavioral, physiological, etc.) that should be reproducible by **image-computable** models of attention. Image-computable models take as input an image, a task, and a system state, and algorithmically compute a function of the image. Image-computability makes models of visual attention testable and comparable to other models.

We provide examples of a few visual tasks in Table 1 below. Operationalization allows us to turn our attention away from arguing over whether a model is a good model of visual attention, to measuring how the model accomplishes a particular visual attention task, or set of tasks.

For example, one behavioral manifestation of stimulus-based bottom-up attention is the pattern of eye movements made by humans on images under some fixed task constraints (e.g. free-viewing). We can operationally define free-viewing as human eye movements when given an image to look at and no specific task instructions.[4] Based on this definition, we can then choose a measurement of eye movements: location of fixations, ordered sequence of fixations (scanpath), first fixation, dwell time per fixation, saccade extent, etc. Once we choose a measurement, this directly provides us with a methodology for evaluating models: for instance, if we are interested in fixation locations, we can use similarity metrics to compare model-predicted likelihoods of fixating different image regions with human fixation maps. Saliency models are image-computable, taking images as input, and returning topographic maps indicating the likelihood of fixation (or "saliency") at each location in the image. Viewed as a distribution over an image, a saliency map can be compared to a human fixation map using various similarity metrics. Another possible model of bottom-up attention might be one that takes as input an image and the previous fixation location (as a system state) and returns the next location of fixation as its output.

[3] For instance, although imperfect, Intelligence Quotient (IQ) tests provide a way of operationalizing intelligence, by reducing it to a measurable quantity that can be studied empirically (for instance, to investigate the impact of a particular condition on intelligence, we can quantify its effect on IQ score). Note that multiple operationalizations can exist (as is the case for intelligence), and thus to facilitate comparison across studies, research findings should make explicit reference to the particular operational definitions employed.

[4] However, different experiments claiming to record free viewing fixations often provide participants with implicit tasks (e.g. to remember an image) and a very constrained experimental setting (e.g. fixed distance to screen, image of finite size displayed for short period of time). This further highlights the difficulty, but necessity, of having operationalized definitions for all experimental components: the phenomena studied, the tasks implemented, the constraints and assumptions used, etc.

Please cite this article in press as: Bylinskii, Z., et al. Towards the quantitative evaluation of visual attention models. *Vision Research* (2015), http://dx.doi.org/10.1016/j.visres.2015.04.007

Table 1
Sample operational definitions for a few visual attention tasks. Under each definition and experimental measurement, the dataset requirements and metrics naturally fall out. Note that many other possibilities exist for each task, for which we only provide a few examples. For instance, we provide behavioral definitions, whereas physiological definitions can also be applied (e.g. a description of the neuronal firing patterns and rates expected under different conditions).

Task	Operational definition	Experimental measurement	Dataset	Evaluation metrics
Free-viewing	Human eye movements when given an image to look at and no task instructions	Fixation locations	Set of images with ground-truth human fixations (see Bylinskii et al., 2014 for a listing of popular ones)	Similarity between human fixations and model saliency map (see Riche et al. (2013) for some examples)
		Scanpath (order of fixations)	Set of images with ground-truth human scanpaths	Similarity between scanpaths (see Le Meur and Baccino, 2010 for some examples)
Visual search with fixed gaze	Human response when given an image with or without a target and distractors and the instructions to respond to target absence/presence without changing fixation (speeded or unspeeded)	Response time	Set of images with varying number of distractors (with and without a target) and accompanying human response times	Similarity between human and predicted response times
		RT $\times$ set size	Set of images (as before) and accompanying human RT slopes[a]	Similarity between human and predicted RT slopes
		Target prediction	Set of images (as before) and ground-truth target locations	Detection accuracy (see Macmillan and Creelman, 1991)
Visual search with free gaze	Human response and eye movements when given an image with target and distractors and the instructions to respond to target absence/presence by fixating target	Same as for visual search with fixed gaze, potentially also fixation locations and scanpaths	Same as for visual search with fixed gaze, but with the additional possibility of peripheral displays in the images, and ground-truth human fixations	Same as for visual search with fixed gaze, potentially also similarity between human fixations and model saliency map

[a] As in Wolfe's visual search data set, available at: http://search.bwh.harvard.edu/new/data_set_files.html.

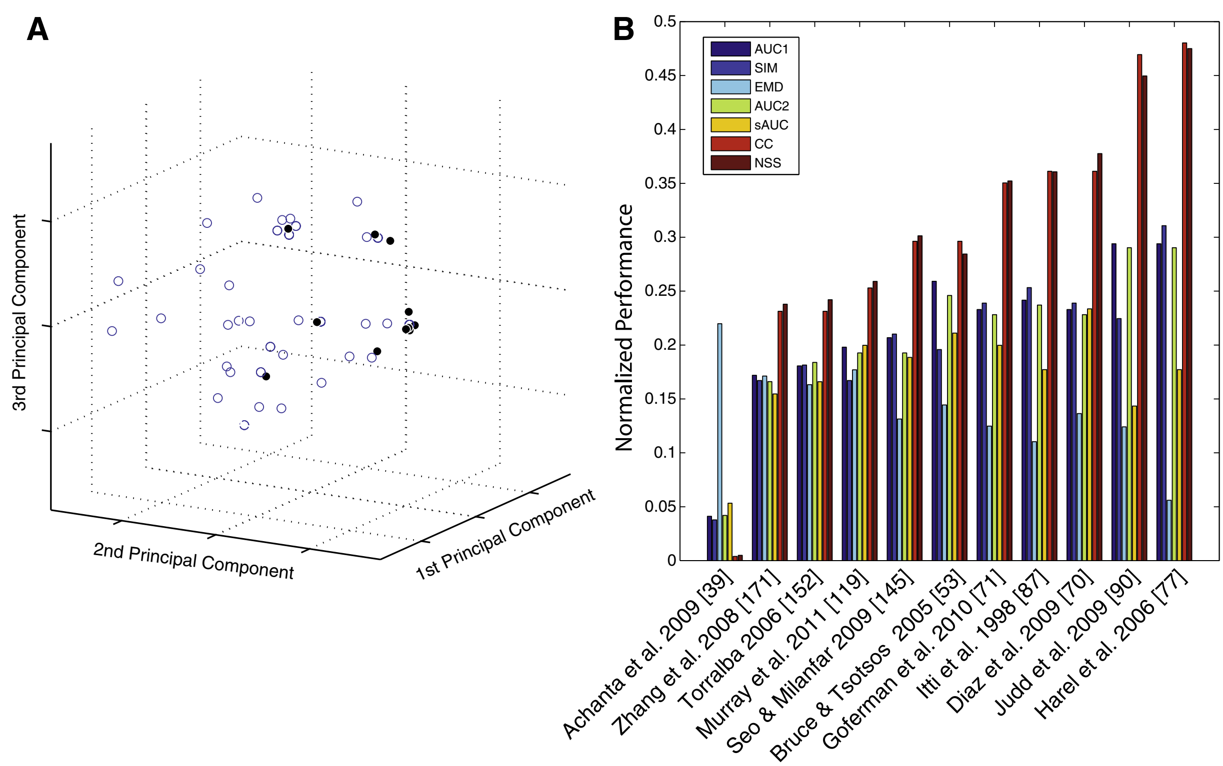

Fig. 3. A subset of the models from Fig. 2 has been evaluated on the MIT Saliency Benchmark (Bylinskii et al., 2014). (A) These models are indicated by the filled-in circles, in the same spatial configuration as in Fig. 2. (B) Scores for the models are plotted according to 7 different metrics: 2 computations of area under precision-recall curve (AUC1 and AUC2) and a shuffled variant that compensates for center bias (sAUC), a similarity computation equivalent to histogram intersection (SIM), an Earth Mover's Distance (EMD), a cross-correlation (CC), and a normalized scanpath saliency (NSS) – all defined in Bylinskii et al. (2014). For the purpose of displaying different metrics with different scales on the same axes, all scores were normalized to lie between 0 and 1. The use of metrics allows us to directly compare and rank models on specific visual attention tasks (in this case, saliency). The results according to different metrics provide us with a window as to where some of the performance may be coming from (e.g. models that are high on AUC but low on sAUC likely have a built-in center bias – see Section 4.1 for a discussion).

