Computational models: Bottom-up and top-down aspects

Laurent Itti and Ali Borji, University of Southern California

1 Preliminaries and definitions

Computational models of visual attention have become popular over the past decade, we believe primarily for two reasons: First, models make testable predictions that can be explored by experimentalists as well as theoreticians; second, models have practical and technological applications of interest to the applied science and engineering communities. In this chapter, we take a critical look at recent attention modeling efforts. We focus on *computational models of attention* as defined by Tsotsos & Rothenstein (2011): Models which can process any visual stimulus (typically, an image or video clip), which can possibly also be given some task definition, and which make predictions that can be compared to human or animal behavioral or physiological responses elicited by the same stimulus and task. Thus, we here place less emphasis on abstract models, phenomenological models, purely data-driven fitting or extrapolation models, or models specifically designed for a single task or for a restricted class of stimuli. For theoretical models, we refer the reader to a number of previous reviews that address attention theories and models more generally (Itti & Koch, 2001a; Paletta *et al.*, 2005; Frintrop *et al.*, 2010; Rothenstein & Tsotsos, 2008; Gottlieb & Balan, 2010; Toet, 2011; Borji & Itti, 2012b).

To frame our narrative, we embrace a number of notions that have been popularized in the field, even though many of them are known to only represent coarse approximations to biophysical or psychological phenomena. These include the *attention spotlight* metaphor (Crick, 1984), the role of focal attention in *binding* features into coherent representations (Treisman & Gelade, 1980a), and the notions of an *attention bottleneck*, a *nexus*, and an *attention hand* as embodiments of the attentional selection process (Rensink, 2000; Navalpakkam & Itti, 2005). Further, we cast the problem of modeling attention computationally as comprising at least three facets: *guidance* (that is, which computations are involved in deciding where or what to attend to next?), *selection* (how is attended information segregated out of other incoming sensory information?), and *enhancement* (how is the information selected by attention processed differently than non-selected information?). While different theories and models have addressed all three aspects, most computational models as defined above have focused on the initial and primordial problem of guidance. Thus guidance is our primary focus, and we refer the reader to previous reviews on selection and enhancement (Allport *et al.*, 1993; Desimone & Duncan, 1995; Reynolds & Desimone, 1999; Driver & Frith, 2000; Robertson *et al.*, 2003; Carrasco, 2011). Note that guidance of attention is often thought of as involving *pre-attentive* computations to attract the focus of attention to the next most behaviorally relevant location (hence, attention guidance models might — strictly speaking — be considered pre-attention rather than attention models).

We explore models for exogenous (or bottom-up, stimulus-driven) attention guidance as well as for endogenous (or top-down, context-driven, or goal-driven) attention guidance. Bottom-up models process sensory information primarily in a feed-forward manner, typically applying successive transformations to visual features received over the entire visual field, so as to highlight those locations which contain the most interesting, important, conspicuous, or so-called *salient* information (Koch & Ullman, 1985; Itti & Koch, 2001a). Many, but not all, of these bottom-up models embrace the concept of a topographic saliency map, which is a spatial map where the map value at every location directly represents visual salience, abstracted from the details of why a location is salient or not (Koch & Ullman, 1985). Under the saliency map hypothesis, the task of a computational model is then to transform an image into its spatially corresponding saliency map, possibly also taking into account temporal relations between successive video frames of a movie (Itti *et al.*, 1998). Many models thus attempt to provide an operational definition of salience in terms of some image transform or some importance operator that can be applied to an image and that directly returns salience at every location, as we further examine below.

By far, the bottom-up, stimulus-driven models of attention have been more developed, probably because they are task-free, and thus often require no learning, training, or tuning to open-ended task or contextual information. This makes the definition of a purely bottom-up importance operator tractable. Another attractive aspect of bottom-up models — and especially saliency map models — is that, once implemented, they can easily be applied to any image and yield some output that can be tested against human or animal experimental data. Thus far, the most widely used test to validate predictions of attention models has been direct comparison between model output and eye movements recorded from humans or animals watching the same stimuli as given to the models (see, e.g., among many others, (Parkhurst *et al.*, 2002; Itti, 2006; Le Meur *et al.*, 2006)). Recently, several standard benchmark image and video datasets with corresponding eye movement recordings from pools of observers have been adopted, which greatly facilitates quantitative comparisons of computational models (Carmi & Itti, 2006; Judd *et al.*, 2009; Toet, 2011; Li *et al.*, 2011; Borji *et al.*, 2012b). It is important to note, however, that several alternative measures have been employed as well (e.g., mouse clicks, search efficiency, reaction time, optimal information gain, scanpath similarity; see (Eckstein *et al.*, 2009; Borji & Itti, 2012b)). One caveat of metrics that compare model predictions to eye movements is that the distinction between covert and overt attention is seldom explicitly addressed by computational models: models usually produce as their final result a saliency map without further concern of how such map may give rise to an eye movement scanpath (Noton & Stark, 1971) (but see (Itti *et al.*, 2003; Ma & Deng, 2009; Sun *et al.*, 2012) and active vision/robotics systems like (Orabona *et al.*, 2005; Frintrop, 2006; Belardinelli *et al.*, 2006; Ajallooeian *et al.*, 2009)). Similarly, biases in datasets, models, and/or behavior may affect the comparison results (Tatler & Vincent, 2009; Tseng *et al.*, 2009; Bonev *et al.*, 2012). While these are important for specialist audiences, here we should just remember that quantitative model evaluation metrics based on eye movements exist and are quite well established, although they remain approximate and should be used with caution (Tatler *et al.*, 2005). Beyond metric issues, one last important consideration to ensure a valid direct comparison between model and eye movements is that conditions should be exactly matched between the model run and the experimental participants, which has not always been the case in previous work, as we discuss below.

While new bottom-up attention models are constantly proposed which rely on novel ways of analyzing images to determine the most interesting or salient locations (we list over 50 of them below), several research efforts have also started to address the more complex problem of building top-down models that are tractable and can be implemented computationally. In the early days, top-down models have been mostly descriptive as they operated at the level of conceptual entities (e.g., collections of objects present in a scene, to be evaluated in terms of a mental model of which objects might be more important to a particular task of interest (Ballard *et al.*, 1995)), and hence have lacked generalized computational implementations (because, for example, algorithms that can robustly recognize objects in scenes were not available). As some of these hurdles have been alleviated by the advent of powerful new theories and algorithms for object recognition, scene classification, decision making under uncertainty, machine learning, and formal description languages that can capture task descriptions and background knowledge, exciting new models are beginning to emerge which significantly surpass purely bottom-up approaches in their ability to predict attention behavior in complex situations. Implementing complete, autonomous computational models of top-down attention is particularly important to our theoretical understanding of task and context effects on attention, as these implementations remind us that some assumptions often made to develop abstract models may not give rise to tractable computational models (e.g., it is easy to note how humans are able to directly fixate a jar of jam when it is the next object required to make a sandwich (Land & Hayhoe, 2001) — but how did they know that a jar is present and where it is located, if not through some previous bottom-up analysis of the visual environment?). As we further discuss in this chapter, these new computational models also often blur the somewhat artificial dichotomy between bottom-up and top-down processing, since the so-called top-down models do rely to a large extent onto a bottom-up flow of incoming information that is merged with goals and task demands to give rise to the decision of where to attend next.

To frame the concepts exposed so far into a broader picture, we refer to Figure 1 as a possible anchor to help organize our thoughts and discussions of how different elements of a visual scene understanding system may work together in the primate brain. In practice, few system-level efforts have included all components mentioned in Figure 1, and most of our discussion will focus on computational models that implement parts of such a system.

In what follows, we first examine in more details the key concepts of early bottom-up attention models

(Section 2), and then provide an overview and comparison of many subsequent models that have provided new exciting insight into defining and computing bottom-up salience and attention (Section 3). We then turn to top-down models (Section 4), first motivating them from experimental evidence, and then examining in turn top-down models that modulate feature gains (Section 4.1), that derive spatial priors from the gist of the visual scene (Section 4.2), and that implement more complex information foraging and decision making schemes (Section 4.3). Finally, we discuss in Section 5 lessons learned from these models, including on the nature and interaction between bottom-up and top-down processes, and promising directions towards the creation of even more powerful combined bottom-up and top-down models.

2 Early bottom-up attention concepts and models

Early attention models have been primarily influenced by the Feature Integration Theory (Treisman & Gelade, 1980b), according to which incoming visual information is first analyzed by early visual neurons which are sensitive to elementary visual features of the stimulus (e.g., colors, orientations, etc). This analysis, operated in parallel over the entire visual field and at multiple spatial and temporal scales, gives rise to a number of cortical feature maps, where each map represents the amount of a given visual feature at any location in the visual field. Attention is then the process by which a region in space is selected and features within that region are re-assembled or bound back together to yield more complex object representations (Figure 2.a). Koch and Ullman (1985) extended the theory by advancing the concept of a single topographic and scalar *saliency map*, receiving inputs from the feature maps, as a computationally efficient representation upon which to operate the selection of where to attend next: A simple maximum-detector or *winner-take-all* neural network (Arbib & Didday, 1971) was proposed which would simply pick the next most salient location as the next attended one, while an active *inhibition-of-return* (Posner, 1980) mechanism would later inhibit that location and thereby allow attention to shift as the winner-take-all network would pick the next most salient location (Figure 2.b). From these ideas, a number of fully computational models started to be developed (e.g., Figure 2.c,d).

At the core of these early models is the notion of visual salience, a signal that is computed in a stimulus-driven manner and which indicates that some location is significantly different from its surroundings and is worthy of attention. Early models computed visual salience from bottom-up features in several feature maps, including luminance contrast, red-green and blue-yellow color opponency, and oriented edges (Itti *et al.*, 1998). While visual salience is sometimes carelessly described as a physical property of a visual stimulus, it is important to remember that salience is the consequence of an interaction of a stimulus with other stimuli, as well as with a visual system (biological or artificial). For example, a color-blind person will have a dramatically different experience of visual salience than a person with normal color vision, even when both look at exactly the same colorful physical scene. Nevertheless, because visual salience is believed to primarily arise from fairly low-level and stereotypical computations in the early stages of visual processing, the factors contributing to salience are generally quite comparable from one observer to the next, leading to similar experiences across a range of observers and of viewing conditions.

The essence of salience lies in enhancing the neural and perceptual representation of locations whose local visual statistics significantly differ from the broad surrounding image statistics, in some behaviorally relevant manner. This basic principle is intuitively motivated as follows. Imagine a simple search array as depicted in Figure 2.e, where one bar pops-out because of its unique orientation. Now imagine examining a feature map which is tuned to stimulus intensity (luminance) contrast: because there are many white bars on a black background, early visual neurons sensitive to local intensity contrast will respond vigorously to each of the bars (distractors and target alike, since all have identical intensity). Based on the pattern of activity in this map, in which essentially every bar elicits a strong peak of activity, one would be hard pressed to pick one location as being clearly more interesting and worthy of attention than any of the others. Intuitively, hence, one might want to apply some normalization operator N(.) which would give a very low overall weight to this map's contribution to the final saliency map. The situation is quite different when examining a feature map where neurons are tuned to local vertical edges. In this map, one location (where the single roughly vertical bar is) would strongly excite the neural feature detectors, while all other locations would elicit much weaker responses. Hence, one location clearly stands out and hence becomes an obvious target for attention. It would be desirable in this situation that the normalization operator N(.) give a high weight to this map's

contribution to the final saliency map (Itti *et al.*, 1998; Itti & Koch, 2000; Itti & Koch, 2001a).

Early bottom-up attention models have created substantial interest and excitement in the community, especially as they were shown to be applicable to an unconstrained variety of stimuli, as opposed to more traditional computer vision approaches at the time, which often had been designed to solve a specific task in a specific environment (e.g., detect human faces in photographs taken from a standing human viewpoint (Viola & Jones, 2001)). Indeed, no parameter tuning nor any prior knowledge related to the contents of the images or video clips to be processed was necessary for any of many early results, as the exact same model processed psychophysical stimuli, filmed outdoors scenes, Hollywood movie footage, video games, and robotic imagery. This gave rise to model prediction results that included, for example, the reproduction by Itti *et al.*'s model of human behavior in visual search tasks (e.g., pop-out versus conjunctive search (Itti & Koch, 2000)); demonstration of strong robustness to image noise (Itti *et al.*, 1998); automatic detection of traffic signs and other salient objects in natural environments filmed by a consumer-grade color video camera (Itti & Koch, 2001b); the detection of pedestrians in natural scenes (Miau *et al.*, 2001); and of military vehicles in overhead imagery (Itti *et al.*, 2001); and — most importantly — the widely demonstrated ability of the model to predict where humans look when freely images or videos that range from search arrays to fractals to satellite images to everyday indoors and outdoors scenes (Parkhurst *et al.*, 2002; Peters *et al.*, 2005).

3 Flourishing of bottom-up models

Following initial success, many research groups started exploring the notions of bottom-up attention and visual salience, which has given rise to many new computational models. We summarize 53 bottom-up models along 13 different factors in Figure 3. A thorough examination of all models is certainly not feasible in this limited space. Instead, we highlight below the main trends in seven categories that span from strong inspiration from biological vision to more abstract mathematical definitions and implementations of the concept of saliency. Models can coarsely be categorized as follows (but also see (Tsotsos & Rothenstein, 2011) for another possible taxonomy). Please note that some models fall under more than one category.

Cognitive models. Research into saliency modeling escalated after Itti *et al.*'s (1998) implementation of Koch and Ullman's (1985) computational architecture based on the Feature Integration Theory (Treisman & Gelade, 1980b). In cognitive models, which were the first ones to approach the problem of algorithmically computing saliency in arbitrary digital images, an input image is decomposed into a set of feature maps across spatial scales which are then linearly or non-linearly normalized and combined to form a master saliency map. An important element of this theory is the idea of center-surround which defines saliency as distinctiveness of an image region to its immediate surroundings. Almost all saliency models are directly or indirectly inspired by cognitive concepts of visual attention (e.g., (Le Meur *et al.*, 2006; Marat *et al.*, 2009)).

Information-theoretic models. Stepping back from biological implementation machinery, models in this category are based on the premise that localized saliency computations serve to maximize information sampled from one's environment. These models assign higher saliency values to scene regions with rare features. Information of visual feature F is $I(F) = -log\ p(F)$ which is inversely proportional to the likelihood of observing F (i.e., $p(F)$). By fitting a distribution $P(F)$ to features (e.g., using Gaussian Mixture Model or Kernels), rare features can be immediately found by computing $P(F)^{-1}$ in an image. While in theory using any feature space is feasible, usually these models (inspired by efficient coding representations in visual cortex) utilize a sparse set of basis functions (using ICA filters) learned from a repository of natural scenes. Some basic approaches in this domain are AIM (Bruce & Tsotsos, 2005), Rarity (MANCAS, 2007), LG (Local + Global image patch rarity) (Borji & Itti, 2012a), and incremental coding length models (Hou & Zhang, 2008).

