State-of-the-art in Visual Attention Modeling

Ali Borji, *Member, IEEE,* and Laurent Itti, *Member, IEEE*

Abstract—Modeling visual attention — particularly stimulus-driven, saliency-based attention — has been a very active research area over the past 25 years. Many different models of attention are now available, which aside from lending theoretical contributions to other fields, have demonstrated successful applications in computer vision, mobile robotics, and cognitive systems. Here we review, from a computational perspective, the basic concepts of attention implemented in these models. We present a taxonomy of nearly 65 models, which provides a critical comparison of approaches, their capabilities, and shortcomings. In particular, thirteen criteria derived from behavioral and computational studies are formulated for qualitative comparison of attention models. Furthermore, we address several challenging issues with models, including biological plausibility of the computations, correlation with eye movement datasets, bottom-up and top-down dissociation, and constructing meaningful performance measures. Finally, we highlight current research trends in attention modeling and provide insights for future.

Index Terms—Visual attention, bottom-up attention, top-down attention, saliency, eye movements, regions of interest, gaze control, scene interpretation, visual search, gist.

1 Introduction

A RICH stream of visual data ($10^8 - 10^9$ bits) enters our eyes every second [1][2]. Processing this data in real-time is an extremely daunting task without the help of clever mechanisms to reduce the amount of erroneous visual data. High-level cognitive and complex processes such as object recognition or scene interpretation rely on data that has been transformed in such a way to be tractable. The mechanism this paper will discuss is referred to as visual attention - and at its core lies an idea of a selection mechanism and a notion of relevance. In humans, attention is facilitated by a retina that has evolved a high-resolution central fovea and a low-resolution periphery. While visual attention guides this anatomical structure to important parts of the scene to gather more detailed information, the main question is on the computational mechanisms underlying this guidance.

In recent decades, many facets of science have been aimed towards answering this question. Psychologists have studied behavioral correlates of visual attention such as change blindness [3][4], inattentional blindness [5], and attentional blink [6]. Neurophysiologists have shown how neurons accommodate themselves to better represent objects of interest [27][28]. Computational neuroscientists have built realistic neural network models to simulate and explain attentional behaviors (e.g., [29][30]). Inspired by these studies, robotists and computer vision scientists have tried to tackle the inherent problem of computational complexity to build systems capable of working in real-time (e.g., [14][15]). Although there are many models available now in the research areas mentioned above, here we limit ourselves to models that can compute saliency maps (please see next section for definitions) from any image or video input. For a review on computational models of visual attention in general, including biased competition [10], selective tuning [15], normalization models of attention [181], and many others, please refer to [8]. Reviews of attention models from psychological, neurobiological, and computational perspectives can be found in [9][77][10][12][202][204][224]. Fig. 1 shows a taxonomy of attentional studies and highlights our scope in this review.

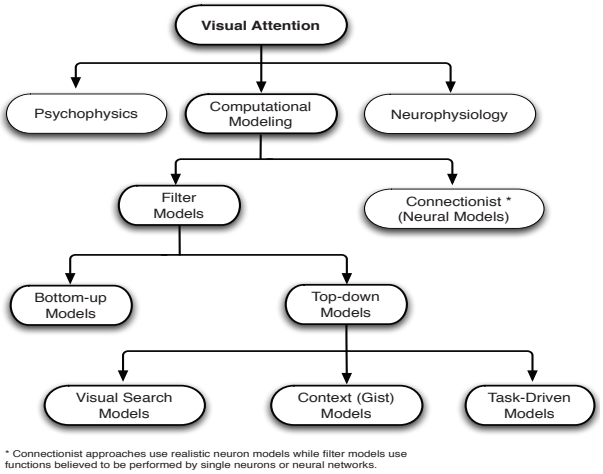

Fig. 1. Taxonomy of visual attention studies. Ellipses with solid borders illustrate our scope in this paper.

1.1 Definitions

While the terms attention, saliency, and gaze are often used interchangeably, each has a more subtle definition that allows their delineation.

Attention is a general concept covering all factors that influence selection mechanisms, whether they be scene-driven bottom-up (BU) or expectation-driven top-down (TD).