In Fig. 3 we include the performances of a subset of the bottom-up models from Fig. 2 that are also available on the MIT Saliency Benchmark (Bylinskii et al., 2014). Seven metrics are provided, allowing for multiple comparisons across, and ranking of, computational models. Whereas taxonomic classifications can point out the differences between models (with respect to their

Please cite this article in press as: Bylinskii, Z., et al. Towards the quantitative evaluation of visual attention models. *Vision Research* (2015), http://dx.doi.org/10.1016/j.visres.2015.04.007

characteristics), metrics allow us to step further and make quantitative judgements about which model performs better on a given task, and under which metric.

Many popular image datasets with human gaze data exist for evaluating saliency models (reviews of some of these can be found in Borji et al. (2013), Judd et al. (2012), and a regularly-updated online listing can be found in Bylinskii et al. (2014)). A discussion of commonly-used evaluation metrics is provided in Riche et al. (2013). Additionally, other behavioral (e.g. human ratings) and physiological (e.g. neural recordings) manifestations of bottom-up attention are amenable to evaluation on separate datasets, and other sets of metrics may be appropriate. We have yet to see the development of benchmark datasets and large-scale evaluation methodologies for bottom-up attention beyond the prediction of eye movements and saliency.

3.1. Additional considerations

In a more general review of computational models of selective attention, Heinke and Humphreys (2005) put forth several possible evaluation metrics, though indicating that comparison between all models is not feasible given the wide variety of model designs and tasks. They note that a chief difficulty lies (a) in the variability of parameter settings, and (b) the lack of a consensus on how to weight biological plausibility when evaluating models. On these two issues, we make the following observations:

(a) Reporting the influence of different parameter choices on model performance (e.g. via sensitivity analysis[5]) is crucial for standardizing evaluation, simply because a single setting of parameters does little to disentangle the power of the underlying model formulation from fine parameter tuning. Not only does the lack of such reporting not help progress, it actually halts it by potentially offering misleading information about what works well and what does not.
(b) To quantify biological plausibility, models should be evaluated both in their ability to explain the most comprehensive behavioral data (e.g. reaction times, gaze patterns, success rate, etc.) as well as the available underlying biological data at various levels of abstraction (e.g. neuronal activity, local field potentials, EEG/MEG, neuro-imaging, etc.). An integrated, complete computational description should unite behavioral findings with the underlying physiological mechanisms.

To summarize, in order to quantitatively measure progress within a given subarea of visual attention, we advocate for the use of computational models evaluated under a rigorous quantitative protocol. This involves first selecting and operationalizing particular visual attention tasks, and specifying evaluation protocols under those task definitions. The model evaluation protocols should additionally detail (1) what data and task constraints are to be used; (2) which metrics models will be compared on; (3) what additional aspects of models should be reported (e.g. parameter choices, model complexity, underlying biological assumptions, etc.).

4. Need for benchmarking datasets (for model comparison)

Given a set of operationally-defined tasks comprising visual attention, the ultimate goal would be to have a model capable of performing well on a maximum number of these tasks. This would hopefully lead to a unified computational understanding of the mechanisms underlying visual attention. Such a model would become the **reference model** of visual attention.[6]

In contrast, in the case of saliency, there has been significant noise in the reporting of successes because many papers introducing novel saliency models report the performance of their model on some choice of dataset(s), under some choice of metric(s), compared against some choice of other models. All these choices afford too much flexibility to the model authors as to how the success of their model's performance is quantified. For instance, in the classification of models provided in Borji and Itti (2013), most of the 63 models considered are included along with the metrics and datasets the models were originally evaluated on. Over 27 distinct datasets are listed and an additional 8 mentions of authors gathering their own data. From this information alone, it is impossible to infer which models are objectively doing well at predicting human fixations. Thus, while some coarse but useful comparisons can be made across models, there is currently no clear conclusion of a winning reference model.

A systematic approach to determining a possible reference model would be to compare all existing models on standard benchmark evaluation datasets under the same set of metrics. Standardized datasets and evaluation metrics have proved successful in the fields of computer vision (Deng et al., 2009; Everingham et al., 2012; Lin et al., 2014; Torralba et al., 2008; Xiao et al., 2010) and natural language processing (NIST, 2013; Voorhees, 2004; Voorhees & Harman, 2005). Unfortunately, benchmark datasets are difficult to construct, and are consequently rare in visual attention. Benchmarks are, however, becoming increasingly popular for the evaluation of saliency models, and this is helping to drive progress and eliminate some of the aforementioned noise in model evaluation. The problem is that saliency benchmark datasets are now where computer vision datasets were a decade ago (Antonio Torralba, personal communication, 4/28/14). By image count alone, computer vision has now moved on to datasets of tens of thousands to many millions of images in size (Deng et al., 2009; Everingham et al., 2012; Lin et al., 2014; Torralba et al., 2008; Xiao et al., 2010), while saliency datasets, in particular, have not grown beyond a few hundred to a few thousand images (Borji et al., 2013; Judd et al., 2012).

4.1. Dataset design considerations

Large datasets alone are insufficient for capturing the space of complex behaviors that are attributable to visual attention. A good benchmark dataset for testing computational models of visual attention should offer multiple tasks on which performance can be reported. As an analog, a popular image benchmark in computer vision (the PASCAL challenge Everingham et al. (2012)) consists of separate competitions that computational models can be evaluated on (e.g. classification, detection, segmentation, etc.). Computing model performance on different tasks may help illuminate what aspects of visual attention the model is most predictive of Koehler, Guo, Zhang, and Eckstein (2014), as well as which models may have complementary functionalities. Testing models on multiple tasks also provides a way of differentiating a solid underlying model framework (capable of generalizing) from one that owes its performance to the implementation details and parameter tuning specific to a task.

[5] Sensitivity analysis can be performed in many different ways, but comes down to testing the robustness of a model under changing inputs and parameter settings, to carefully measure effects on performance and quantify uncertainty (Saltelli et al., 2008 provides a thorough treatment of the subject).

[6] By reference model we mean a standardized model that would be the accepted reference point for other models to compare against, at a fixed point in time. To obtain this status, a model would need to be capable of explaining the most complete set of observations on the broadest set of tasks.