Graphical models. Graphical models are generalized Bayesian models which have been employed for modeling complex attention mechanisms over space and time. Torralba (Torralba, 2003) proposed a Bayesian approach for modeling contextual effects on visual search which was later adopted in the SUN model (Zhang *et al.*, 2008) for fixation prediction in free viewing. Itti and Baldi (Itti & Baldi, 2005) defined surprising stimuli as those which significantly change beliefs of an observer. Harel *et al.* (GBVS) (Harel *et al.*, 2007) propagated similarity of features in a fully connected graph to build a saliency map. Avraham (Avraham & Lindenbaum, 2010), Jia Li *et al.*, (Li & Fei-Fei, 2010), and Tavakoli *et al.* (Rezazadegan Tavakoli *et al.*, 2011), have also exploited Bayesian concepts for saliency modeling.

Decision theoretic models. This interpretation states that attention is driven optimally with respect to the end task. Gao and Vasconcelos (Gao & Vasconcelos, 2004) argued that for recognition, salient features are those that best distinguish a class of objects of interest from all other classes. Given some set of features $X = \{X_1, \cdots, X_d\}$, at locations l, where each location is assigned a class label Y with $Y_l = 0$ corresponding to background and $Y_l = 1$ indicates objects of interest, saliency is then a measure of mutual information (usually Kullback-Leibler divergence (KL)), computed as $I(X, Y) = \sum_{i=1}^{d} I(X_i, Y)$. Besides having good accuracy in predicting eye fixations, these models have been very successful in computer vision applications (e.g., anomaly detection and object tracking).

Spectral-analysis models. Instead of processing an image in the spatial domain, these models derive saliency in the frequency domain. This way, there is no need for image processing operations such as center-surround or segmentation. Hou and Zhang (Hou & Zhang, 2007) derive saliency for an image with amplitude $\mathcal{A}(f)$ and phase $\mathcal{P}(f)$ as follows. The log spectrum $\mathcal{L}(f)$ is computed from the down-sampled image. From $\mathcal{L}(f)$, the spectral residual $\mathcal{R}(f)$ is obtained by multiplying $\mathcal{L}(f)$ with $h_n(f)$ which is an $n \times n$ local average filter and subtracting the result from itself. Saliency map is then the inverse Fourier transform of the exponential of amplitude plus phase (i.e., $\mathcal{S}(x) = \mathcal{F}^{-1}\big[exp\big(\mathcal{R}(f) + \mathcal{P}(f)\big)\big]$). The saliency of each point is squared to indicate the estimation error and is then smoothed with a Gaussian filter for better visual effect. Bian and Zhang (Bian & Zhang, 2009) and Guo *et al.* (Guo & Zhang, 2010) proposed spatio-temporal models in the spectral domain.

Pattern classification models. Models in this category use machine learning techniques to learn "stimuli-saliency" mappings from image features to eye fixations. They estimate saliency s; $p(s|f)$ where f is a feature vector which could be the contrast of a location and its surrounding neighborhood. Kienzle *et al.* (Kienzle *et al.*, 2007), Peters and Itti (Peters & Itti, 2007), and Judd *et al.* (2009) used image patches, scene gist, and a vector of several features at each pixel, respectively and used classical SVM and Regression classifiers for learning saliency. In an extension of Judd model, Borji (2012) showed that using a richer set of features including bottom-up saliency maps of other models and within-object regions (e.g., eye within faces) along with a boosting classifier leads to higher fixation predicting accuracy. Tavakoli *et al.* (Rezazadegan Tavakoli *et al.*, 2011), used sparse sampling and kernel density estimation to estimate the above probability in a Bayesian framework. Note that some of these models may not be purely bottom-up since they use features that guide top-down attention, for example faces or text (Judd *et al.*, 2009; Cerf *et al.*, 2008).

Other models. Some other models exist that do not easily fit into our categorization. For example, Seo and Milanfar (Seo & Milanfar, 2009) proposed self-resemblance of local image structure for saliency detection. The idea of decorrelation of neural response was used for a normalization scheme in the Adaptive Whitening Saliency (AWS) model (Garcia-Diaz *et al.*, 2009). Kootstra *et al.* (Kootstra *et al.*, 2008) developed symmetry operators for measuring saliency and Goferman *et al.* (Goferman *et al.*, 2010) proposed a context-aware saliency detection model with successful applications in re-targeting and summarization.

An important trend to consider is that over the past years starting from (Liu *et al.*, 2007), models have begun to diverge into two different classes: models of *fixation prediction* and models of *salient region detection*. While the goal of the former models is to predict locations that grab attention, the latter models attempt to segment the most salient object or region in a scene. A saliency operator is usually used to estimate the extent of the object that is predicted to be the most likely first attended object. Evaluation is often done by measuring precision-recall of saliency maps of a model against ground-truth data (explicit saliency judgments of subjects by annotating salient objects or clicking on locations). Some models in two categories have compared themselves against each other, without being aware of the distinction.

Figure 3 shows a list of models and their properties according to thirteen qualitative criteria derived from behavioral and computational studies. The majority (53 out of 65) of covered attention models consists of bottom-up models, indicating that at least from a computational perspective it is easier to formulate attention guidance mechanisms based on low-level image features. This is reinforced by the existence of several established benchmark datasets and standard evaluation scores for bottom-up models. The situation is the opposite for top-down attention modeling although we have recently initiated an effort to share data and code (Borji *et al.*, 2012a; Borji *et al.*, 2012c).

A brief comparison of saliency maps of 26 models on a few test images (Figure 4) shows large differences in appearance of the maps generated by different models. Some models generate very sparse maps while others are smoother. This makes fair model comparison a challenge since some scores may be influenced by

smoothness of a map (Tatler *et al.*, 2005). Recently, Borji *et al.* (2012b) performed a detailed investigation of models to quantify their correlations with human attentional behavior. This study suggests that so far the so-called "shuffled AUC (Area Under the ROC Curve)" score (Zhang *et al.*, 2008) is the most robust (this score uses distributions of human fixations on other stimuli than the one being scored to establish a baseline, which attenuates the effects of certain biases in eye movements datasets, the strongest being a bias towards looking preferentially near the center of any image). Results are shown in Figure 5. This model evaluation shows a gap between current models and human performance. This gap is smaller for some datasets, but overall exists. Discovering and adding more top-down features to models will hopefully boost their performance. The analysis also shows that some models are very effective (e.g., HouNIPS, Bian, HouCVPR, Torralba, and Itti-CIO2 in Figure 5) and also very fast providing a trade-off between accuracy and speed necessary for many applications.

Despite past progress in bottom-up saliency modeling and fixation prediction while freely viewing natural scenes, several open questions remain that should be answered in the future. The most confusing one is that of "center bias," whereby humans often appear to preferentially look near an image's center. It is believed to be largely caused by stimulus bias (e.g., photographer bias, whereby photographers tend to frame interesting objects near the image center). Collecting fixation datasets with no or less center bias, and studying its role on model evaluation needs to be addressed with natural scenes (see, e.g., (Parkhurst *et al.*, 2002; Peters *et al.*, 2005) for unbiased artificial datasets of fractal images). As opposed to saliency modeling on static scenes, the domain of spatio-temporal attention remains less explored (See (Dorr *et al.*, 2010; Wang *et al.*, 2012) for examples). Emphasis should be on finding cognitive factors (e.g., actor, non-actor) rather than simple bottom-up features (e.g., motion, flicker, or focus of expansion), and some of the top-down models discussed below have started to explore how more semantic scene analysis can influence attention. Another aspect is the study of attention on affective and emotional stimuli. Although a database of fixations on emotional images has been gathered by Ramanathan *et al.* (2010), it is still not clear whether current models can be extended to explain such fixations.

4 Top-down guidance of attention by task demands

Research towards understanding the mechanisms of top-down attention has given rise to two broad classes of models: models which operate on semantic content, and models which operate on raw pixels and images. Models in the first category are not fully computational in the sense used in the present chapter, in that they require that an external expert (typically, one or more humans) first pre-processes raw experimental recordings, often to create semantic annotations (e.g., translate from recorded video frames and gaze positions into sequences that describe which objects were being looked at). For example, in a block copying task (Ballard *et al.*, 1995), the observers' algorithm for completing the task was revealed by their pattern of eye movements: first select a target block in the model by fixating it, then find a matching block in the resource pool, then revisit the model to verify the block's position, then fixate the workspace to place the new block in the corresponding position. Other studies have used naturalistic interactive or immersive environments to give high-level accounts of gaze behavior in terms of objects, agents, "gist" of the scene, and short-term memory (Yarbus, 1967; Henderson & Hollingworth, 1999; Rensink, 2000; Land & Hayhoe, 2001; Sodhi *et al.*, 2002; Hayhoe *et al.*, 2003), to describe, for example, how task-relevant information guides eye movements while subjects make a sandwich (Land & Hayhoe, 2001; Hayhoe *et al.*, 2003) or how distractions such as setting the radio or answering a phone affect eye movements while driving (Sodhi *et al.*, 2002).

While such perceptual studies have provided important constraints regarding goal-oriented high-level vision, additional work is needed to translate these descriptive results into fully-automated computational models that can be used in the application domains mentioned above. That is, although the block copying task reveals observers' algorithm for completing the task, it does so only in the high-level language of "workspace" and "blocks" and "matching." In order for a machine vision system to replicate human observers' ability to understand, locate, and exploit such visual concepts, we need a "compiler" to translate such high-level language into the assembly language of vision—that is, low-level computations on a time-varying array of raw pixels. Unfortunately, a general computational solution to this task is tantamount to solving computer vision.

From behavioral and in particular eye-tracking experiments during execution of real-world tasks, several

key computational factors can be identified which can be implemented in computational models (Figure 6):

- **Spatial biases**, whereby a given high-level task or top-down set may make some region of space more likely to contain relevant information. For example, when the task is to drive, it is important to keep our eyes on the road (Figure 6.a). We describe below how bottom-up attention models can be enhanced by considering such task-driven spatial constraints, for example to suppress salient stimuli that lie outside the task-relevant region of visual space. These models are motivated by both psychophysical and physiological evidence of spatial biasing of attention based on both short-term and long-term top-down cues (Chun & Jiang, 1998; Summerfield *et al.*, 2011), resulting in enhancement of attended visual regions and suppression of the un-attended ones (Brefczynski & DeYoe, 1999; Kastner *et al.*, 1999);

- **Feature biases**, whereby the task may dictate that some visual features (e.g., some colors) are more likely associated with items of interest than other features (Figure 6.b). Bottom-up models can also be enhanced to account for feature biases, for example by modulating according to top-down goals the relative weights by which different feature maps contribute to a saliency map (e.g., when searching for a blue item, increase the gain of blue-selective feature maps). These models also are motivated by experimental studies of so-called feature-based attention (Treue & Martinez Trujillo, 1999; Saenz *et al.*, 2002; Zhou & Desimone, 2011; Martinez-Trujillo, 2011) and, in particular, recent theories and experiments investigating the role of the pulvinar nucleus in carrying out such biases (Baluch & Itti, 2011; Saalmann *et al.*, 2012);

- **Object-based and cognitive biases**, whereby knowing about objects, about how they may interact with each other, and about how they obey the laws of physics such as gravity and friction, may help humans make more efficient decisions of where to attend next to achieve a certain top-down goal (e.g., playing cricket, Figure 6.c, or making a sandwich, Figure 6.d). Models can be taught how to recognize objects and possibly other aspects of the world, to enable semantic reasoning that may give rise to these more complex top-down attention behaviors. These models are also motivated by recent experimental findings (Võ & Henderson, 2009; Schmidt & Zelinsky, 2009; Hwang *et al.*, 2011).

We review top-down models that have implemented these strategies below. Although spatial biases have historically been studied first, we start with feature biasing models as those are conceptually simpler extensions to the bottom-up models described in the previous sections.

4.1 Top-down biasing of bottom-up feature gains

A simple strategy to include top-down influences in a computational attention model is to modulate the low-levels of visual processing of the model according to the top-down task demands. This embodies the concept of *feature-based attention*, whereby increased neural response can be detected in monkeys and humans to visual locations which contain features that match a feature of current behavioral interest (e.g., locations that contain upward moving dots when the animal's task is to monitor upward motion (Treue & Martinez Trujillo, 1999; Saenz *et al.*, 2002; Zhou & Desimone, 2011).

While the idea of top-down feature biases was already present in early conceptual models like the *Guided Search* theory (Wolfe, 1994) and *FeatureGate* (Cave, 1999), the question for computational modelers has been how exactly the feature gains should be adjusted to yield optimal expected enhancement of a desired target among unwanted distractors (Figure 7). Earlier models have used supervised learning techniques to compute feature gains from example images where targets of interest had been manually indicated (Itti & Koch, 2001b; Frintrop *et al.*, 2005; Borji *et al.*, 2011). A more recent approach uses eye movement recordings to determine these weights (Zhao & Koch, 2011). Interestingly, it has recently been proposed that an optimal set of weight can be computed in closed-form given distributions of expected features for both targets of interest to the task and irrelevant clutter or distractors. In this approach, each feature map is characterized by a target-to-distractor response ratio (or signal-to-noise ratio, SNR), and feature maps are simply assigned a weight that is inversely proportional to their SNR (Navalpakkam & Itti, 2006; Navalpakkam & Itti, 2007). In addition to giving rise to a fully computational model, this theory has also been found to explain many aspects of human guided search behavior (Navalpakkam & Itti, 2007; Serences & Saproo, 2010)

Note how the models of Figure 7 start introducing a blur between the notions of bottom-up and top-down processing. In these models, indeed, the effect of top-down knowledge is to modulate the way in

which bottom-up computations are carried out. This opens the question of whether a pure bottom-up state may ever exist, and what the corresponding gain values may be (e.g., unity as has been assumed in many models?). These questions are also being raised by a number of recent experiments (Theeuwes, 2010; Awh *et al.*, 2012), and are further discussed below.

4.2 Spatial priors and scene context

In a recent model, Ehinger *et al.* (2009) investigated whether a model that, in addition to a bottom-up saliency map, learns spatial priors about where people may appear and a feature prior about what people may look like, would better predict gaze patterns of humans searching for people. Indeed, several earlier studies had suggested that bottom-up models, while widely demonstrated to correlate with human fixations during free viewing, may not well predict fixations of participants once they are given a top-down task, for example a search task (Zelinsky *et al.*, 2006; Foulsham & Underwood, 2007; Henderson *et al.*, 2007; Einhäuser *et al.*, 2008). One may argue, however, especially in the light of our above discussion of whether pure bottom-up salience is a valid concept, that in these experiments the models were at an unfair disadvantage: Human participants had been provided with some information which had not been communicated to models (e.g., search for a specific target, shown for 1 second before the search (Zelinsky *et al.*, 2006); search of objects in a category or for a specific object (Foulsham & Underwood, 2007); search for the small bullseye pattern or for a local higher-contrast region (Einhäuser *et al.*, 2008); or count people (Henderson *et al.*, 2007)). Ehinger *et al.* addressed this by proposing a model that combines three sources of information (Figure 8.a): First, a "scene context" map was derived from learning the associations between holistic or global scene features (coarsely capturing the gist of the scene (Torralba, 2003)) and the locations where humans appeared in scenes with given holistic features (trained over 1880 example images). This map, which is of central interest to this section of our chapter, thus learned the typical locations where humans were expected to appear in different views of street scenes. This learning step produces a prior on locations that can be used to filter out salient responses in locations that are highly unlikely to contain the target (e.g., in the sky, assuming that no human was seen flying in the training dataset). Second, a person detector was run in a sliding window manner over the entire image, creating a "target features" map that highlighted locations that closely look like humans. This provides an alternative to learning feature gains as discussed above; instead, an object detector algorithm is trained for the desired type of target. While possibly more efficient than gain modulation, this approach suffers from lower biological plausibility. (See (Rao *et al.*, 2002; Orabona *et al.*, 2005) for related models). Third and finally, a standard bottom-up "saliency map" provided additional candidate locations (also see (Oliva *et al.*, 2003) for earlier related work, integrating only saliency and scene context). Ehinger *et al.* found that the model which combined all three maps outperformed any of the three component models taken alone (Figure 8.b).