Saliency intuitively characterizes some parts of a scene — which could be objects or regions — that appear to an observer to stand out relative to their neighboring parts. The term "salient" is often considered in the context of bottom-up computations [18][14].

Gaze, a coordinated motion of the eyes and head, has often been used as a proxy for attention in natural behavior

- *Authors are with the Department of Computer Science, University of Southern California (USC), Los Angeles, CA, 90089. E-mail: {borji,itti}@usc.edu*

- *Manuscript received November 2010.*

Digital Object Indentifier 10.1109/TPAMI.2012.89

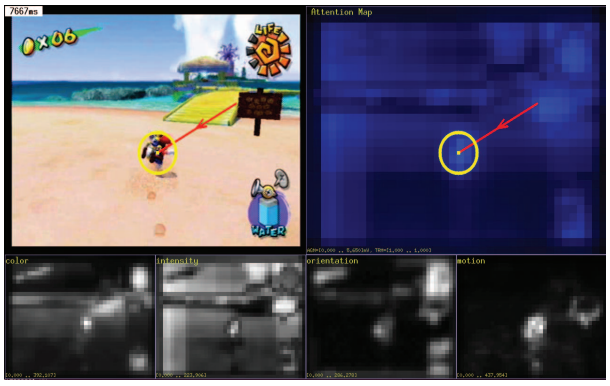

Fig. 2. Neuromorphic Vision C++ Toolkit (iNVT) developed at iLab, USC, http://ilab.usc.edu/toolkit/. A saccade is targeted to the location that is different from its surroundings in several features. In this frame from a video, attention is strongly driven by motion saliency.

(see [99]). For instance, a human or a robot has to interact with surrounding objects and control the gaze to perform a task while moving in the environment. In this sense, gaze control engages vision, action, and attention simultaneously to perform sensorimotor coordination necessary for the required behavior (e.g., reaching and grasping).

1.2 Origins

The basis of many attention models dates back to Treisman & Gelade's [81] "Feature Integration Theory" where they stated which visual features are important and how they are combined to direct human attention over pop-out and conjunction search tasks. Koch and Ullman [18] then proposed a feed-forward model to combine these features and introduced the concept of a saliency map which is a topographic map that represents conspicuousness of scene locations. They also introduced a winner-take-all neural network that selects the most salient location and employs an inhibition of return mechanism to allow the focus of attention to shift to the next most salient location. Several systems were then created implementing related models which could process digital images [15][16][17]. The first complete implementation and verification of the Koch & Ullman model was proposed by Itti *et al.* [14] (see Fig. 2) and was applied to synthetic as well as natural scenes. Since then, there has been increasing interest in the field. Various approaches with different assumptions for attention modeling have been proposed and have been evaluated against different datasets. In the following sections, we present a unified conceptual framework in which we describe the advantages and disadvantages of each model against one another. We give the reader insight into the current state of the art in attention modeling and identify open problems and issues still facing researchers.

The main concerns in modeling attention are how, when, and why we select behaviorally-relevant image regions. Due to these factors, several definitions and computational perspectives are available. A general approach is to take inspiration from the anatomy and functionality of the early human visual system, which is highly evolved to solve these problems (e.g., [14][15][16][191]). Alternatively, some studies have hypothesized what function visual attention may serve and have formulated it in a computational framework. For instance, it has been claimed that visual attention is attracted to the most informative [144], the most surprising scene regions [145], or those regions that maximize reward regarding a task [109].

1.3 Empirical Foundations

Attentional models have commonly been validated against eye movements of human observers. Eye movements convey important information regarding cognitive processes such as reading, visual search, and scene perception. As such, they often are treated as a proxy for shifts of attention. For instance, in scene perception and visual search, when the stimulus is more cluttered, fixations become longer and saccades become shorter [19]. The difficulty of the task (e.g., reading for comprehension versus reading for gist, or searching for a person in a scene versus looking at the scene for a memory test) obviously influences eye movement behavior [19]. Although both attention and eye movement prediction models are often validated against eye data, there are slight differences in scope, approaches, stimuli, and level of detail. Models for eye movement prediction (saccade programming) try to understand mathematical and theoretical underpinnings of attention. Some examples include search processes (e.g., optimal search theory [20]), information maximization models [21], Mr. Chips: an ideal-observer model of reading [25], EMMA (Eye Movements and Movement of Attention) model [139], HMM model for controlling eye movements [26], and constrained random walk model [175]). To that end, they usually use simple controlled stimuli, while on the other hand, attention models utilize a combination of heuristics, cognitive and neural evidence, and tools from machine learning and computer vision to explain eye movements in both simple and complex scenes. Attention models are also often concerned with practical applicability. Reviewing all movement prediction models is beyond the scope of this paper. The interested reader is referred to [22][23][127] for eye movement studies and [24] for a breadth-first survey of eye tracking applications.

