Vision Research 116 (2015) 258–268

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 5 May 2015

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

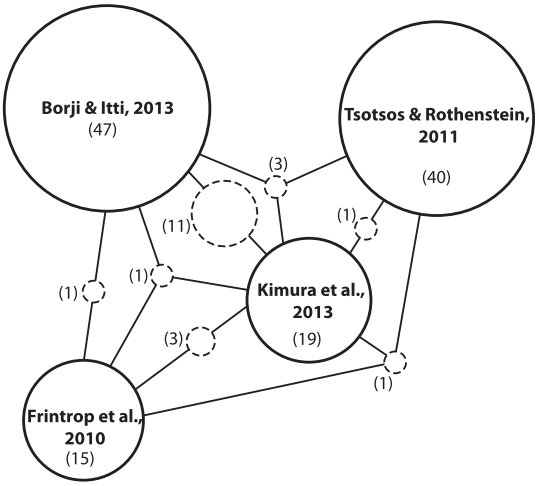

Fig. 1. There are many logical ways of carving up the space of models in the visual attention literature, and different taxonomies/categorizations consider different subsets of models. Here we include four categorizations that cover a total of 142 models (listed in the appendix). Many, but not all, of 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).