In a related model, Peters & Itti (Peters & Itti, 2007) also used a combination of bottom-up saliency maps and top-down spatial maps derived from the holistic gist of the scene, but their top-down maps were directly learned from eye movements of human observers, playing the same 3D video games as would be used for testing (games included driving, exploration, flight combat, etc.; note that since players control the game's virtual camera viewpoint, each run of such game gives rise to a unique set of viewpoints and of generated scenes). The bottom-up component of this model is based on the Itti-Koch saliency model (Itti *et al.*, 1998), which predicts interesting locations based on low-level visual features such as luminance contrast, color contrast, orientation, and motion. The top-down component is based on the idea of "gist," which in psychophysical terms is the ability of people to roughly describe the type and overall layout of an image after only a very brief presentation (Li *et al.*, 2002), and to use this information to guide subsequent target searches (Torralba, 2003). This model (Figure 8.c) decomposes each video frame into a low-level image signature intended to capture some of the properties of "gist" (Siagian & Itti, 2007), and learns to pair the low-level signatures from a series of video clips with the corresponding eye positions; once trained, it generates predicted gaze density maps from the gist signatures of previously unseen video frames. To test these bottom-up and top-down components, we compared their predicted gaze density maps with the actual eye positions recorded while people interactively played video games (Figure 8.d).

4.3 More complex top-down models

Many top-down models have been proposed which include higher degrees of cognitive scene understanding. Already in the late 1990's several models included a top-down component that decided where to look next based on what had been observed so far (e.g., (Rybak *et al.*, 1998; Schill *et al.*, in press); also see (Itti & Koch, 2001a) for review). In robotics, the notion of combining or alternating between different behaviors (such as exploration versus search, or bottom-up versus top-down) has also led to several successful models (Sprague & Ballard, 2003; Forssén *et al.*, 2008; Burattini *et al.*, 2010; Xu *et al.*, 2010). More recently, and our focus here, probabilistic inference and reasoning techniques, very popular in computer vision, have started to be used in attention models.

In many recent models, the saliency map of bottom-up models is conserved as a data-driven source of information for an overarching top-down system (more complicated than the feature or spatial biasing described above). For example, Boccignone & Ferraro (2004) develop an overt attention system where the top-down component is a random walker that follows an information foraging strategy over a bottom-up saliency map. They demonstrate simulated gaze patterns that better match human distributions (Tatler *et al.*, 2011). Interesting related models have been proposed where bottom-up and top-down attention interact through object recognition (Ban *et al.*, 2010; Lee *et al.*, 2011), or by formulating a task as a classification problem with missing features, with top-down attention then providing a choice process over the missing features (Hansen *et al.*, 2011).

Of growing recent interest is the use of probabilistic reasoning and graphical models to explore how several sources of bottom-up and top-down information may combine in a Bayesian-optimal manner. For example, the model of Akamine *et al.* (2012) (also see (Kimura *et al.*, 2008)), which employs probabilistic graphical modeling techniques and considers the following factors, interacting in a dynamic Bayesian network (Figure 9.a): On the one hand, input video frames give rise to deterministic saliency maps. These are converted into stochastic saliency maps via a random process that affects the shape of salient blobs over time (e.g., dynamic Markov random field (Kimura *et al.*, 2008)). An eye focusing map is then created which highlights maxima in the stochastic saliency map, additionally integrating top-down influences from an eye movement pattern (a stochastic selection between passive and active state with a learned transition probability matrix). The authors use a particle filter with Markov chain Monte-Carlo (MCMC) sampling to estimate the parameters; this technique often used in machine learning allows for fast and efficient estimation of unknown probability density functions. Although the top-down component is quite simple in this version of the model, it is easy to see how more sophisticated top-down and contextual influences could be integrated into the dynamic Bayesian network framework of Kimura *et al.* Several additional recent related models using graphical models have been proposed (e.g., (Chikkerur *et al.*, 2010)).

Although few have been implemented as fully computational models, several efforts have started to develop models that perform reasoning over objects or other scene elements to make a cognitive decision of where to look next (Navalpakkam & Itti, 2005; Yu *et al.*, 2008; Beuter *et al.*, 2009; Yu *et al.*, 2012).

A a recent example, using probabilistic reasoning and inference tools, Borji *et al.* (Borji *et al.*, 2012a) introduced a framework to model top-down overt visual attention based on reasoning, in a task-dependent manner, about objects present in the scene and about previous eye movements. They designed a Dynamic Bayesian Network (DBN) that infers probability distributions over attended objects and spatial locations directly from observed data. Two basic concepts in this model are 1) taking advantage of the sequence structure of tasks, which allows to predict the future fixations from past fixations and knowledge about objects present in the scene. Graphical models have indeed been very successful in the past to model sequences with applications in different domains, including biology, time series modeling, and video processing, and 2) computing attention at the object level. Since objects are essential building blocks in scenes, it is reasonable to assume that humans have instantaneous access to task-driven object-level variables (as opposed to only gist-like, scene-global, representations). Briefly, the model works by defining a Bayesian network over object variables that matter for the task. For example, in a video game where one runs a hot-dog stand and has to serve multiple hungry customers while managing the grill, those include raw sausages, cooked sausages, buns, ketchup, etc.(Figure 9.b). Then, existing objects in the scene, as well as the previous attended object, provide evidence toward the next attended object (Figure 9.b). The model also allows to read out which spatial location will be attended, thus allowing one to verify its accuracy against the next actual fixation of the human player. The parameters of the network are learned directly from training data in the same

form as the test data (human players playing the game). This object-based model was significantly more predictive of eye fixations compared to simpler classifier-based models, also developed by the same authors, that map a signature of a scene to eye positions, several state-of-the-art bottom-up saliency models, as well as brute-force algorithms such as mean eye position (Figure 9.c). This points toward the efficacy of this class of models for modeling spatio-temporal visual data in presence of a task and hence a promising direction for future. Probabilistic inference in this model is performed over object-related functions which are fed from manual annotations of objects in video scenes or by state-of-the-art object detection models. (Also see (Sun & Fisher, 2003; Sun *et al.*, 2008) for models that consider objects, although they do not reason about object identities and task-dependent roles).

Finally, several computational models have started to explore making predictions that go beyond simply the next attended location. For example, Peters & Itti (Peters & Itti, 2008) developed a model that monitors in an online manner video frames and eye gaze of humans engaged in 3D video games, computing instantaneous measures of how well correlated the eye is with saliency predictions and with gist-based top-down predictions. They then learn to detect specific patterns in these instantaneous measures, which allows them to predict — up to several seconds in in advance — when players are about to fire a missile in a flight combat game, or to shift gears in a driving game. This model hence in essence estimates the intentions and predicts the future actions of the player. A related recent model was proposed by Doshi & Trivedi (Doshi & Trivedi, 2010) for active vehicle safety and driver monitoring. The system both computes bottom-up and top-down saliency maps from a video feed of the driver's view, and monitors the eye movements of the driver to better predict driver attention and gaze by estimating online the cognitive state and level of distraction of the driver. Using similar principles and adding pattern classification techniques, Tseng *et al.* (2009) have recently introduced a model that uses machine learning to classify, from features collected at the point of gaze over a few minutes of television viewing, control subjects from patients with disorders that affect the attention and oculomotor systems. The model has been successfully applied to elderly subjects (classifying patients with Parkinson's disease vs. controls) as well as children (classifying children with Attention Deficit Hyperactivity Disorder vs. Fetal Alcohol Spectrum Disorder vs. controls well above chance). These recent efforts suggest that eye movement patterns in complex scenes do contain — like a drop of saliva — latent individual biomarkers, which the latest attention modeling and pattern classification techniques are now beginning to reliably decode.

5 Discussion and outlook

Our review shows that tremendous progress has been made in modeling both bottom-up and top-down aspects of attention computationally. Tens of new models have been developed, each bringing new insight into the question of what makes some stimuli more important to visual observers than other stimuli.

Our quantitative comparison of many existing computational models on three standard datasets (Figure 5) prompts at least two reactions: First, it is encouraging to see that several models perform significantly better than trivial models (e.g., a central Gaussian blob) or than older models (e.g., Itti-CIO2 model in Figure 5). Second, however, it is surprising that the ranking of model scores is quite substantially different from the chronological order in which models were published. Indeed, one would typically expect that for a new model to be recognized, it should demonstrate superior performance compared to the state of the art, and actually often this is the case — just using possibly different datasets, scoring metrics, etc. Thus one important conclusion of our study is that carrying out standardized evaluations is important to ensure that the field keeps moving forward (see (Borji, 2010) for a web-based effort in this direction).

Another important aspect of model evaluation is that currently almost all model comparisons and scoring are based on average performance over a dataset of images or video clips, where often the dataset has been hand-picked and may contain significant biases (Torralba & Efros, 2011). It may be more fruitful in the future to focus scoring on the most dramatic mistakes a model might make, or the worst-case disagreement between model and human observers. Indeed, average measures can easily be dominated by trivial cases if those happen often (e.g., we discussed earlier the notion of center bias and how a majority of saccades which humans make are aimed towards the centers of images), and models may be developed which perform well in these cases but miss conceptually important understanding of how attention may operate in the minority of non-trivial cases. In addition, departure from average performance measures may provide richer information

about which aspects of attention are better captured by a given model (e.g., some models may perform better on some sub-categories or even instances of images than others).

As we described more models, and in particular started moving from bottom-up models to those which include top-down biases, the question arose of whether purely bottom-up models are indeed relevant to real life. In other words, is there such a state of human cognition where a default or unbiased form of salience may be computed and may guide gaze. Many experiments have assumed that free viewing, just telling observers to "watch and enjoy" stimuli presented to them, might be an acceptable approximation to this canonical unbiased state. However, it is trivially clear from introspection that cognition is not turned off during free viewing, and that what we look at in one instant triggers a range of memories, emotions, desires, cognitive inferences, etc. which all will ultimately influence where we look next. In this regard, it has been recently suggested that maybe only the initial volley of activity through visual cortex following stimulus onset may represent such canonical bottom-up saliency representation (Theeuwes, 2010). If such is the case, then maybe comparing model predictions to sometimes rather long sequences of eye movements may be not be the best measure of how well a model captures this initial purely bottom-up attention.

Models where top-down influences serve to bias the bottom-up processing stages have also blurred the line between bottom-up and top-down. In fact, this is an important reminder that bottom-up and top-down influences are not mutually exclusive and do not sum to give rise to attention control (Awh *et al.*, 2012). Instead, bottom-up and top-down often agree: The actor cognitively identified as the protagonist in a video clip may also move in such ways that he is the most salient. In fact, today's bottom-up may be nothing more than our former generations' top-down. Indeed, some bottom-up models have successfully integrated high-level features such as human face detectors into their palette of feature maps (Judd *et al.*, 2009), which blurs again the line between bottom-up and top-down (these features are computed in a bottom-up manner from the image, but their very presence in a model is based on top-down knowledge that humans do strongly tend to look at faces in images (Cerf *et al.*, 2009)).

When there is a task, top-down influences on attention are often believed to dominate, though this remains controversial and depends both on the task and on the quantification method used (Zelinsky *et al.*, 2006; Foulsham & Underwood, 2007; Henderson *et al.*, 2007; Einhäuser *et al.*, 2008; Greene *et al.*, 2012). In human vision, we should not forget the following: Purely top-down attention (i.e., making a purely volitional eye movement unrelated to any visual stimulus) is not a generally viable model, except maybe in blind persons. Others may make pure top-down eye movements from time to time, but certainly not always — that is, no matter how strongly one believes that top-down influences dominate, in the end controlling visual attention is a visually-guided behavior, and, as such, it is dependent on visual stimuli. This is important for future modeling efforts, as they attempt to tackle more complex tasks and situations, such as making a sandwich (Figure 6.d): A person may indeed look at the jar of jam because it is the next required object for the task (top-down guidance towards the jam). However, how did that person know where the jam is? In most cases, some bottom-up analysis (maybe in the past) must have provided that information (except maybe if the person was told where the jam is). Thus, modeling human behavior in complex tasks will likely require very careful control over the experimental setup, so that human participants are not given more information or additional priors that are not communicated to models (e.g., let a person look around before the task begins; see (Foulsham, 2012) for recent relevant data). This consideration echoes our earlier remark about making fair comparisons: If a model is not given the same information as a human participant (e.g., the model is not biased towards a search target or is not allowed to explore a scene before the task begins), likely the model will not perform as well, but we will also learn very little from such an experiment.

Another important challenge for models briefly mentioned above is dealing with sequence in eye movement data (i.e., scanpath) and with how to capture temporality in saccades. When comparing models, a model might be favored not only if it can predict exact saccade locations, but also their ordering and their individual times of occurrence. In free viewing, in spite of past efforts (Privitera & Stark, 2000), it is still not clear whether such sequential information is a strong factor of attention control and to what extent it depends on the subject or the asked question (Yarbus, 1967). Despite this, recently some researchers (e.g., (Wang *et al.*, 2011)) have tried to develop models and scores to explain sequences of saccades. As opposed to free viewing, it seems that there is much more temporal information in saccades in presence of a task. For instance, assume an observer is viewing videos of two different tasks such as sandwich making or driving. It probably should not be very difficult to decode the task just from the sequential pattern of eye movements. This means that the task governs sequence of saccades when there is a task. In free-viewing, however, when subjects are asked

to watch a static scene freely there might not be a unique instruction making them to saccade sequentially to certain places. Even if subjects are asked to watch the scene under different questions, chances are that sequence may not help to decode the task (Greene *et al.*, 2012).

Our survey shows that the remaining gap between man and machine seems to a large extent to be in 3D+time scene understanding, which includes reconstruction of the 3D geometry of the scene, understanding temporal sequences of events, simulation and extrapolation of physics over time in that 3D environment (e.g., to extrapolate the trajectory of a ball as in Figure 6.c), and so on. This requires some degree of machine vision and scene understanding which is not yet solved in the general case. This means that future computational models of attention will need to bring to bear sophisticated machine vision algorithms for scene understanding, to provide the necessary parsing of visual inputs into tokens that can be reasoned upon and prioritized by attention.