Please cite this article in press as: Bylinskii, Z., et al. Towards the quantitative evaluation of visual attention models. *Vision Research* (2015), http://dx.doi.org/10.1016/j.visres.2015.04.007

Furthermore, summary performance numbers may be insufficient to tease models apart: is a model performing reasonably across all conditions, or is its performance confined to a small set of successes? Do different models make similar mistakes, or are they complementary in the data they can explain? Reporting multiple performance numbers on each dataset may reveal common and complementary aspects of different models.

Additionally, reporting results on multiple datasets can help alleviate dataset bias, which can have severe implications on model evaluation (Andreopoulos & Tsotsos, 2012; Torralba & Efros, 2011). For instance, center bias has been a prevalent problem in saliency model evaluation (Borji et al., 2013; Judd et al., 2012), with a simple center prior model, completely ignorant of image content, often outperforming many other models. One potential cause is photographer bias (placing the main content in the center), which could be avoided, for instance, by a more careful selection of images. An alternative to modifying the data is designing evaluation metrics to compensate for dataset biases. For instance, the shuffled AUC (sAUC) metric (Borji et al., 2013; Zhang et al., 2008) assigns chance performance to a center prior model. However, Tatler (2007) has found that observer bias is a major contributing factor to center bias (a kind of optimal viewing strategy). Wloka and Tsotsos have shown that center-bias seems to be a feature of fixed-size image viewing, not present in natural free-head viewing (Wloka & Tsotsos, 2013). In this case, center bias may be natural under the task constraints of no head movements and fixed-size input. In any case, center bias remains an issue to resolve.

These are the types of questions that must be carefully considered when putting together benchmark datasets and evaluation methodologies: what are the possible biases? Are they a property of the data or the task? Should dataset bias be compensated for by testing on different datasets or using appropriate metrics? Should task bias be a property of the models?

In Table 1, we provide some examples of how benchmarking can be applied to different visual attention tasks under the operationalized definitions proposed. A benchmark provides models with a standard dataset of inputs composed of images, task constraints, and system state (e.g. current context). Model output is then evaluated according to standardized evaluation metrics on ground-truth measurements (human fixations, neuronal recordings, etc.). In some cases, the ground-truth measurements are held out and known only by the benchmark curators. This prevents models from being specifically tailored to fit the data. In other cases, part of the ground-truth measurements may be released for training models, but a held-out test set is ultimately what models are evaluated on. This is the case for the MIT Saliency Benchmark (Bylinskii et al., 2014), an online benchmarking website which publicly releases only images, and not the ground-truth human fixations on those images. Model designers submit saliency maps computed on this set of images, and the benchmark curators evaluate the saliency maps against human fixation maps according to 7 standardized metrics (as of this manuscript's publication date). The website makes available the code for the metrics and for optimization, to allow model designers to re-adjust saliency map parameters to compensate for dataset biases. The website also includes an up-to-date list of other fixation datasets that can be used for model evaluation.

Similarly, for standardizing model comparison in other subareas of visual attention, we propose that benchmark datasets should be constructed: (1) to cover multiple visual attention tasks; (2) to include multiple measures of performance and comparison per task; and (3) to permit model testing on multiple datasets to minimize the effects of dataset bias. The additional benefits of putting benchmarks online are discussed in the following section.

5. Future directions

Established evaluation protocols and benchmark datasets can help organize and categorize the wide range of phenomena and corresponding computational models in visual attention research. While comprehensive reviews are immensely useful for performing this synthesis, we believe that an important next step is to establish a systematic, high-throughput and rapidly-updated evaluation protocol. We propose having online, up-to-date competitions, whereby all model entries are evaluated in the same manner, via the same metrics, on the same data, compared against the same set of models. The latest progress in the field can thus be documented. Such a benchmark already exists for saliency modeling (Bylinskii et al., 2014), but most other subareas of visual attention have no such benchmark available. Benchmarks have been crucial for pushing progress in the computer vision (Deng et al., 2009; Everingham et al., 2012; Lin et al., 2014; Torralba et al., 2008; Xiao et al., 2010) and natural language processing (NIST, 2013; Voorhees, 2004; Voorhees & Harman, 2005) communities by having regularly-updated lists of the best-ranking models. Note that we are not advocating for the reduction of all model evaluation to one summary number, as this is uninformative and a poor measure of success. Instead, as discussed in the previous section, evaluation should be multi-faceted – covering multiple tasks, datasets, and metrics. In such a way, evaluation is better able to capture a broader set of capabilities of a model. Making evaluation quantitative is precisely what will allow us to make comparisons rigorous, serving as a better indicator of progress.

Given the established groundwork and success of computer vision competitions, this evaluation paradigm is most easily implemented by studying attentional effects as human responses to sequences of image stimuli. This suggests that the first-generation visual attention benchmarks should focus on evaluating image-computable models. The ground truth for this evaluation paradigm can be multi-layered, spanning data from single-unit neuronal responses all the way to behavioral outputs. For instance, models of visual search may attempt to predict behavioral patterns of human subjects performing visual search experiments, as well as neurons in the non-human primate attention network that modulate their activity based on search behavior. Already, there are many published results on human and non-human primate visual search that can be used to bootstrap the benchmarking process. Good models of visual search should be capable of explaining as comprehensive a set of the published results as possible.

When consensus about the specific behavioral measures and similarity metrics cannot be attained, or when new measures are introduced, multiple measures can be considered simultaneously. Similarly, the establishment of a competition-style benchmarking system allows for relatively rapid updating with new data and new models. Finally, by augmenting the test data[7] on a yearly basis, established models will be constantly challenged and must prove their value to remain in contention for the position of a reference model.

The computational modeling of visual attention is a rapidly-developing area. In addition to making significant contributions to our understanding of visual attention, these models have made significant impacts on other scientific domains, via applications including efficient image search for object detection and recognition, surveillance, image segmentation, image

[7] As long as the focus is on evaluating image-computable models, augmenting the dataset is as simple as adding images and corresponding measurements, potentially from different experimental settings. Each experiment may come with an independent set of data, and evaluation can also proceed independently, but with the end result of producing a final set of performance numbers.

compression, photo retargeting, infographic design, user-interface design, brain–machine interfaces, medical diagnosis, robot navigation, as well as educational and promotional content design (discussed in greater detail by Judd (2011), Borji & Itti (2013)).

The growing interest in applications of computational models of attention is likely, as a byproduct, to stimulate progress in this field. As the number of computational models available continue to increase, it may become increasingly difficult to find order and structure among them, making scientific progress difficult to evaluate. Thus, now is a crucial time to establish standardized evaluation methodologies. In this paper, we have offered some approaches for standardization, borrowing ideas from other computational fields that have proven successful.

6. Origin of this opinion paper

This presentation comes out of the culmination of a seminar course on "Understanding Visual Attention through Computation".[8] The course details are included in the Appendix. The authors come from five different institutions and represent a total of eight different research labs with diverse backgrounds and research interests, ranging from physiological to behavioral and computational. Their own research areas served as launching pads for initiating discussion about visual attention.

Acknowledgements

We would like to thank the guest lecturers: Salva Ardid, Neil Bruce, Robert Desimone, Mazyar Fallah, Ruth Rosenholtz, Thomas Serre, Antonio Torralba, Jeremy Wolfe, and Thilo Womelsdorf for their time and stimulating presentations.