Acknowledgements

Supported by the National Science Foundation (grant numbers BCS-0827764 and CMMI-1235539), the Army Research Office (W911NF-11-1-0046 and 62221-NS), the U.S. Army (W81XWH-10-2-0076), and Google. The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States government or any agency thereof.

References

Achanta, R., Hemami, S., Estrada, F., & Susstrunk, S. 2009. Frequency-tuned salient region detection. *Pages 1597–1604 of: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE.

Ajallooeian, M., Borji, A., Araabi, BN, Ahmadabadi, M.N., & Moradi, H. 2009. Fast hand gesture recognition based on saliency maps: An application to interactive robotic marionette playing. *Pages 841–847 of: Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on*. IEEE.

Akamine, K., Fukuchi, K., Kimura, A., & Takagi, S. 2012. Fully Automatic Extraction of Salient Objects from Videos in Near Real Time. *The Computer Journal*, **55**(1), 3–14.

Allport, A., Meyer, DE, & Kornblum, S. 1993. Attention and control: Have we been asking the wrong questions? A critical review of twenty-five years. *Attention and Performance XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience, MIT Press, Cambridge, Mass*, 183–218.

Arbib, M.A., & Didday, R.L. 1971. The Organization of Action-Oriented Memory for a Perceiving System. Part I: The Basic Model. *Journal of Cybernetics*, **1**(1), 3–18.

Avraham, T., & Lindenbaum, M. 2010. Esaliency (extended saliency): Meaningful attention using stochastic image modeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **32**(4), 693–708.

Awh, E., Belopolsky, A.V., & Theeuwes, J. 2012. Top-down versus bottom-up attentional control: a failed theoretical dichotomy. *Trends in Cognitive Sciences*.

Aziz, M.Z., & Mertsching, B. 2008. Fast and robust generation of feature maps for region-based visual attention. *Image Processing, IEEE Transactions on*, **17**(5), 633–644.

Ballard, D.H., Hayhoe, M.M., & Pelz, J.B. 1995. Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, **7**(1), 66–80.

Baluch, F., & Itti, L. 2011. Mechanisms of Top-Down Attention. *Trends in Neurosciences*, **34**(March), 210–224. Front cover, April 2011.

Ban, S.W., Lee, I., & Lee, M. 2008. Dynamic visual selective attention model. *Neurocomputing*, **71**(4), 853–856.

Ban, S.W., Kim, B., & Lee, M. 2010. Top-down visual selective attention model combined with bottom-up saliency map for incremental object perception. *Pages 1–8 of: Neural Networks (IJCNN), The 2010 International Joint Conference on*. IEEE.

Belardinelli, A., Pirri, F., & Carbone, A. 2006. Robot task-driven attention. *Pages 117–128 of: Proceedings of the 2006 international symposium on Practical cognitive agents and robots*. ACM.

Beuter, N., Lohmann, O., Schmidt, J., & Kummert, F. 2009. Directed attention-a cognitive vision system for a mobile robot. *Pages 854–860 of: Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on*. IEEE.

Bian, P., & Zhang, L. 2009. Biological plausibility of spectral domain approach for spatiotemporal visual saliency. *Advances in Neuro-Information Processing*, 251–258.

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

Bonev, B., Chuang, LL, & Escolano, F. 2012. How do image complexity, task demands and looking biases influence human gaze behavior? *Pattern Recognition Letters*.

Borji, A. 2010. *https://sites.google.com/site/saliencyevaluation/*.

Borji, A. 2012. Boosting Bottom-up and Top-down Visual Features for Saliency Estimation. *Pages 438–445 of: 2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

Borji, A., & Itti, L. 2012a (Jun). Exploiting Local and Global Patch Rarities for Saliency Detection. *In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, Rhode Island*.

Borji, A., & Itti, L. 2012b. State-of-the-art in Visual Attention Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Borji, A., Ahmadabadi, M.N., Araabi, B.N., & Hamidi, M. 2010. Online learning of task-driven object-based visual attention control. *Image and Vision Computing*, **28**(7), 1130–1145.

Borji, A., Ahmadabadi, M.N., & Araabi, B.N. 2011. Cost-sensitive learning of top-down modulation for attentional control. *Machine Vision and Applications*, **22**(1), 61–76.

Borji, A., Sihite, D. N., & Itti, L. 2012a (Aug). An Object-based Bayesian Framework for Top-down Visual Attention. *Pages 1529–1535 of: Proc. Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI-12), Toronto, Canada*.

Borji, A., Sihite, D. N., & Itti, L. 2012b. Quantitative Analysis of Human-Model Agreement in Visual Saliency Modeling: A Comparative Study. *IEEE Transactions on Image Processing*.

Borji, A., Sihite, D. N., & Itti, L. 2012c. What/Where to Look Next? Modeling Top-down Visual Attention in Complex Interactive Environments. *IEEE Transactions on Systems, Man, and Cybernetics, Part A - Systems and Humans*.

Brefczynski, J A, & DeYoe, E A. 1999. A physiological correlate of the 'spotlight' of visual attention. *Nat Neurosci*, **2**(4), 370–374.

Bruce, Neil D. B., & Tsotsos, John K. 2005. Saliency Based on Information Maximization. *In: NIPS*.

Burattini, E., Rossi, S., Finzi, A., & Staffa, M. 2010. Attentional modulation of mutually dependent behaviors. *From Animals to Animats 11*, 283–292.

Butko, N.J., & Movellan, J.R. 2009. Optimal scanning for faster object detection. *Pages 2751–2758 of: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE.

Carmi, R., & Itti, L. 2006. The Role of Memory in Guiding Attention during Natural Vision. *Journal of Vision*, **6**(9), 898–914.

Carrasco, M. 2011. Visual attention. *Vision research*.

Cave, K R. 1999. The FeatureGate Model of Visual Selection. *Psychol Res*, **62**, 182–194.

Cerf, M., Harel, J., Einhäuser, W., & Koch, C. 2008. Predicting human gaze using low-level saliency combined with face detection. *Advances in neural information processing systems*, **20**.

Cerf, M., Frady, E.P., & Koch, C. 2009. Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision*, **9**(12).

Chikkerur, S., Serre, T., Tan, C., & Poggio, T. 2010. What and where: A Bayesian inference theory of attention. *Vision research*, **50**(22), 2233–2247.

Chun, M.M., & Jiang, Y. 1998. Contextual Cueing: Implicit Learning and Memory of Visual Context Guides Spatial Attention. *Cognitive Psychology.*, **36**, 28–71.

Crick, F. 1984. Function of the thalamic reticular complex: the searchlight hypothesis. *Proc Natl Acad Sci USA*, **81**, 4586–4590.

Desimone, R., & Duncan, J. 1995. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, **18**(1), 193–222.

Dorr, M., Martinetz, T., Gegenfurtner, K.R., & Barth, E. 2010. Variability of eye movements when viewing dynamic natural scenes. *Journal of vision*, **10**(10).

Doshi, A., & Trivedi, M.M. 2010. Attention estimation by simultaneous observation of viewer and view. *Pages 21–27 of: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE.

Driver, J, & Frith, C. 2000. Shifting baselines in attention research. *Nature Reviews Neuroscience*, **1**, 147–148.

Eckstein, M.P., Peterson, M.F., Pham, B.T., & Droll, J.A. 2009. Statistical decision theory to relate neurons to behavior in the study of covert visual attention. *Vision research*, **49**(10), 1097–1128.

Ehinger, K.A., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. 2009. Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition*, **17**(6-7), 945–978.

Einhäuser, W., Rutishauser, U., & Koch, C. 2008. Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, **8**(2).

Forssén, P.E., Meger, D., Lai, K., Helmer, S., Little, J.J., & Lowe, D.G. 2008. Informed visual search: Combining attention and object recognition. *Pages 935–942 of: Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*. IEEE.

Foulsham, T. 2012. 'Eyes Closed' and 'Eyes Open' Expectations Guide Fixations in Real-World Search. *Pages 1–6 of: Proc. CogSci 2012*.

Foulsham, T., & Underwood, G. 2007. How does the purpose of inspection influence the potency of visual salience in scene perception? *Perception - London*, **36**(8), 1123.

Frintrop, S. 2006. *VOCUS: A visual attention system for object detection and goal-directed search*. Vol. 3899. Springer-Verlag New York Inc.

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

Frintrop, S., Rome, E., & Christensen, H.I. 2010. Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception (TAP)*, **7**(1), 6.

Gao, D., & Vasconcelos, N. 2004. Discriminant saliency for visual recognition from cluttered scenes. *Advances in neural information processing systems*, **17**(481-488), 1.

Gao, D., Han, S., & Vasconcelos, N. 2009. Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **31**(6), 989–1005.

Garcia-Diaz, A., Fdez-Vidal, X., Pardo, X., & Dosil, R. 2009. Decorrelation and distinctiveness provide with human-like saliency. *Pages 343–354 of: Advanced Concepts for Intelligent Vision Systems*. Springer.

Goferman, S., Zelnik-Manor, L., & Tal, A. 2010. Context-aware saliency detection. *Pages 2376–2383 of: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE.

Gottlieb, J., & Balan, P. 2010. Attention as a decision in information space. *Trends in cognitive sciences*, **14**(6), 240–248.

Greene, M.R., Liu, T., & Wolfe, J.M. 2012. Reconsidering Yarbus: A failure to predict observers task from eye movement patterns. *Vision Research*.

Guo, C., & Zhang, L. 2010. A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression. *IEEE Trans. on Image Processing.*, **19**.

Guo, C., Ma, Q., & Zhang, L. 2008. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. *Pages 1–8 of: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE.

Hansen, L.K., Karadogan, S., & Marchegiani, L. 2011. What to measure next to improve decision making? On top-down task driven feature saliency. *Pages 1–7 of: Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB), 2011 IEEE Symposium on*. IEEE.

Harel, J., Koch, C., & Perona, P. 2007. Graph-based visual saliency. *Advances in neural information processing systems*, **19**, 545.

Hayhoe, M.M., Shrivastava, A., Mruczek, R., & Pelz, J.B. 2003. Visual memory and motor planning in a natural task. *Journal of Vision*, **3**(1).

Henderson, J.M., & Hollingworth, A. 1999. High-level scene perception. *Annual review of psychology*, **50**(1), 243–271.

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

Hou, X., & Zhang, L. 2007. Saliency Detection: A Spectral Residual Approach. *In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hou, X., & Zhang, L. 2008. Dynamic Visual Attention: Searching for Coding Length Increments. *Advances in Neural Information Processing Systems (NIPS).*, 681–688.

Hwang, A.D., Wang, H.C., & Pomplun, M. 2011. Semantic guidance of eye movements in real-world scenes. *Vision research*.

Itti, L. 2006. Quantitative Modeling of Perceptual Salience at Human Eye Position. *Visual Cognition*, **14**(4-8), 959–984.

Itti, L., & Baldi, P. F. 2005 (Jun). A Principled Approach to Detecting Surprising Events in Video. *Pages 631–637 of: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Itti, L., & Koch, C. 2000. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, **40**(10-12), 1489–1506.

Itti, L., & Koch, C. 2001a. Computational Modelling of Visual Attention. *Nature Reviews Neuroscience*, **2**(3), 194–203.

Itti, L., & Koch, C. 2001b. Feature Combination Strategies for Saliency-Based Visual Attention Systems. *Journal of Electronic Imaging*, **10**(1), 161–169.

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**(11), 1254–1259.

Itti, L., Gold, C., & Koch, C. 2001. Visual Attention and Target Detection in Cluttered Natural Scenes. *Optical Engineering*, **40**(9), 1784–1793.

Itti, L., Dhavale, N., & Pighin, F. 2003. Realistic Avatar Eye and Head Animation Using a Neurobiological Model of Visual Attention. *Pages 64–78 of:* Bosacchi, B., Fogel, D. B., & Bezdek, J. C. (eds), *Proc. SPIE 48th Annual International Symposium on Optical Science and Technology*, vol. 5200. Bellingham, WA: SPIE Press.

Jodogne, S., & Piater, J. 2007. Closed-loop learning of visual control policies. *Journal of Artificial Intelligence Research*, **28**(1), 349–391.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. 2009. Learning to predict where humans look. *Pages 2106–2113 of: Computer Vision, 2009 IEEE 12th International Conference on*. IEEE.

Kastner, S., Pinsk, M.A., De Weerd, P., Desimone, R., & Ungerleider, L.G. 1999. Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron*, **22**(4), 751–761.

Kienzle, W., Wichmann, F., Schölkopf, B., & Franz, M. 2007. A nonparametric approach to bottom-up visual saliency. *Pages 1–8 of: Proc. NIPS*.

Kimura, A., Pang, D., Takeuchi, T., Yamato, J., & Kashino, K. 2008. Dynamic Markov random fields for stochastic modeling of visual attention. *Pages 1–5 of: Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE.

Koch, C, & Ullman, S. 1985. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol*, **4**(4), 219–27.

Kootstra, G., Nederveen, A., & De Boer, B. 2008. Paying attention to symmetry. *Pages 1115–1125 of: Proceedings of the British Machine Vision Conference (BMVC2008)*.

Land, M.F., & Hayhoe, M. 2001. In what ways do eye movements contribute to everyday activities? *Vision research*, **41**(25-26), 3559–3565.

Land, M.F., & McLeod, P. 2000. From eye movements to actions: how batsmen hit the ball. *Nature neuroscience*, **3**(12), 1340–1345.

Le Meur, O., Le Callet, P., Barba, D., & Thoreau, D. 2006. A coherent computational approach to model bottom-up visual attention. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **28**(5), 802–817.

Lee, S., Oh, J., Park, J., Kwon, J., Kim, M., & Yoo, H.J. 2011. A 345 mW heterogeneous many-core processor with an intelligent inference engine for robust object recognition. *Solid-State Circuits, IEEE Journal of*, **46**(1), 42–51.

Li, F.F., VanRullen, R., Koch, C., & Perona, P. 2002. Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences*, **99**(14), 9596.

Li, J., Tian, Y., Huang, T., & Gao, W. 2011. Multi-Task Rank Learning for Visual Saliency Estimation. *Circuits and Systems for Video Technology, IEEE Transactions on*, 1–1.

Li, L.J., & Fei-Fei, L. 2010. Optimol: automatic online picture collection via incremental model learning. *International journal of computer vision*, **88**(2), 147–168.

Li, Y., Zhou, Y., Yan, J., Niu, Z., & Yang, J. 2010. Visual Saliency Based on Conditional Entropy. *Computer Vision–ACCV 2009*, 246–257.

Liu, T., Sun, J., Zheng, N., Tang, X., & Shum, H. 2007. Learning to Detect a Salient Object. *IEEE Conference on Computer Vision and Pattern Recognition.*

Ma, X., & Deng, Z. 2009. Natural eye motion synthesis by modeling gaze-head coupling. *Pages 143–150 of: Virtual Reality Conference, 2009. VR 2009. IEEE.* IEEE.

Ma, Y.F., Hua, X.S., Lu, L., & Zhang, H.J. 2005. A generic framework of user attention model and its application in video summarization. *Multimedia, IEEE Transactions on*, **7**(5), 907–919.

Mahadevan, V., & Vasconcelos, N. 2010. Spatiotemporal saliency in dynamic scenes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **32**(1), 171–177.

MANCAS, M. 2007. *Computational Attention Towards Attentive Computers.* Presses univ. de Louvain.

Marat, S., Ho Phuoc, T., Granjon, L., Guyader, N., Pellerin, D., & Guérin-Dugué, A. 2009. Modelling spatio-temporal saliency to predict gaze direction for short videos. *International Journal of Computer Vision*, **82**(3), 231–243.

Martinez-Trujillo, J. 2011. Searching for the Neural Mechanisms of Feature-Based Attention in the Primate Brain. *Neuron*, **70**(6), 1025–1028.

McCallum, A.K. 1996. *Reinforcement learning with selective perception and hidden state.* Ph.D. thesis, University of Rochester.

Miau, F., Papageorgiou, C., & Itti, L. 2001. Neuromorphic algorithms for computer vision and attention. *Pages 12–23 of:* Bosacchi, B., Fogel, D. B., & Bezdek, J. C. (eds), *Proc. SPIE 46 Annual International Symposium on Optical Science and Technology*, vol. 4479. Bellingham, WA: SPIE Press.

Milanese, R., Wechsler, H., Gill, S., Bost, J.M., & Pun, T. 1994. Integration of bottom-up and top-down cues for visual attention using non-linear relaxation. *Pages 781–785 of: Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on.* IEEE.

Mourant, R.R., & Rockwell, T.H. 1970. Mapping eye-movement patterns to the visual scene in driving: An exploratory study. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, **12**(1), 81–87.

Murray, N., Vanrell, M., Otazu, X., & Parraga, C.A. 2011. Saliency estimation using a non-parametric low-level vision model. *Pages 433–440 of: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on.* IEEE.

Navalpakkam, V., & Itti, L. 2005. Modeling the influence of task on attention. *Vision Research*, **45**(2), 205–231.

Navalpakkam, V., & Itti, L. 2006 (Jun). An Integrated Model of Top-down and Bottom-up Attention for Optimal Object Detection. *Pages 2049–2056 of: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Navalpakkam, V., & Itti, L. 2007. Search goal tunes visual features optimally. *Neuron*, **53**(4), 605–617. Also see commentary / preview entitled "Paying Attention to Neurons with Discriminating Taste" by A. Pouget and D. Bavelier, Neuron 2007;53(4):473-475.

Navalpakkam, V., Arbib, M. A., & Itti, L. 2005. Attention and Scene Understanding. *Pages 197–203 of:* Itti, L., Rees, G., & Tsotsos, J. K. (eds), *Neurobiology of Attention.* San Diego, CA: Elsevier.

Noton, D, & Stark, L. 1971. Scanpaths in eye movements during pattern perception. *Science*, **171**(968), 308–11.

Oliva, A., Torralba, A., Castelhano, M.S., & Henderson, J.M. 2003. Top-down control of visual attention in object detection. *Pages I–253 of: Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, vol. 1. IEEE.

Orabona, F., Metta, G., & Sandini, G. 2005. Object-based visual attention: a model for a behaving robot. *Pages 89–89 of: Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*. IEEE.

Ouerhani, N., & Hügli, H. 2003. Real-time visual attention on a massively parallel SIMD architecture. *Real-Time Imaging*, **9**(3), 189–196.

Paletta, L., Rome, E., & Buxton, H. 2005. Attention architectures for machine vision and mobile robots. *Neurobiology of attention*, 642–648.

Pang, D., Kimura, A., Takeuchi, T., Yamato, J., & Kashino, K. 2008. A stochastic model of selective visual attention with a dynamic Bayesian network. *Pages 1073–1076 of: Multimedia and Expo, 2008 IEEE International Conference on*. IEEE.

Parkhurst, D., Law, K., & Niebur, E. 2002. Modeling the role of salience in the allocation of overt visual attention. *Vision Res*, **42**(1), 107–123.

Peters, R. J., & Itti, L. 2007 (Jun). Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. *In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Peters, R. J., & Itti, L. 2008. Congruence between model and human attention reveals unique signatures of critical visual events. *Pages 1145–1152 of: Advances in Neural Information Processing Systems, Vol. 20 (NIPS*2007)*. Cambridge, MA: MIT Press.

Peters, R. J., Iyer, A., Itti, L., & Koch, C. 2005. Components of bottom-up gaze allocation in natural images. *Vision Research*, **45**(8), 2397–2416.

Posner, M.I. 1980. Orienting of attention. *Quarterly journal of experimental psychology*, **32**(1), 3–25.

Privitera, C.M., & Stark, L.W. 2000. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **22**(9), 970–982.

Rajashekar, U., van der Linde, I., Bovik, A.C., & Cormack, L.K. 2008. GAFFE: A gaze-attentive fixation finding engine. *Image Processing, IEEE Transactions on*, **17**(4), 564–573.

Ramanathan, S., Katti, H., Sebe, N., Kankanhalli, M., & Chua, T.S. 2010. An eye fixation database for saliency detection in images. *Computer Vision–ECCV 2010*, 30–43.

Ramström, O., & Christensen, H. 2002. Visual attention using game theory. *Pages 462–471 of: Biologically Motivated Computer Vision*. Springer.

Rao, R. P. N., Zelinsky, G. J., Hayhoe, M. M., & Ballard, D. H. 2002. Eye movements in iconic visual search. *Vision research*, **42**(11), 1447–1463.

Rao, R.P.N., Zelinsky, G.J., Hayhoe, M.M., & Ballard, D.H. 1996. Modeling saccadic targeting in visual search. *Advances in neural information processing systems*, 830–836.

Renninger, L.W., Coughlan, J., Verghese, P., & Malik, J. 2005. An information maximization model of eye movements. *Advances in neural information processing systems*, **17**, 1121–1128.

Rensink, R.A. 2000. The dynamic representation of scenes. *Visual Cognition*, **7**(1-3), 17–42.

Reynolds, J H, & Desimone, R. 1999. The role of neural mechanisms of attention in solving the binding problem. *Neuron*, **24**(1), 19–29, 111–25.

Rezazadegan Tavakoli, H., Rahtu, E., & Heikkilä, J. 2011. Fast and efficient saliency detection using sparse sampling and kernel density estimation. *Image Analysis*, 666–675.

Robertson, L.C., *et al.*. 2003. Binding, spatial attention and perceptual awareness. *Nature Reviews Neuroscience*, **4**(2), 93–102.

Rosin, P.L. 2009. A simple method for detecting salient regions. *Pattern Recognition*, **42**(11), 2363–2371.

Rothenstein, A.L., & Tsotsos, J.K. 2008. Attention links sensing to recognition. *Image and Vision Computing*, **26**(1), 114–126.

Rybak, I A, Gusakova, V I, Golovan, A V, Podladchikova, L N, & Shevtsova, N A. 1998. A model of attention-guided visual perception and recognition. *Vision Res*, **38**(15-16), 2387–2400.

Saalmann, Y.B., Pinsk, M.A., Wang, L., Li, X., & Kastner, S. 2012. The Pulvinar Regulates Information Transmission Between Cortical Areas Based on Attention Demands. *Science*, **337**(6095), 753–756.

Saenz, M., Buracas, G. T., Boynton, G. M., *et al.*. 2002. Global effects of feature-based attention in human visual cortex. *Nature neuroscience*, **5**(7), 631–632.

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. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **24**(3), 420–425.

Schill, K, Umkehrer, E, Beinlich, S, Krieger, G, & Zetzsche, C. in press. Scene analysis with saccadic eye movements: top-down and bottom-up modeling. *J Electronic Imaging*.

Schmidt, J., & Zelinsky, G.J. 2009. Search guidance is proportional to the categorical specificity of a target cue. *The Quarterly Journal of Experimental Psychology*, **62**(10), 1904–1914.

Seo, H., & Milanfar, P. 2009. Static and Space-time Visual Saliency Detection by Self-resemblance. *Journal of Vision.*, **9**(12), 1–27.

Serences, J.T., & Saproo, S. 2010. Population response profiles in early visual cortex are biased in favor of more valuable stimuli. *Journal of neurophysiology*, **104**(1), 76–87.

Shic, F., & Scassellati, B. 2007. A behavioral analysis of computational models of visual attention. *International Journal of Computer Vision*, **73**(2), 159–177.

Siagian, C., & Itti, L. 2007. Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **29**(2), 300–312.

Sodhi, M., Reimer, B., Cohen, JL, Vastenburg, E., Kaars, R., & Kirschenbaum, S. 2002. On-road driver eye movement tracking using head-mounted devices. *Pages 61–68 of: Proceedings of the 2002 symposium on Eye tracking research & applications*. ACM.

Sprague, N., & Ballard, D. 2003. Eye movements for reward maximization. *Advances in neural information processing systems*, **16**.

Summerfield, J.J., Rao, A., Garside, N., & Nobre, A.C. 2011. Biasing Perception by Spatial Long-Term Memory. *The Journal of Neuroscience*, **31**(42), 14952–14960.

Sun, X., Yao, H., & Ji, R. 2012. What Are We Looking For: Towards Statistical Modeling of Saccadic Eye Movements and Visual Saliency. *Pages 1552–1559 of: Proc IEEE-CVPR*.

Sun, Y., & Fisher, R. 2003. Object-based visual attention for computer vision. *Artificial Intelligence*, **146**(1), 77–123.

Sun, Y., Fisher, R., Wang, F., & Gomes, H.M. 2008. A computer vision model for visual-object-based attention and eye movements. *Computer Vision and Image Understanding*, **112**(2), 126–142.

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. 2005. Visual correlates of fixation selection: effects of scale and time. *Vision Res*, **45**(5), 643–659.

Tatler, B.W., & Vincent, B.T. 2009. The prominence of behavioural biases in eye guidance. *Visual Cognition*, **17**(6-7), 1029–1054.

Tatler, B.W., Hayhoe, M.M., Land, M.F., & Ballard, D.H. 2011. Eye guidance in natural vision: Reinterpreting salience. *Journal of vision*, **11**(5).

Theeuwes, J. 2010. Top–down and bottom–up control of visual selection. *Acta psychologica*, **135**(2), 77–99.

Toet, A. 2011. Computational versus Psychophysical Bottom-Up Image Saliency: A Comparative Evaluation Study. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **33**(11), 2131–2146.

Torralba, A. 2003. Modeling global scene factors in attention. *JOSA A*, **20**(7), 1407–1418.

Torralba, A., & Efros, A.A. 2011. Unbiased look at dataset bias. *In: IEEE Conference on Computer Vision and Pattern Recognition*.

Treisman, A, & Souther, J. 1985. Search asymmetry: a diagnostic for preattentive processing of separable features. *J Exp Psychol Gen*, **114**(3), 285–310.

Treisman, A M, & Gelade, G. 1980a. A feature-integration theory of attention. *Cognit Psychol*, **12**(1), 97–136.

Treisman, A.M., & Gelade, G. 1980b. A Feature Integration Theory of Attention. *Cognitive Psychology.*, **12**, 97–136.

Treue, S, & Martinez Trujillo, J C. 1999. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, **399**(6736), 575–579.

Tseng, P.H., Carmi, R., Cameron, I.G.M., Munoz, D.P., & Itti, L. 2009. Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision*, **9**(7).

Tsotsos, J.K., & Rothenstein, A. 2011. Computational models of visual attention. *Scholarpedia*, **6**(1), 6201.

Verma, M., & McOwan, P.W. 2009. Generating customised experimental stimuli for visual search using genetic algorithms shows evidence for a continuum of search efficiency. *Vision research*, **49**(3), 374–382.

Viola, P., & Jones, M. 2001. Rapid object detection using a boosted cascade of simple features. *Pages I–511 of: Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE.

Võ, M.L.H., & Henderson, J.M. 2009. Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of Vision*, **9**(3).

Walther, D., & Koch, C. 2006. Modeling attention to salient proto-objects. *Neural Networks*, **19**(9), 1395–1407.

Wang, H.X., Freeman, J., Merriam, E.P., Hasson, U., & Heeger, D.J. 2012. Temporal eye movement strategies during naturalistic viewing. *Journal of Vision*, **12**(1).

Wang, W., Chen, C., Wang, Y., Jiang, T., Fang, F., & Yao, Y. 2011. Simulating human saccadic scanpaths on natural images. *Pages 441–448 of: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE.

Wolfe, J.M. 1994. Guided search 2.0 A revised model of visual search. *Psychonomic bulletin & review*, **1**(2), 202–238.

Xu, T., Kuhnlenz, K., & Buss, M. 2010. Autonomous behavior-based switched top-down and bottom-up visual attention for mobile robots. *Robotics, IEEE Transactions on*, **26**(5), 947–954.

Yarbus, A. 1967. *Eye Movements and Vision*. New York: Plenum Press.

Yu, Y., Mann, G.K.I., & Gosine, R.G. 2008. An object-based visual attention model for robots. *Pages 943–948 of: Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*. IEEE.

Yu, Y., Mann, G., & Gosine, R. 2012. A Goal-Directed Visual Perception System Using Object-Based Top-Down Attention. *Autonomous Mental Development, IEEE Transactions on*, **4**(1), 87–103.

Zelinsky, G., Zhang, W., Yu, B., Chen, X., & Samaras, D. 2006. The role of top-down and bottom-up processes in guiding eye movements during visual search. *Advances in neural information processing systems*, **18**, 1569.

Zhai, Y., & Shah, M. 2006. Visual attention detection in video sequences using spatiotemporal cues. *Pages 815–824 of: Proceedings of the 14th annual ACM international conference on Multimedia*. ACM.

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).

Zhang, L., Tong, M.H., & Cottrell, G.W. 2009. SUNDAy: Saliency using natural statistics for dynamic analysis of scenes. *Pages 2944–2949 of: Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Zhao, Q., & Koch, C. 2011. Learning a saliency map using fixated locations in natural scenes. *Journal of Vision*, **11**(3).

Zhou, H., & Desimone, R. 2011. Feature-based attention in the frontal eye field and area V4 during visual search. *Neuron*, **70**(6), 1205–1217.