JKT is grateful for the support of Robert Desimone and James DiCarlo during his sabbatical at MIT, during which this course was presented and this paper developed.

JKT, ZB, and RR acknowledge support from the Natural Sciences and Engineering Research Council of Canada. JKT additionally acknowledges support from the Canada Research Chairs Program. JZ is supported by the China Scholarship Council and National Natural Science Fund of China (Grant No. 61233011).

Appendix A. Models included in Figure 1

A.1. From Kimura et al. (2013)

Achanta et al. (2008), Amari, Cichocki, and Yang (1995), Avraham and Lindenbaum (2010), Borji et al. 2012, Cerf et al. (2007), Eckstein et al. (2000), Frintrop (2006), Hyvarinen et al. (2001), Itti and Baldi (2005), Itti, Dhavale, and Pighin (2003), Itti, Koch, and Niebur (1998), Jeong, Ban, and Lee (2008), Judd et al. (2009), Kanan et al. (2009), Kienzle et al. (2009), Kimura et al. (2010), Koch and Ullman (1985), Leung et al. (2007), Li et al. (2010), Ma and Zhang (2003), Maki, Nordlund, and Eklundh (1996), Marat et al. (2009), Miyazato et al. (2009), Muller, Humphreys, and Donnelly (1994), Nakayama (1990), Nakayama and Silverman (1986), Olshausen, Anderson, and Van Essen (1993), Ouerhani and Hugli (2000), Pang et al. (2008), Peters and Itti (2007), Rao et al. (2002), Renninger et al. (2005), Sandon (1990), Sun et al. (2010), Torralba et al. (2006), Verghese (2001), Yamada et al. (2010), Zelinsky et al. (2006), Zhang et al. (2008).

[8] Course 9.S913, offered in the Spring of 2014, in the Dept. of Brain and Cognitive Sciences, Massachusetts Institute of Technology.

A.2. From Frintrop et al. (2010)

Backer, Mertsching, and Bollmann (2001), Bundesen (1990), Eckstein et al. (2000), Eriksen and St. James (1986), Frintrop (2006), Green and Swets (1966), Hamker (2005), Heidemann et al. (2004), Humphreys and Muller (1993), Itti et al. (1998), Logan (1996), Milanese (1993), Olshausen et al. (1993), Palmer, Ames, and Lindsey (1993), Phaf, Van der Heijden, and Hudson (1990), Postma (1994), Rensink (2000), Sun and Fisher (2003), Tsotsos et al. (1995), van Oeffelen and Vos (1982), Verghese (2001).

A.3. Tsotsos and Rothenstein (2011)

Ahmad (1992), Anderson and Van Essen (1987), Bajcsy (1985), Ballard (1991), Broadbent (1958), Bundesen (1990), Burt (1988), Clark and Ferrier (1988), Corchs and Deco (2001), Deco and Zihl (2001), Desimone and Duncan (1995), Deutsch and Deutsch (1963), Fukushima (1986), Grossberg (1982), Hamker (2005), Hummel and Biederman (1992), Humphreys and Muller (1993), Itti and Baldi (2005), Itti and Koch (2000), Itti et al. (1998), Kelly (1971), Knudsen (2007), Koch and Ullman (1985), Lee and Maunsell (2009), Moravec (1981), Muerle and Allen (1968), Navalpakkam and Itti (2005), Niebur, Koch, and Rosin (1993), Nowlan and Sejnowski (1995), Olshausen et al. (1993), Postma (1994), Reynolds and Heeger (2009), Sandon (1990), Shipp (2004), Taylor and Rogers (2002), Treisman (1964), Treisman and Gelade (1980), Treue and Martinez-Trujillo (1999), Tsotsos (1992), Tsotsos et al. (1980), Tsotsos et al. (1995), Ullman (1984), Walther et al. (2002), Wiesmeyer and Laird (1990), Zhang, Tong, Marks, Shan, and Cottrell (2008).

A.4. Borji and Itti (2013)

Achanta et al. (2009), Avraham and Lindenbaum (2010), Ban, Lee, and Lee (2008), Bian and Zhang (2009), Boccignone and Ferraro (2004), Borji, Ahmadabadi, and Araabi (2011), Bruce and Tsotsos (2005), Butko and Movellan (2009), Cerf, Harel, Einhauser, and Koch (2007), Chikkerur et al. (2010), Garcia-Diaz et al. (2009), Frintrop (2006), Gao and Vasconcelos (2004), Gao, Han, and Vasconcelos (2009), Goferman, Zelnik-Manor, and Tal (2010), Guo, Ma, and Zhang (2008), Guo and Zhang (2010), Harel, Koch, and Perona (2006), Hou and Zhang (2008), Hou and Zhang (2007), Itti and Baldi (2005), Itti et al. (1998), Itti et al. (2003), Li et al. (2010), Jodogne and Piater (2007), Judd et al. (2009), Kienzle et al. (2009), Kootstra, Nederveen, and de Boer (2008), Le Meur, Le Callet, and Barba (2007), Le Meur et al. (2006), Liu et al. (2011), Ma et al. (2005), Mahadevan and Vasconcelos (2009), Mancas (2007), Marat, Guironnet, and Pellerin (2007), McCallum (1996), Murray et al. (2011), Navalpakkam and Itti (2006), Navalpakkam and Itti (2005), Ouerhani et al. (2003), Paletta, Fritz, and Seifert (2005), Pang et al. (2008), Peters and Itti (2007), Privitera and Stark (2000), Rajashekar et al. (2008), Ramstrom and Christensen (2002), Rao et al. (2002), Renninger et al. (2005), Rosin (2009), Salah, Alpaydin, and Akarun (2002), Seo and Milanfar (2009), Shic and Scassellati (2007), Sprague and Ballard (2003), Sun and Fisher (2003), Torralba et al. (2006), Verma and McOwana (2009), Walther and Koch (2006), Wang et al. (2011), Li et al. (2009), Zhai and Shah (2006), Zhang, Tong, and Cottrell (2009), Zhang et al. (2008).

Appendix B. Course information: understanding visual attention through computation

A full syllabus for the course taught by the senior author on "Understanding Visual Attention through Computation" can be found online.[9] It is available as a resource and possible inspiration for future courses on visual attention. This highly interdisciplinary course explored many of the different approaches and perspectives in the current literature, within the historical context of research of the field. The intent was to develop a 'big picture' view of what this thing called visual attention might entail and how can we best further deepen our understanding. The course used different themes to explore different viewpoints and theories in order to develop an appreciation of both strengths and weaknesses, not only of the research but also the methodologies. Here we provide an outline of the lecture topics, guest lecturers, and course readings.