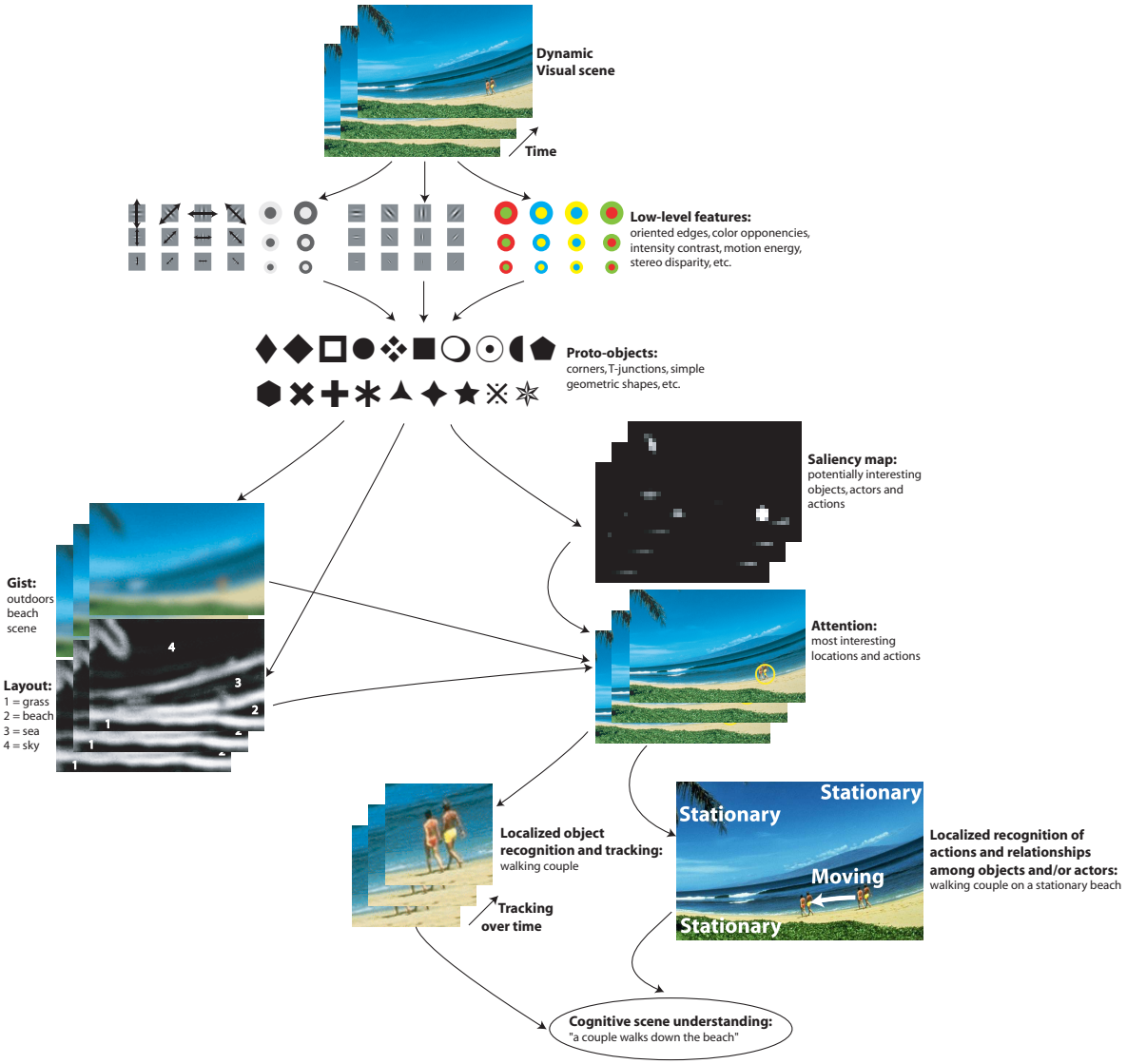

Figure 1: Minimal attention-based architecture for complex dynamic visual scene understanding. This diagram augments the triadic architecture of Rensink (2000), which identified three key components of visual processing: volatile (instantaneous) and parallel pre-attentive processing over the entire visual field, from the lowest-level features up to slightly more complex proto-object representations (top), identification of the setting (scene gist and layout; left), and attentional vision including detailed and more persistent object recognition within the spatially circumscribed focus of attention (right). Here we have extended Rensink's architecture to include a saliency map to guide attention bottom-up towards salient image locations, and action recognition in dynamic scenes. Also see Navalpakkam *et al.* (2005).

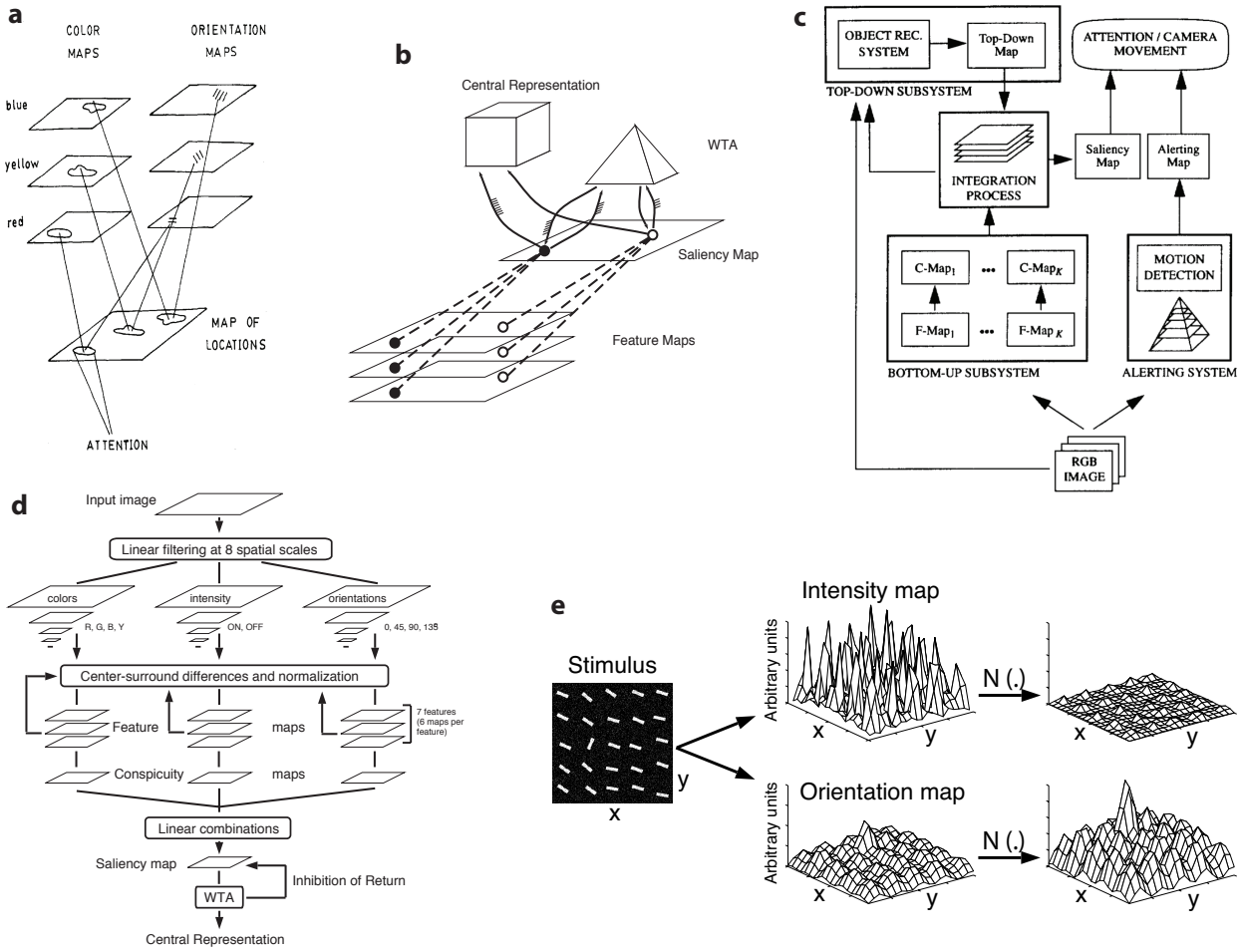

Figure 2: Early bottom-up attention theories and models. **(a)** Feature integration theory of Treisman & Gelade (1980) posits several feature maps, and a focus of attention that scans a map of locations and collects and binds features at the currently attended location. (from Treisman & Souther (1985)). **(b)** Koch & Ullman (1985) introduced the concept of a saliency map receiving bottom-up inputs from all feature maps, where a winner-take-all (WTA) network selects the most salient location for further processing. **(c)** Milanese *et al.* (1994) provided one of the earliest computational models. They included many elements of the Koch & Ullman framework, and added new components, such as an alerting subsystem (motion-based saliency map) and a top-down subsystem (which could modulate the saliency map based on memories of previously recognized objects). **(d)** Itti *et al.* (1998) proposed a complete computational implementation of a purely bottom-up and task-independent model based on Koch & Ullman's theory, including multiscale feature maps, saliency map, winner-take-all, and inhibition of return. **(e)** One of the key elements of Itti *et al.*'s model is to clearly define an attention interest operator, here denoted N(.), whereby the weight by which each feature map contributes to the final saliency map depends on how busy the feature map is. This embodies the idea that feature maps where one location significantly stands out from all others (as is the case in the orientation map shown) should strongly contribute to salience because they clearly vote for a particular location in space as the next focus of attention. In contrast, feature maps where many locations elicit comparable responses (e.g., intensity map shown) should not strongly contribute because they provide no clear indication of which location should be looked at next.

	No	Model	Year	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13
Bottom-up Visual Saliency Models for fixation prediction or salient region detection)	1	Itti et al.	1998	+	-	-	+	-	-	+	f	+	CIO	C	-	-
	2	Privitera & Stark	2000	+	-	-	+	-	-	+	f	+	-	O	-	Stark and Choi
	3	Salah et al.	2002	+	+	-	+	-	-	+	-	+	O	G	DR	Digit & Face
	4	Itti et al.	2003	+	-	+	+	+	+	+	f	+	CIOFM	C	-	-
	5	Torralba	2003	-	+	-	+	-	-	+	s	+	CI	B	DR	Torralba et al.
	6	Sun & Fisher	2003	+	-	-	+	-	-	+	-	-	CIO	G	-	-
	7	Gao & Vasconcelos	2004	-	+	-	+	-	-	+	s	-	DCT	D	DR	Brodatz, Caltech
	8	Ouerhani et al.	2004	+	-	-	+	-	-	+	f	+	CIO+Corner	C	CC	Ouerhani
	9	Boccignone & Ferraro	2004	+	-	+	-	+	-	+	f	-	Optical Flow	B	-	BEHAVE
	10	Frintrop	2005	+	+	+	+	+	+	+	f/s	+/-	CIOM	C	-	-
	11	Itti & Baldi	2005	+	-	+	+	+	+	-	f	+	CIOFM	B	KL, AUC	ORIG-MTV
	12	Ma et al.	2005	+	-	+	+	-	-	+	f	+	M*	O	-	-
	13	Bruce & Tsotsos	2006	+	-	-	+	-	+	+	f	+	DOG, ICA	I	KL, ROC	Bruce and Tsotsos
	14	Navalpakkam & Itti	2006	-	+	-	+	-	+	+	s	+	CIO	C	-	-
	15	Zhai & Shah	2006	+	-	+	+	+	-	+	f	+	SIFT	O	-	-
	16	Harel et al.	2006	+	-	-	+	-	-	+	f	+	IO	G	AUC	Bruce and Tsotsos
	17	Le Meur et al.	2006	+	-	-	+	-	-	+	f	+	LM*	C	CC, KL	Le Meur et al.
	18	Walther & Koch	2006	+	-	-	+	-	+	+	f	+/-	CIO	C	-	-
	19	Peters & Itti	2007	+	+	+	+	+	-	+	i	+	CIOFM	P	KL, NSS	Peters and Itti
	20	Liu et al.	2007	+	-	-	+	-	-	+	f	-	Liu*	G	F-measure	Regional
	21	Shic & Scassellati	2007	+	-	+	+	+	-	+	f	+	CIOM	C	ROC	Shic and Scassellati
	22	Hou & Zhang	2007	+	-	-	+	-	+	+	f	+	FFT, DCT	S	NSS	DB of Hou and Zhang, 2007
	23	Cerf et al.	2007	+	+	-	+	-	+	+	f/s	+	CIO :)	C	AUC	Cerf et al.
	24	Le Meur et al.	2007	+	-	+	+	+	-	+	f	+	LM*	C	CC, KL	Le Meur et al.
	25	Mancas	2007	+	-	+	+	+	+	+	f	+	CI	I	CC	Le Meur et al.
	26	Guo et al.	2008	+	-	-	+	-	-	+	f	+	CIO	D	CC	Self data
	27	Zhang et al.	2008	+	-	-	+	-	+	+	f	+	DOG, ICA	B	KL, AUC	Bruce and Tsotsos
	28	Hou & Zhang	2008	+	-	+	+	+	-	+	f	+	ICA	I	AUC, KL	Bruce and Tsotsos, ORIG
	29	Pang et al.	2008	+	+	+	+	+	-	+	f	+	CIOM	G	NSS	ORIG, Self data
	30	Kootstra et al.	2008	+	-	-	+	-	-	+	f	+	Symmetry	C	CC	Kootstra et al.
	31	Ban et al.	2008	+	-	+	+	+	-	+	f	+	CIO+SYM	I	-	-
	32	Rajashekar et al.	2008	+	-	-	+	-	-	+	f	+	R*	S	CC	Rajashekar et al.
	33	Aziz and Mertsching	2008	+	-	-	+	-	+	+	f	+	CO,size, sym	I	_	-
	34	Kienzle et al.	2009	+	-	-	+	-	-	+	f	+	I	P	K*	Kienzle et al.
	35	Marat et al.	2009	+	-	+	+	+	-	+	f	+	SM*	C	NSS	Marat et al.
	36	Judd et al.	2009	+	-	-	+	-	-	+	f	+	J*	P	AUC	Judd et al.
	37	Seo & Milanfar	2009	+	-	+	+	+	+	+	f	+	LSK	I	AUC, KL	Bruce and Tsotsos, ORIG
	38	Rosin	2009	+	-	-	+	-	-	+	f	+	C+ Edge	O	PR, F-measure	DB of Liu et al, 2007
	39	Yin Li et al.	2009	-	+	+	+	+	+	+	s	+	RGB	S	DR	DB of Hou and Zhang, 2007
	40	Bian & Zhang	2009	+	-	+	+	+	+	+	f	+	FFT	S	AUC	Bruce and Tsotsos
	41	Diaz et al.	2009	+	-	-	+	-	+	+	f	+	CIO	O	AUC	Bruce and Tsotsos
	42	Zhang et al.	2009	+	-	+	-	+	-	+	f	+	DOG, ICA	B	KL, AUC	Bruce and Tsotsos
	43	Achanta et al.	2009	+	-	-	+	-	-	+	f	+	DOG	S	PR	DB of Liu et al, 2007
	44	Gao et al.	2009	+	-	+	+	+	+	+	f	+	CIO	D	AUC	Bruce and Tsotsos
	45	Chikkerur et al.	2010	+	+	-	+	-	+	+	f/s	+/-	CIO	B	AUC	Bruce and Tsotsos, Chikkerur
	46	Mahadaven & Vasconcelos	2010	+	-	+	-	+	-	+	-	+	I	D	DR, AUC	SVCL background data
	47	Avraham & Lindenbaum	2010	+	+	-	+	-	+	+	f/s	+/-	CIO	G	DR, CC	UWGT, Ouerhani et al.
	48	Jia Li et al.	2010	-	+	+	+	+	-	+	f	+	CIO	B	AUC	RSD, MTV, ORIG, Peters and Itti
	49	Guo et al.	2010	+	-	+	+	+	+	+	f/s	+/-	FFT	S	DR	Self data
	50	Borji et al.	2010	-	+	-	+	-	+	+	s	+/-	CIO	O	DR	-
	51	Goeferman et al.	2010	+	-	-	+	-	-	+	-	+	C :)	O	AUC	DB of Hou and Zhang, 2007
	52	Murray et al.	2011	+	-	-	+	-	-	+	f	+	CIO	C	AUC, KL	Bruce and Tsotsos, Judd et al.
	53	Wang et al.	2011	+	-	-	+	-	-	+	f	+	ICA	I	AUC	Self data
Top-down (general models)	54	McCallum	1995	-	+	-	+	-	+	-	i	+	-	R	-	Self data
	55	Rao et al.	1995	-	+	-	+	-	-	+	s	+	CIO	O	-	Self data
	56	Ramstrom & Christiansen	2002	-	+	-	+	-	-	+	-	+	CI	O	-	-
	57	Sprague & Ballard	2003	-	+	+	-	+	+	+	i	-	S*	R	-	-
	58	Renninger et al.	2004	-	+	-	+	-	+	-	s	-	Edgelet	I	DR	Self data
	59	Navalpakkam & Itti	2005	-	+	-	+	-	+	+	-	+	CIO	C	-	Self data
	60	Paletta et al.	2005	-	+	-	+	-	-	+	-	-	SIFT	R	DR	COIL-20, TSG-20
	61	Jodogne & Piater	2007	-	+	-	+	-	-	+	i	-	SIFT	R	-	-
	62	Butko & Movellan	2009	-	+	+	+	+	+	+	s	-	-	R	-	-
	63	Verma & McOwan	2009	+	-	-	+	-	+	-	s	-	CIO	O	-	-
	64	Borji et al.	2010	-	+	-	+	-	-	+	i	-	CIO	R	-	self data
	65	Borji et al.	2012	-	+	-	+	+	-	+	i	-	CIO	B	AUC, NSS	self data