Appendix C. Lectures

1. Introduction to Attention, *J.K. Tsotsos* (Tsotsos, 2011)
2. Computational Foundations of Visual Attention, *J.K. Tsotsos* (Tsotsos, 2011)
3. Selective Tuning Part I, *J.K. Tsotsos* (Tsotsos, 2011)
4. Biased Competition, *R. Desimone* (Armstrong, Fitzgerald, & Moore, 2006; Desimone & Duncan, 1995; Reynolds & Heeger, 2009; Sundberg, Mitchell, & Reynolds, 2009)
5. Attention and Search, *J. Wolfe and R. Rosenholtz* (Rosenholtz, Kuzmova, & Sherman, 2011; Rosenholtz et al., 2012; Treisman & Gelade, 1980; Treisman, 2006; Wolfe, 1998; Wolfe, 2007; Wolfe, 1994; Wolfe et al., 2011)
6. Neurobiology of Attention, *M. Fallah* (Baluch & Itti, 2011; Corbetta, Patel, & Shulman, 2008; Fallah, Stoner, & Reynolds, 2007; Krauzlis, Lovejoy, & Zenon, 2013; Maunsell & Treue, 2006; Petersen & Posner, 2012; Salazar et al., 2012; Squire et al., 2013; Sundberg et al., 2012)
7. Saliency Map Models, *N.D.B. Bruce* (Borji & Itti, 2013; Bruce & Tsotsos, 2009;Koehler et al., 2014; Riche et al., 2013)
8. Selective Tuning Part II, *J.K. Tsotsos* (Tsotsos, 2011)
9. Dynamical systems models, *S. Ardid* (Ardid, Wang, & Compte, 2007; Ardid et al., 2010; Borgers, Epstein, & Kopell, 2008; Buia & Tiesinga, 2008; Mante et al., 2013)
10. The Roles of Gist, Context and Task, *A. Torralba* (Isola et al., 2011; Oliva et al., 2003; Oliva & Torralba, 2006; Torralba et al., 2006; Torralba, 2003)
11. Bayesian Methods for Attention, *T. Serre* (Angela & Dayan, 2004; Chikkerur et al., 2010; Dayan, 2009; Dayan & Yu, 2003; Lee & Mumford, 2003)
12. Final Integrative Discussion, *J.K. Tsotsos*

References

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*(1), 110–126.

Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision, 9*.

Borji, A., Sihite, D., & Itti, L. (2012). Probabilistic learning of task-specific visual attention. In *Proc IEEE conference on computer vision and pattern recognition*.

Borji, A., Tavakoli, H. R., Sihite, D.N. & Itti, L. (2013). Analysis of scores, datasets, and models in visual saliency prediction. In *Proc. International Conference on Computer Vision (ICCV), Sydney, Australia*.

Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A, et al. (2014). MIT Saliency Benchmark Results. http://saliency.mit.edu/

Corbetta, M. (1998). Frontoparietal cortical networks for directing attention and the eye to visual locations: Identical, independent, or overlapping neural systems? *Proceedings of the National Academy of Sciences, 95*, 831–838.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*, 193–222.

[9] http://www.cse.yorku.ca/tsotsos/Tsotsos/Archives.html.

Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modeling search for people in 900 Scenes: A combined source model of eye guidance. *Visual Cognition, 17*, 945–978.

Everingham, M., Van Gool, L.,. Williams, C. K. I, Winn, J. & Zisserman, A. (2012). The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

Filipe, S., & Alexandre, L. A. (2013). From the human visual system to the computational models of visual attention: a survey. *Artificial Intelligence Review*.

Foley, N. C., Grossberg, S., & Mingolla, E. (2012). Neural dynamics of object-based multifocal visual spatial attention and priming: Object cueing, useful-field-of-view, and crowding. *Cognitive Psychology, 65*, 77–117.

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

Heinke, D., & Humphreys, G. W. (2005). *Computational models of visual selective attention: A review*. Psychology Press. 273–312.

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research, 49*.

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

Judd, T., Durand, F., & Torralba, A. (2012). A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*.

Kastner, S., & Ungerleider, L. G. (2000). Mechanisms of visual attention in the human cortex. *Annual Review of Neuroscience, 23*, 315–341.

Kimura, A., Yonetani, R., & Hirayama, T. (2013). Computational models of human visual attention and their implementations: A survey. *IEICE Trans Inf. and Syst*.

Lin,T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). Microsoft coco: Common objects in context. In *ECCV*.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. Cambridge University Press.

Le Meur, O., & Baccino, T. (2010). Methods for comparing scanpaths and saliency maps: Strengths and weaknesses. *Behavior Research Methods, 42*(2).

Miller, E. K., & Buschman, T. J. (2013). Cortical circuits for the control of attention. *Current Opinion in Neurobiology, 23*, 216–222.

National Institute of Standards and Technology (NIST) (2013). *Proceedings of the Sixth Text Analysis Conference (TAC 2013)*, Gaithersburg, Maryland, USA.

Rensink, R. A. (2000). The dynamic representation of scenes. *Visual Cognition, 7*, 17–42.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., et al. (2008). *Global Sensitivity Analysis: The Primer*. Wiley.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception, 28*.

Tatler, B. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision, 7*(14).

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision, 7*(14).

Torralba, A. & Efros, A. (2011). Unbiased Look at Dataset Bias. In *Proceedings of the 2011 IEEE conference on computer vision and pattern recognition, CVPR '11*, pp. 1521–1528. IEEE Computer Society.

Torralba, A., Fergus, R., & Freeman, W. T. (2008). 80 million tiny images: A large dataset for non-parametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30*(11), 1958–1970.

Tsotsos, J. K., & Rothenstein, A. (2011). Computational models of visual attention. *Scholarpedia, 6*.

Voorhees, E. M. (2004). TREC: Text Retrieval Conference Tracks. http://trec.nist.gov/data.html.

Voorhees, E. M., & Harman, D. K. (Eds.). (2005). *TREC: Experiment and Evaluation in Information Retrieval*. The MIT Press.

Wloka, C., & Tsotsos, J. (2013). Overt fixations reflect a natural central bias. *Journal of Vision, 13*(9).

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

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

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*(7), 1–20.

References used in Figure 1

Achanta, R, Estrada, F., Wils, P., & Susstrunk, S. (2008). Salient region detection and segmentation. In *Proceedings of international conference on computer vision systems*.

Achanta, R., Estrada, F., Wils, P., & Susstrunk, S. (2008). Salient region detection and segmentation. In *Proceedings of international conference on computer vision systems*.

Ahmad, S. (1992). VISIT: a neural model of covert visual attention. In *Neural information processing systems*.

Amari, S., Cichocki, A., & Yang, H. H. (1995). A new learning algorithm for blind signal separation. In *Proceedings of conference on neural information processing systems*.

Anderson, C., Van Essen, D. (1987). Shifter Circuits: A computational strategy for dynamic aspects of visual processing. In *Proceedings of national academy of science*.

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.
Backer, G., Mertsching, B., & Bollmann, M. (2001). Data- and model-driven gaze control for an active-vision system. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*, 1415–1429.
Bajcsy, R. (1985). Active perception vs passive perception. In *Proceedings of IEEE workshop on computer vision: representation and control*.
Ballard, D. (1991). Animate vision. *Artificial Intelligence, 48*, 57–86.
Ban, S. W., Lee, I., & Lee, M. (2008). Dynamic visual selective attention model. *Neurocomputing, 71*.
Bian, P., & Zhang, L. (2009). Biological plausibility of spectral domain approach for spatiotemporal visual saliency. *Advances in Neuro-information Processing, 5506*, 251–258.
Boccignone, G., & Ferraro, M. (2004). Modeling gaze shift as a constrained random walk. *Physica A, 331*.
Borji, A., Ahmadabadi, M. N., & Araabi, B. N. (2011). Cost-sensitive learning of top-down modulation for attentional control. *Machine Vision and Applications, 22*, 61–76.
Broadbent, D. (1958). *Perception and communication*. Pergamon Press.
Bruce, N. D. B. & Tsotsos, J. K. (2005). Saliency based on information maximization. In *Proceedings of Advances in Neural Information Processing Systems*.
Bundesen, C. (1990). A theory of visual attention. *Psychological Review, 97*, 523–547.
P. Burt. Attention mechanism for vision in a dynamic world. In *Proceedings of 9th international conference on pattern recognition*.
N.J. Butko & J.R. Movellan (2009). Optimal scanning for faster object detection. In *Proceedings of IEEE Conference Computer Vision and Pattern Recognition*.
Cerf, M., Harel, J., Einhauser, W. & Koch, C. (2007). Predicting human gaze using low-level saliency combine with face detection. In *Proceedings of Conference on Neural Information Processing Systems*.
Clark, J. J. & Ferrier, N. (1988). Modal control of an attentive vision system. In *Proc. ICCV*.
Corchs, S., & Deco, G. (2001). A neurodynamical model for selective visual attention using oscillators. *Neural Networks, 14*, 981–990.
Deco, G., & Zihl, J. (2001). A neurodynamical model of visual attention: feedback enhancement of spatial resolution in a hierarchical system. *Journal of Computer Neuroscience, 10*.
Deutsch, J., & Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological Review, 70*, 80–90.
Eckstein, M., Thomas, J., Palmer, J., & Shimozaki, S. (2000). A signal detection model predicts the effects of set size on visual search accuracy for feature, conjunction, triple conjunction, and disjunction displays. *Perception and Psychophysics, 63*, 425–451.
Eriksen, C. W., & St. James, J. D. (1986). Visual attention within and around the field of focal attention: A zoom lens model. *Perception and Psychophysics, 40*, 225–240.
Frintrop, S. (2006). VOCUS: A visual attention system for object detection and goal-directed search. PhD thesis, Rheinische Friedrich-Wilhelms-Universitat Bonn.
Fukushima, K. (1986). A neural network model for selective attention in visual pattern recognition. *Biological Cybernetics, 55*, 5–15.
Gao, D., Han, S., & Vasconcelos, N. (2009). Discriminant Saliency, the detection of suspicious coincidences, and applications to visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(6), 989–1005.
Gao, D. & Vasconcelos, N. (2004). Discriminant saliency for visual recognition from cluttered scenes. In *Proceedings of Advances in Neural Information Processing Systems*.
Garcia-Diaz, A., Fdez-Vidal, X. R., Pardo, X. M., & Dosil, R. (2009). Decorrelation and distinctiveness provide with human-like saliency. In *Proceedings of advanced concepts for intelligent vision systems*.
Goferman, S., Zelnik-Manor, L., & Tal, A. (2010). Context-aware saliency detection. In *Conference on computer vision and pattern recognition*.
Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
Grossberg, S. (1982). A psychophysiological theory of reinforcement, drive, motivation, and attention. *Journal of Theoretical Neurobiology, 18*, 263–327.
Guo, C., Ma, Q., & Zhang, L. (2008). Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform. In *Proceedings of IEEE conference on computer vision and pattern recognition*.
Guo, C., & Zhang, L. (2010). A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Transactions on Image Processing, 19*, 185–198.
Hamker, F.H. (2005). The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision. *Journal of Computer Vision and Image Understanding (CVIU), Special Issue on Attention and Performance*, 100, 64–106.
Harel, J., Koch, C. & Perona, P. (2006). Graph-based visual saliency. In *Advances in neural information processing systems*.
Heidemann, G., Rae, R., Bekel, H., Bax, I., & Ritter, H. (2004). Integrating context-free and context-dependent attentional mechanisms for gestural object reference. *Machine Vision and Applications, 16*, 64–73.
Hou, X., & Zhang, L. (2007). Saliency detection: A spectral residual approach. In *Conference on computer vision and pattern recognition*.
Hou, X. & Zhang, L. (2008). Dynamic visual attention: searching for coding length increments. In *Proceedings of Advances in Neural Infomation Processing Systems*.
Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review, 99*, 480–517.
Humphreys, G., & Muller, H. (1993). Search via recursive rejection (SERR): A connectionist model of visual search. *Cognitive Psychology, 25*, 45–110.
Hou, X., & Zhang, L. (2008). Dynamic visual attention: searching for coding length increments. In *Proceedings advances in neural infomation processing systems*.
Itti, L., & Baldi, P. F. (2005). Bayesian surprise attracts human attention. In *Advances in neural information processing systems* (pp. 547–554). Cambridge, MA: MIT Press.
Itti, L., Dhavale, N., & Pighin, F. (2003). Realistic avatar eye and head animation using a neurobiological model of visual attention. In *SPIE international symposium on optical science and technology*.
Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40*.
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.
Jeong, S., Ban, S. W., & Lee, M. (2008). Stereo saliency map considering affective factors and selective motion analysis in a dynamic environment. *Neural Networks, 21*, 1420–1430.
Jodogne, S. R., & Piater, J. H. (2007). Closed-loop learning of visual control policies. *Journal of Artificial Intelligence Research, 28*, 349–391.
Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *Proceedings of international conference on computer vision*.
Kanan, C., Tong, M., Zhang, L., & Cottrell, G. (2009). SUN: Top-down saliency using natural statistics. *Visual Cognition, 17*, 979–1003.
Kelly, M. (1971). Edge detection in pictures by computer using planning. *Machine Intelligence, 6*, 397–409.
Kienzle, W., Franz, M. O., Scholkopf, B., & Wichmann, F. A. (2009). Center-surround patterns emerge as optimal predictors for huamn saccade targets. *Journal of Vision, 9*, 1–15.
Kimura, A., Pang, D., Takeuchi, T., Miyazato, K., Yamato, J. & Kashino, K. (2010). A stochastic model of human visual attention with a dynamic Bayesian netword. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
Knudsen, E. (2007). Fundamental components of attention. *Annual Review of Neuroscience, 30*, 57–78.
Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, 4*, 219–227.
Kootstra, G., Nederveen, A. & de Boer, B. (2008). Paying attention to symmetry. In *Proceedings of British machine vision conference*.
Lee, J., & Maunsell, J. H. (2009). A normalization model of attentional modulation of single unit responses. *PLoS One, 4*.
Leung, C., Kimura, A., Takeuchi, T. & K. Kashino (2007). A computational model of saliency depletion/recovery phenomena for the salient region extraction of videos. In *Proceedings on IEEE international conference on multimedia and expo*.
Li, J., Tian, Y., Huang, T., & Gao, W. (2010). Probabilistic multi-task learning for visual saliency estimation in video. *International Journal of Computer Vision, 90*, 150–165.
Li, Y., Zhou, Y., Yan, J. & Yang J. (2009). Visual saliency based on conditional entropy. In *Proceedings of Ninth Asian conference on computer vision*.
Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., et al. (2011). Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*, 353–367.
Logan, G. D. (1996). The CODE theory of visual attention: an integration of space-based and object-based attention. *Psychological Review*, 603–649.
Ma, Y. F., Hua, X., Lu, L., & Zhang, H. J. (2005). A generic framework of user attention model and its application in video summarization. *IEEE Transactions of Multimedia, 7*, 907–919.
Ma, Y.F. & Zhang, H. J. (2003). Contrast-based image attention analysis by using fuzzy growing. In *Proc. ACM international conference on multimedia (ACMMM)*.
Mahadevan, V., & Vasconcelos, N. (2009). Saliency based discriminant tracking. In *Proceedings of IEEE conference on computer vision and pattern recognition*.
Maki, A., Nordlund, P., & Eklundh, J. (1996). A computational model of depth-based attention. In *Proc. ICPR*.
Mancas, M. (2007). Computational attention: modelisation and application. PhD thesis, Universite de Mons.
Marat, S., Guironnet, M., & Pellerin, D. (2007). Video summarization using a visual attention model. In *Proceedings of 15th European signal processing conference*.
Marat, S., Ho Phuoc, T., Granjon, L., Guyader, N., Pellerin, D., & Guerin-Dugue, A. (2009). Modeling spatio-temporal saliency to predict gaze direction for short videos. *International Journal of Computer Vision, 82*, 231–243.
McCallum, R. (1996). Reinforcement learning with selective perception and hidden state. PhD thesis, Computer Science Department, University of Rochester.
Le Meur, O., Le Callet, P., & Barba, D. (2007). Predicting visual fixations on video based on low-level visual features. *Vision Research, 47*(19), 2483–2498.
Le Meur, O., Le Callet, P., Barba, D., & Thoreau, D. (2006). A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*, 802–817.
Milanese, R. (1993). Detecting salient regions in an image: From biological evidence to computer implementation. PhD thesis, University of Geneva.
Miyazato, K., Kimura, A., Takagi, S. & Yamato, J. (2009). Real-time estimation of human visual attention with dynamic Bayesian network and MCMC-based particle filter. In *Proceedings of IEEE international conference on multimedia and expo*.
Moravec, H. (1981). Rover visual obstacle avoidance. In *IJCAI*.
Muerle, J. & Allen, D. (1968). Experimental evaluation of techniques for automatic segmentation of objects in a complex scene. In *Pictorial pattern recognition*.