Figure 3: Survey of bottom-up and top-down computational models, classified according to 13 factors. Factors in order are: Bottom-up (f_1), Top-down (f_2), Spatial (-)/Spatio-temporal (+) (f_3), Static (f_4), Dynamic (f_5), Synthetic (f_6) and Natural (f_7) stimuli, Task-type (f_8), Space-based(+)/Object-based(-) (f_9), Features (f_{10}), Model type (f_{11}), Measures (f_{12}), and Used dataset (f_{13}). In Task type (f_8) column: free-viewing (f); target search (s); interactive (i). In Features (f_{10}) column: CIO: color, intensity and orientation saliency; CIOFM: CIO plus flicker and motion saliency; M* = motion saliency, static saliency, camera motion, object (face) and aural saliency (Speech-music); LM* = contrast sensitivity, perceptual decomposition, visual masking and center-surround interactions; Liu* = center-surround histogram, multi-scale contrast and color spatial-distribution; R* = luminance, contrast, luminance-bandpass, contrast-bandpass; SM* = orientation and motion; J* = CIO, horizontal line, face, people detector, gist, etc; S* = color matching, depth and lines; :) = face. In Model type (f_{11}) column, R means that a model is based RL. In Measures (f_{12}) column: K* = used Wilcoxon-Mann-Whitney test (The probability that a random chosen target patch receives higher saliency than a randomly chosen negative one); DR means that models have used a measure of detection/classification rate to determine how successful was a model. PR stands for Precision-Recall. In dataset (f_{13}) column: Self data means that authors gathered their own data. For a detailed definition of these factors please refer to Borji & Itti (2012 PAMI).

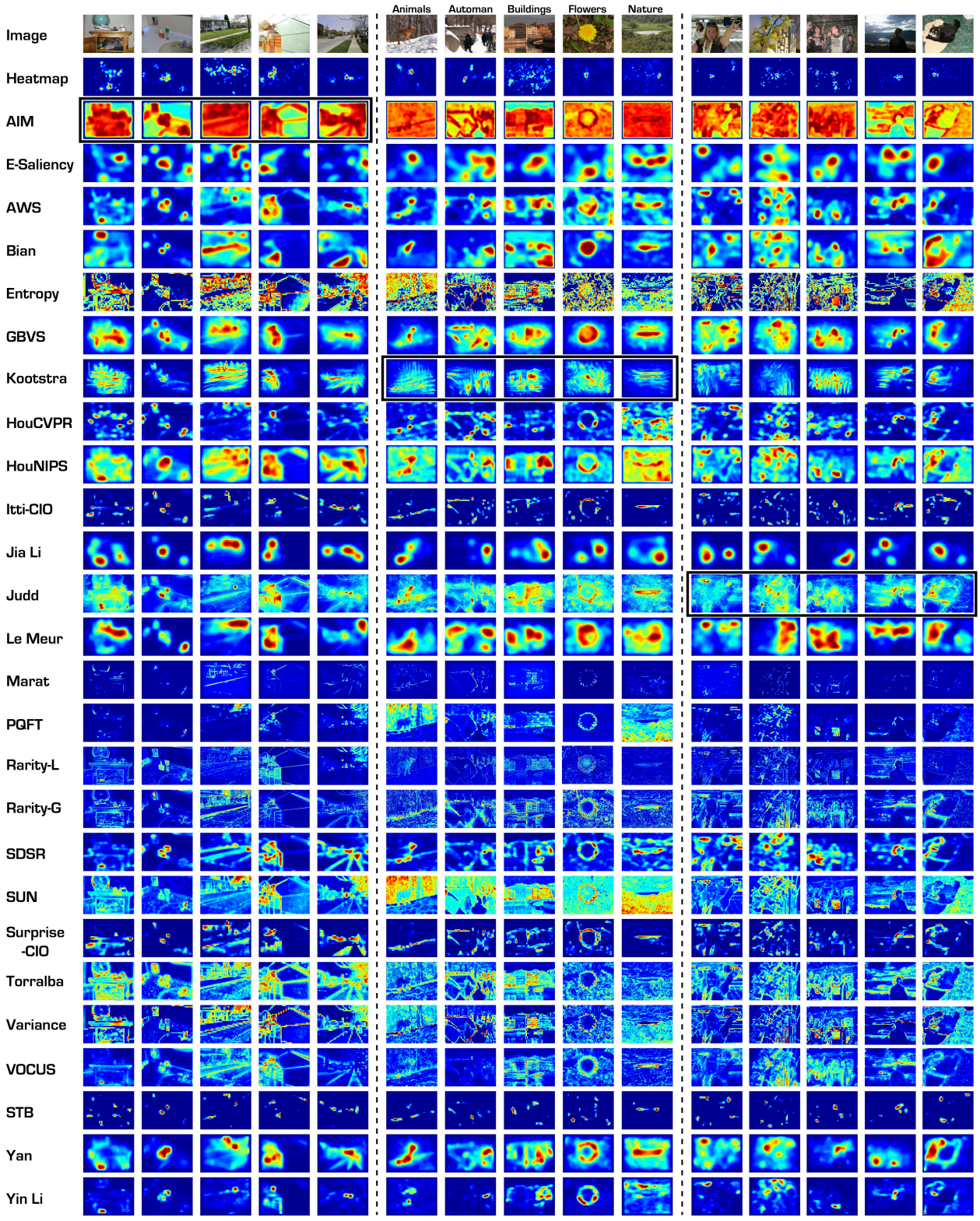

Figure 4: Example images (first row), human eye movement heatmaps (second row), and saliency maps from 26 computational models. The three vertical dashed lines separate the three datasets used (Bruce & Tsotsos, Kootstra & Schomacker, and Judd *et al.*). Black rectangles indicate the model originally associated with a given image dataset. Please see Borji *et al.* (2012 TIP) for additional details.

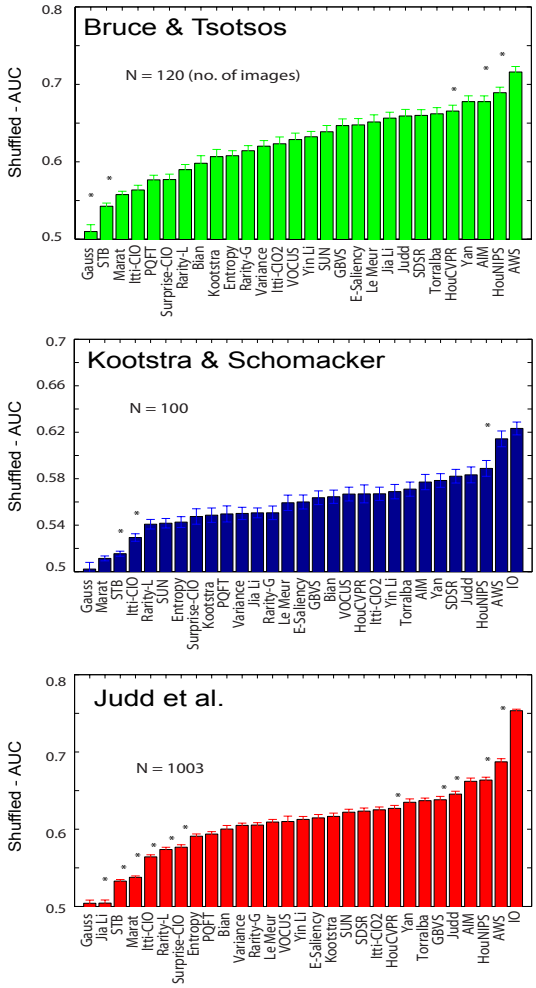

Figure 5: Ranking visual saliency models over three image datasets. Left column: Bruce & Tsotsos (2005), Middle column: Kootstra & Shomaker (2008), and Right column: Judd *et al.* (2009) using shuffled AUC score. Stars indicate statistical significance using t-test (95%, $p \leq 0.05$) between consecutive models. Error bars indicate standard error of the mean (SEM): $\frac{\sigma}{\sqrt{N}}$, where σ is the standard deviation and N is the number of images. The Judd model uses center feature, gist and horizon line, and object detectors for cars, faces, and human body. Itti-CIO2 is the approach proposed by Itti *et al.* (1998) that uses a normalization scheme known as Maxnorm: For each feature map, find the global max M and find the average m of all other local maxima. Then just weight the map by $(M-m)^2$. In the Itti-CIO method (Itti & Koch, 2000), normalization is: Convolve each map by a Difference of Gaussian(DoG) filter, cut off negative values, and iterate this process for a few times. As results show the Maxnorm normalization scheme performs better. In the literature, majority of models have been compared against Itti-CIO model. Please see Borji *et al.* (2012 TIP) for additional details on these results.

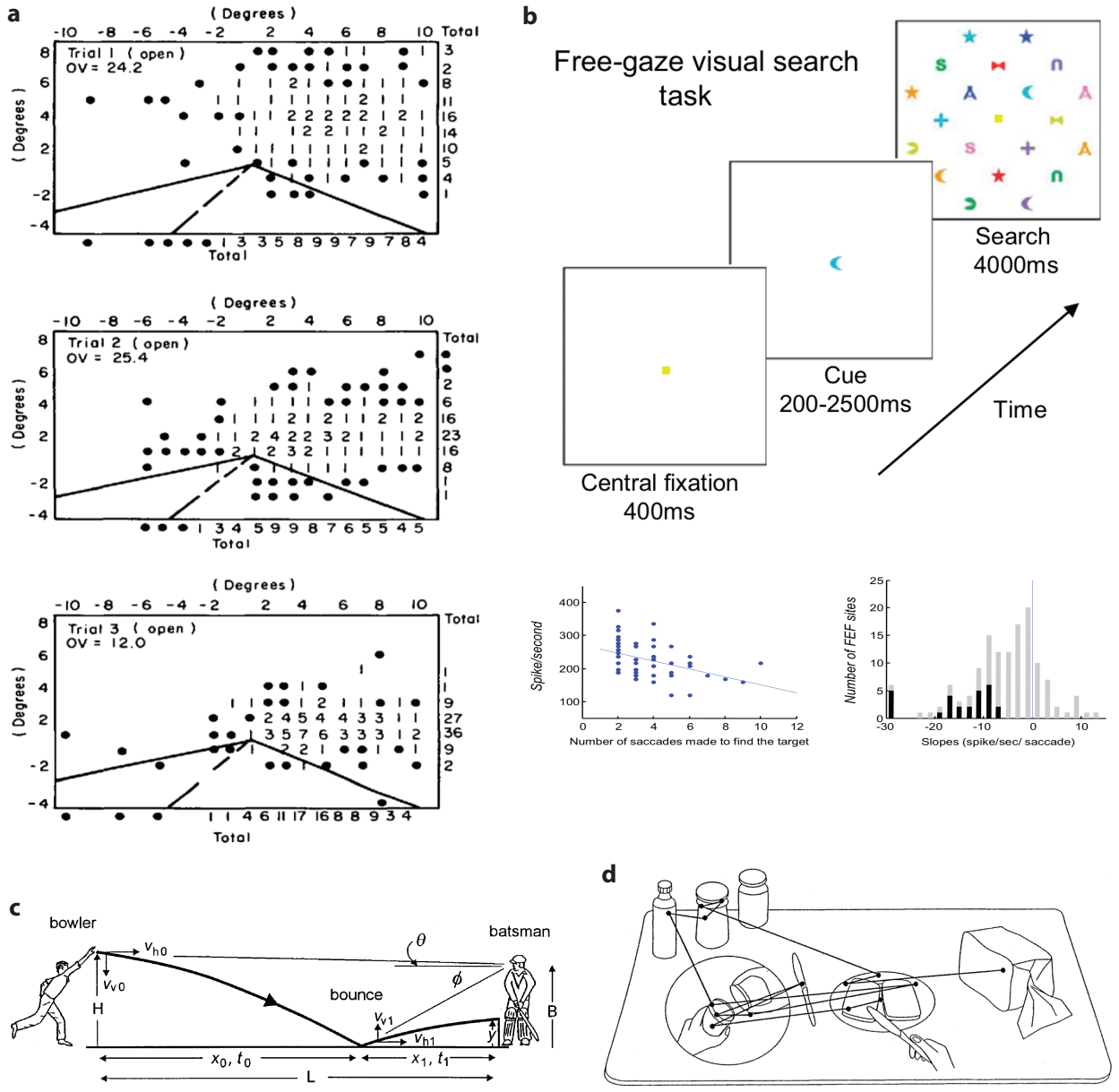

Figure 6: Experimental motivation to explore spatial biases, feature biases, and more complex semantic top-down models. **(a)** Percentage of fixations onto different road locations while human drivers drove at 50 miles/hour on an open road three times (once each panel; dots indicate non-zero percentages smaller than 1). We can see that over the three repetitions (top to bottom panels), eye fixations started clustering more tightly around the left side of the horizon line (from Mourant & Rockwell, 1970). This motivates top-down models to learn over time how a task may induce some spatial biases in the deployment of attention. **(b)** Neural recordings in the frontal eye fields (FEF) as monkeys searched from a cued target among various distractors reveals that the number of saccades the animal made to find the target is negatively correlated with neural firing at the target location around (±50ms) the onset of the first saccade (scatter plot shows examples, one dot per trial, from one recording site, and the negatively-sloped linear regression; histogram shows distributions of slopes over all recording sites, with the significant ones in black and all others in gray). This suggests that top-down models can also exploit biasing for specific features of a search target to attempt to guide attention faster towards the target (from Zhou & Desimone, 2011). **(c)** Eye movement recordings of cricket batsmen revealed that their "eye movements monitor the moment when the ball is released, make a predictive saccade to the place where they expect it to hit the ground, wait for it to bounce, and follow its trajectory for 100-200 ms after the bounce" (Land & McLeod, 2000). This suggests that some knowledge of physics, gravity, bouncing, etc. may be necessary to fully understand human gaze behavior in this more complex scenario. **(d)** Eye movement recordings while making a sandwich are clearly aimed towards the next required item during the unfolding of the successive steps required by the task, with very little searching or exploration (Land & Hayhoe, 2001). Thus, recognition and memorization of objects in the scene is also likely to be required of top-down models to tackle such more complex scenario.