Muller, H., Humphreys, G., & Donnelly, N. (1994). SEarchviPerception Rejection (SERR): Visual search for single and dual form conjunction targets. *Journal of Experimental Psychology—Human Perception and Performance, 20*, 235–258.

Murray, N., Vanrell, M., Otazu, X. & Parraga, C.A. (2011). Saliency estimation using a non-parametric low-level vision model. In *IEEE conference on computer vision and pattern recognition*.

Nakayama, K. (1990). The iconic bottleneck and the tenuous link between early visual processing and perception. *Vision: Coding and Efficiency, 25*, 1545–1555.

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

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research, 45*, 205–231.

Navalpakkam, V. & Itti, L. (2006). An integrated model of top-down and bottom-up attention for optimizing detection speed. In *Proceedings of IEEE conference on computer vision and pattern recognition*.

Niebur, E., Koch, C., & Rosin, C. (1993). An oscillation-based model for the neural basis of attention. *Vision Research, 33*, 2789–2802.

Nowlan, S., & Sejnowski, T. (1995). A selection model for motion processing in area MT of primates. *The Journal of Neuroscience, 15*, 1195–1214.

Olshausen, B., Anderson, C., & Van Essen, D. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience, 13*, 4700–4719.

Ouerhani, N. & Hugli, H. (2000). Computing visual attention from scene depth. In *Proc. ICPR*.

Ouerhani, N., von Wartburg, R., Hugli, H., & Muri, R. M. (2003). Empirical validation of saliency-based model of visual attention. *Electronic Letters Computer Vision and Image Analysis, 3*, 13–24.

Paletta, L., Fritz, G., & Seifert, C. (2005). Q-Learning of sequential attention for visual object recognition from informative local descriptors. In *Proceedings of 22nd international conference on machine learning*.

Palmer, J., Ames, C., & Lindsey, D. (1993). Measuring the effect of attention on simple visual search. *Journal of Experimental Psychology. Human Perception and Performance, 19*, 108–130.

Pang, D., Kimura, A., Takeuchi, T., Yamato, J., & Kashino, K. (2008). A stochastic model of selective visual attention with a dynamic Bayesian network. In *Proceedings of IEEE international conference on multimedia and expo*.

Peters, R. & Itti, L. (2007). Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attentnion. In *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)*.

Phaf, R. H., Van der Heijden, A. H., & Hudson, P. T. W. (1990). SLAM: A connectionist model for attention in visual selection tasks. *Cognitive Psychology, 22*, 273–341.

Postma, E. (1994). Scan: A neural model of covert attention. PhD thesis, Rijksuniversiteit Limburg, Wageningen.

Privitera, C. M., & Stark, L. W. (2000). Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*, 970–982.

Rajashekar, U., van der Linde, I., Bovik, A. C., & Cormack, L. K. (2008). GAFFE: A gaze-attentive fixation finding engine. *IEEE Transactions on Image Processing, 17*, 564–573.

Ramstrom, O. & Christensen, H. I. (2002). Visual attention using game theory. In *2nd international workshop on biologically motivated computer vision*.

Rao, R., Zelinsky, G., Hayhoe, M., & Ballard, D. (2002). Eye mmovement in iconic visual search. *Vision Research, 42*, 1447–1463.

Renninger, L. W., Coughlan, J., Verghese, P., & Malik, J. (2005). An information maximization model of eyemovements. In *Advances in neural information processing systems*.

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

Rosin, P. L. (2009). A simple mothod for detecting salient regions. *Pattern Recognition, 42*, 2363–2371.

Salah, A. A., Alpaydin, E., & Akarun, L. (2002). A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*, 420–425.

Sandon, P. (1990). Simulating visual attention. *Journal of Cognitive Neuroscience, 2*, 213–231.

Seo, H. J., & Milanfar, P. (2009). Static and space-time visual saliency detection by self-resemblance. *Journal of Vision, 9*, 1–27.

Shic, F., & Scassellati, B. (2007). A behavioral analysis of computational models of visual attention. *International Journal of Computere Vision, 73*, 159–177.

Shipp, S. (2004). The brain circuitry of attention. *Trends in Cognitive Sciences, 8*, 223–230.

Sprague, N. & Ballard, D. (2003). Eye movements for reward maximization. In *Neural Information Process Systems*.

Sun, X., Yao, H., Ji, R., Xu, P., Liu, X. & Liu, S. (2010). Visual saliency as sequential eye fixation probability. In *Proceedings of IEEE International Conference on Image Processing*.

Sun, Y., & Fisher, R. (2003). Object-based visual attention for computer vision. *Artificial Intelligence, 146*, 77–123.

Taylor, J. G., & Rogers, M. (2002). A control model of the movement of attention. *Neural Networks, 15*, 309–326.

Treisman, A. (1964). The effect of irrelevant material on the efficiency of selective listening. *American Journal of Psychology, 77*, 533–546.

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

Treue, S., & Martinez-Trujillo, J. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature, 399*, 575–579.

Tsotsos, J., Mylopoulos, J., Covvey, H., & Zucker, S. (1980). A framework for visual motion understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 2*, 563–573.