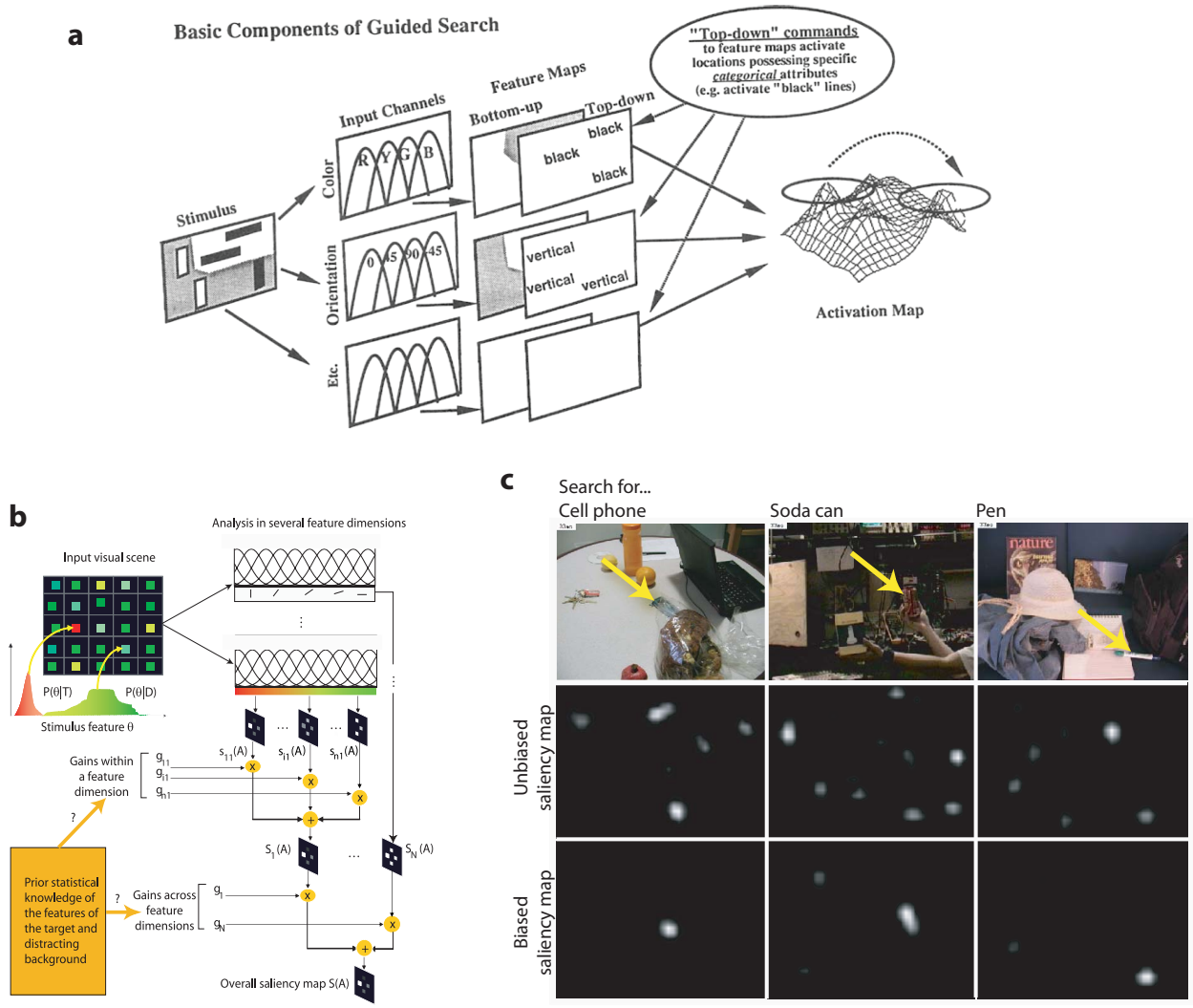

Figure 7: Top-down models that modulate feature gains. **(a)** The Guided Search theory of Wolfe (1994) predicted that bottom-up feature maps can be weighted and modulated by top-down commands. **(b)** Computational framework to optimally compute the weight of each feature map, given top-down knowledge of the expected distribution of feature values θ for both target objects ($P(\theta|T)$) and distractors ($P(\theta|D)$). The gain of each feature is set inversely proportionally to the expected target-to-distractor signal-to-noise ratio given these distributions (Navalpakkam & Itti, 2006; 2007). **(c)** Examples of images (top row), naive un-weighted saliency maps (middle row), and optimally-biased saliency maps based on feature distributions gathered from sample training images (bottom row). Although the object of interest was not the most salient according to a purely bottom-up model in these examples, it becomes the most salient once the model is biased using top-down gains computed from the target and distractor feature distributions (Navalpakkam & Itti, 2006).

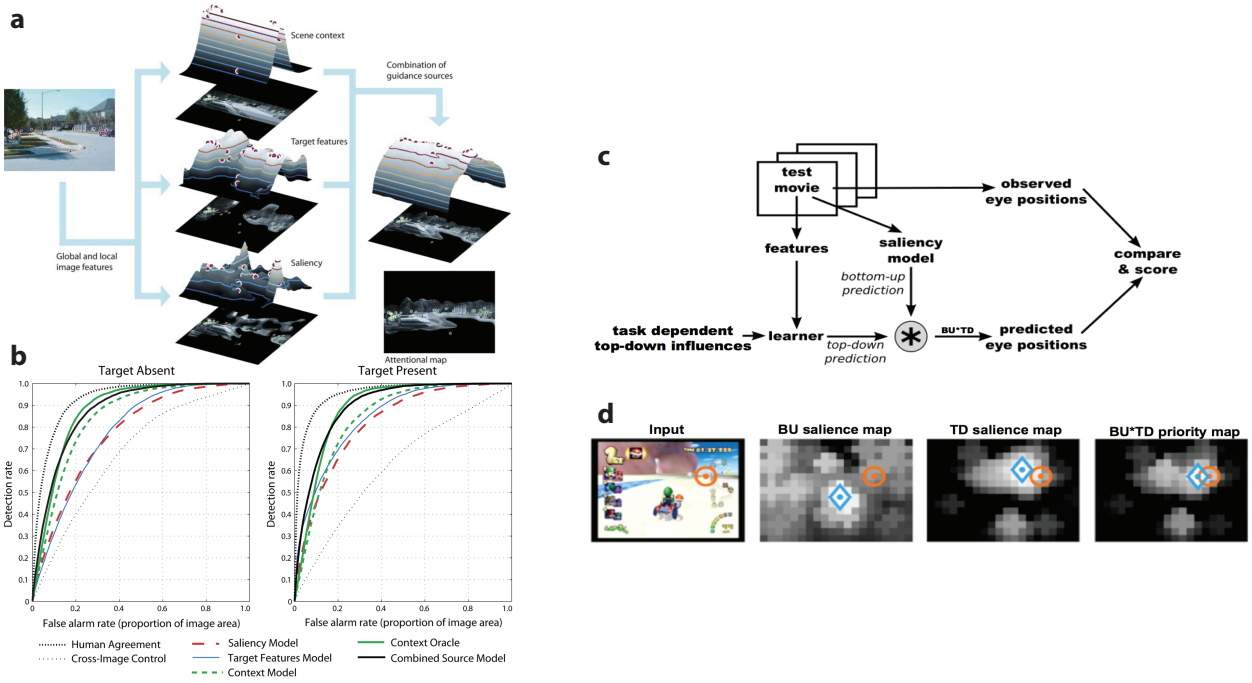

Figure 8: Top-down models that involve spatial modulation. **(a)** Model of Ehinger *at al.* (2009) where an attention map is informed by three guidance sources, given the task of finding people: (1) a spatial map that, based on the coarse scene structure, provides a spatial prior on where humans might appear in the given scene (e.g., they might appear on sidewalks); (2) A map that indicates where visual features in the image that resemble the features of the desired targets are observed; (3) a bottom-up saliency map. **(b)** The combined-source model performs best and significantly better than any of the three component models taken alone, and also performs better than an empirical context oracle (where a set of humans manually indicated where humans might appear in the given scenes). **(c)** A model that learns top-down priors from human gaze behavior while engaged in complex naturalistic tasks, such as driving (Peters & Itti, 2007). A task-dependent learner component builds, during a training phase, associations between distinct coarse types of scenes and observed eye movements (e.g., drivers tend to look to the left when the road turns left). During testing, exposure to similar scenes gives rise to a top-down salience map, which is combined with a standard bottom-up salience map to give rise to the final (BU*TD) priority map that guides attention. **(d)** Example results from the model of (c) applied to a driving video game. Blue diamonds represent the peak location in each map and orange circles represent the current eye position of the human driver. Here the bottom-up (BU) salience map considers that the main character is the most interesting scene element, but, as more correctly predicted by the top-down (TD) map, the driver is looking into the road's turn and on the horizon line. From Peters & Itti (2007).

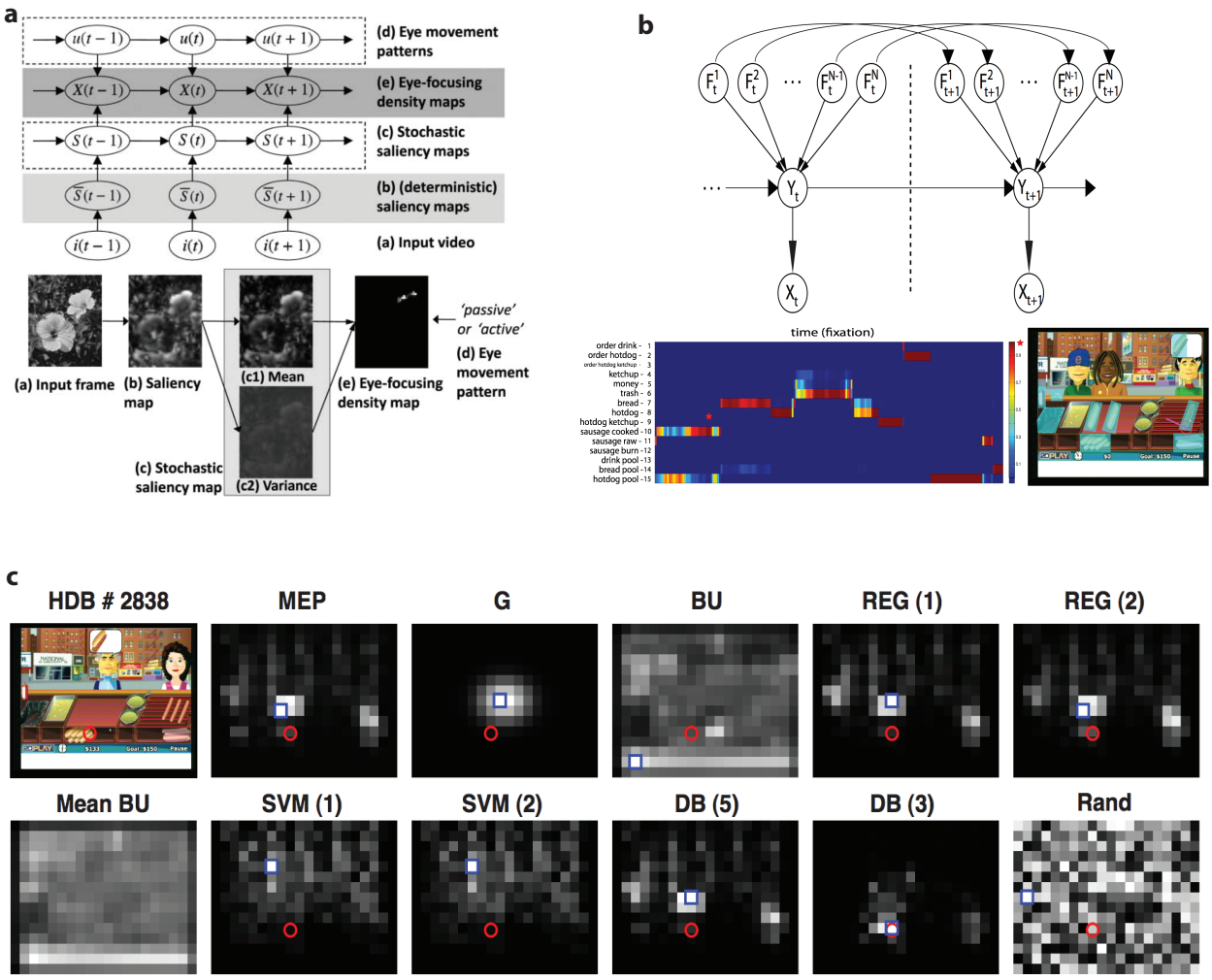

Figure 9: Examples of recent more complex top-down models. **(a)** Model of Akamine *et al.* (2012) which combines bottom-up saliency influences and top-down active/passive state influences over space and time using a dynamic Bayesian network. Although the top-down state is quite simple in this model (active vs. passive), the proposed mathematical framework could easily extend to more complex top-down influences. **(b)** Graphical representation of the DBNs approach of Borji *et al.* (2012 AAAI) unrolled over two time-slices. X_t is the current saccade position, Y_t is the currently attended object, and F_t^i is the function that describe object i at the current scene. All variables are discrete. It also shows a time series plot of probability of objects being attended and a sample frame with tagged objects and eye fixation overlaid. **(c)** Sample predicted saccade maps of the DBN model (shown in b). Each red circle indicates the observers eye position superimposed with each maps peak location (blue squares). Smaller distance indicates better prediction. Images from top-left to bottom-right are: a sample frame from the hot-dog bush game where the player has to serve customers food and drink, MEP stands for the mean eye position over all frames during the game play, G is just a trivial Gaussian map at the image center, BU is the bottom-up saliency map of the Itti model, REG(1) is a regression model which maps the previous attended object to the current attended object and fixation location, REG(2) is similar to REG(1) but the input vector consists of the available objects at the scene augmented with the previously attended object, SVM(1) and SVM(1) correspond to REG(1) and REG(2) but using an SVM classifier, Mean BU is the average BU map showing which regions are salient throughout the game course, Similarly DBN(1) and DBN(2) correspond to REG(1) and REG(2) meaning that in DBN(1) network slice consists of just one node for previously attended object while in DBN(2) each network slice consists of the previously attended object as well information of the previous objects in the scene, and finally Rand is a white noise random map.