Tsotsos, J. K. (1992). On the relative complexity of passive vs active visual search. *International Journal of Computer Vision, 7*, 127–141.

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.

Ullman, S. (1984). Visual routines. *Cognition, 18*, 97–159.

van Oeffelen, M. P., & Vos, P. G. (1982). Configurational effects on the enumeration of dots: counting by groups. *Memory & Cognition, 10*, 396–404.

Verghese, P. (2001). Visual search and attention: A signal detection theory approach. *Neuron, 31*, 523–535.

Verma, M., & McOwana, P. W. (2009). Generating customised experimental stimuli for visual search using genetic algorithms shows evidence for a continuum of search efficiency. *Vision Research, 49*, 374–382.

Walther, D., Itti, L., Riesenhuber, M., Poggio, T., & Koch, C. (2002). Attentional selection for object recognition – A gentle way. In *Biologically motivated computer vision*.

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Network, 19*, 1395–1407.

Wang, W., Chen, C., Wang, Y., Jiang, T., Fang, F., & Yao, Y. (2011). Simulating human saccadic scanpaths on natural images. *IEEE Conference on Computer Vision and Pattern Recognition, 11*, 1–15.

Wiesmeyer M., & Laird, J. (1990). A computer model of 2D visual attention. In *Proceedings of the twelfth annual conference of the cognitive science society*.

Yamada, K., Sugano, Y., Okabe, T., Sato, Y., Sugimoto, A. & Hiraki, K. (2010). Can saliency map mmodel predict human egocentric visual attention. In *Proceedings of Asian conference on computer vision*.

Zelinsky, G., Zhang, W., Yu, B., & Chen, X. (2006). The role of top-down and bottom-up processes in guiding eye mmovement during visual search. In *Proceedings of conference on neural information processing systems*.

Zhai, Y., & Shah, M. (2006). Visual attention detection in video sequences using spatiotemporal cues. In *Proceedings of ACM International Conference on Multimedia*.

Zhang, L., Tong, M. H. & Cottrell, G. W. (2009). SUNDAy: Saliency using natural statistics for dynamic analysis of scenes. In *Proceedings of 31th annual cognitive Science Society Conference*, 2009.

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*, 1–20.

Course Readings

Angela, J Y. & Dayan, P. (2004). Inference, attention, and decision in a Bayesian neural architecture. In *Advances in neural information processing systems*, pp. 1577–1584.

Ardid, S., Wang, X., Gomez-Cabrero, D., & Compte, A. (2010). Reconciling coherent oscillation with modulation of irregular spiking activity in selective attention: Gamma-range synchronization between sensory and executive cortical areas. *Journal of Neuroscience*.

Ardid, S., Wang, X. J., & Compte, A. (2007). An integrated microcircuit model of attentional processing in the neocortex. *Journal of Neuroscience, 27*, 8486–8495.

Armstrong, K. M., Fitzgerald, J. K., & Moore, T. (2006). Changes in visual receptive fields with microstimulation of frontal cortex. *Neuron, 50*, 791–798.

Baluch, F., & Itti, L. (2011). Mechanisms of top-down attention. *Trends in Neuroscience, 34*, 210–224.

Borgers, C., Epstein, S., & Kopell, N. (2008). Gamma oscillations mediate stimulus competition and attentional selection in a cortical network model. *PNAS*.

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(185–207), 110–126.

Bruce, N. D. B., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision, 9*, 1–24.

Buia, C. I., & Tiesinga, P. H. (2008). Role of interneuron diversity in the cortical microcircuit for attention. *Journal of Neurophysiology, 99*, 2158–2182.

Chikkerur, S., Serre, T., Tan, C., & Poggio, T. (2010). What and where: A Bayesian inference theory of attention. *Vision Research, 50*, 2233–2247.

Corbetta, M., Patel, G., & Shulman, G. L. (2008). The reorienting system of the human brain: from environment to theory of mind. *Neuron, 58*, 306–324.

Dayan, P. (2009). Load and attentional bayes. In *Advances in neural information processing system*.

Dayan, P., & Yu, A. J. (2003). Uncertainty and learning. *IETE Journal of Research, 49*, 171–182.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective attention. *Annual Review of Neuroscience, 18*, 193–222.

Fallah, M., Stoner, G., & Reynolds, J. (2007). Stimulus-specific competitive selection in macaque extrastriate visual area V4. *PNAS, 104*, 4165–4169.

Isola, P., Xiao, J., Torralba, A., & Oliva, A. (2011). What makes an image memorable? In *IEEE conference on CVPR*.

Koehler, K., Guo, F., Zhang, S., & Eckstein, M. P. (2014). What do saliency models predict? *Journal of Vision, 14*, 1–27.

Krauzlis, R. J., Lovejoy, L. P., & Zenon, A. (2013). Superior colliculus and visual spatial attention. *Annual Review of Neuroscience, 36*, 165–182.

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

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature, 503*, 78–85.

Maunsell, J., & Treue, S. (2006). Feature-based attention in visual cortex. *Trends in Neurosciences, 29–6*, p317–p322.

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., Castelhano, M.S., & Henderson, J.M. (2003). Top-down control of visual attention in object detection. In *ICIP*.

Petersen, S. E., & Posner, M. I. (2012). The attention system of the human brain: 20 years after. *Annual Review of Neuroscience, 35*, 73–89.

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

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 *Proceedings of the 14th international conference on computer vision*.

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*, 3–13.

Rosenholtz, R., Kuzmova, Y. L., & Sherman, A. M. (2011). Visual search for arbitrary objects in real scenes. *Attention, Perception and Psychophysics, 970*.

Salazar, R. F., Dotson, N. M., Bressler, S. L., & Gray, C. M. (2012). Content-specific fronto-parietal synchronization during visual working memory. *Science, 338*, 1097–1100.

Squire, R. F., Noudoost, B., Schafer, R. J., & Moore, T. (2013). Prefrontal contributions to visual selective attention. *Annual Review of Neuroscience, 36*, 451–466.

Sundberg, K. A., Mitchell, J. F., Gawne, T. J., & Reynolds, J. H. (2012). Attention influences single unit and local field potential response latencies in visual cortex. *Journal of Neuroscience, 10*, 10–15.

Sundberg, K. A., Mitchell, J. F., & Reynolds, J. H. (2009). Spatial attention modulates center-surround interactions in macaque visual area v4. *Neuron, 61*, 952–963.

Torralba, A. (2003). Modeling global scene factors in attention. *JOSA A, 20*, 1407–1418.

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. (2006). How the deployment of attention determines what we see. *Visual Cognition, 14*, 411–443.

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

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

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. *Attention*, 13–74.

Wolfe, J. M. (2007). Guided Search 4.0: Current Progress with a model of visual search. *Integrated Models of Cognitive Systems, 10*, 99–119.

Wolfe, J. W., Alvarez, G., Rosenholtz, R., Kuzmova, Y. L., & Sherman, A. M. (2011). Visual search for arbitrary objects in real scenes. *Attention, Perception and Psychophysics*.

Please cite this article in press as: Bylinskii, Z., et al. Towards the quantitative evaluation of visual attention models. *Vision Research* (2015), http://dx.doi.org/10.1016/j.visres.2015.04.007