Note that eye movements do not always tell the whole story and there are other metrics which can be used for model evaluation. For example, accuracy in correctly reporting a change in an image (i.e., search-blindness [5]), or predicting what attention grabbing items one will remember, show important aspects of attention which are missed by sole analysis of eye movements. Many attention models in visual search have also been tested by accurately estimating reaction times (RT) (e.g., RT/setsize slopes in pop-out and conjunction search tasks [224][191]).

1.4 Applications

In this paper, we focus on describing the attention models themselves. There are, however, many technological applications of these models which have been developed over the years and which have further increased interest in attention modeling. We organize the applications of attention modeling into three categories: vision and graphics, robotics, and those in other areas as shown in Fig. 3.

1.5 Statement and Organization

Attention is difficult to define formally in a way that is universally agreed upon. However, from a computational

Category	Application	References
Computer Vision and Graphics	Image segmentation	Mishra and Aloimonos, 2009, Maki et al., 2000
	Image quality assessment	Ma and Zhang, 2008, Ninassi et al., 2007
	Image matching	Walther et al., 2006, Siagian and Itti, 2009, Frintrop and Jensfelt, 2008
	Image rendering	DeCarlo and Santella, 2002
	Image and video compression	Ouerhani et al., 2003, Itti, 2004, Guo and Zhang, 2010.
	Image thumbnailing	Marchesotti et al., 2009, Le Meur et al., 2006, Suh et al., 2003
	Image super-resolution	Jacobson et al, 2010
	Image re-targeting (thumbnailing)	Setlur et al., 2005, Chamaret et al., 2008, Goferman et al., 2010, Achanta et al., 2009, Marchesotti et al., 2009, Le Meur et al., 2006, Suh et al., 2003
	Image superresolution	Sadaka and Karam, 2009
	Video summarization	Marat et al., 2007, Ma et al., 2005
	Scene classification	Siagian and Itti, 2009
	Object detection	Frintrop, 2006, Navalpakkam and Itti, 2006, Fritz et al., 2005, Butko and Movellan, 2009, Viola and Jones, 2004, Ehinger et al., 2009.
	Salient object detection	Liu et al, 2007, Goferman et al. 2010, Achanta et al., 2009, Rosin, 2009.
	Object recognition	Salah et al., 2002, Walther et al., 2006 and 2007, Frintrop, 2006, Mitri et al., 2005, Gao and Vasconcelos, 2004 and 2009, Han and Vasconcelos 2010, Paletta et al., 2005.
	Visual tracking	Mahadevan and Vasconcelos, 2009, Frintrop, 2010
	Dynamic lighting	Seif El-Nasr, 2009
	Video shot detection	Boccignone et al., 2005
	Interest point detection	Kadir and Brady, 2001, Kienzle et al., 2007.
	Automatic collage creation	Goferman et al., 2010, Wang et al., 2006.
	Face segmentation and tracking	Li and Ngan, 2008
Robotics	Active vision	Mertsching et al., 1999, Vijaykumar et al., 2001, Dankers, 2007, Borji et al., 2010
	Robot Localization	Siagian and Itti, 2009, Ouerhani et al., 2005
	Robot Navigation	Baluja and Pomerlau, 1997, Scheier and Egner, 1997
	Human-robot interaction	Breazeal, 1999, Heidemann et al., 2004, Belardinelli, 2008, Nagai, 2009, Muhl, 2007
	Synthetic vision for simulated actors	Courty and Marchand, 2003
Others	Advertising	Rosenholtz et al. 2011, Liu et al., 2008
	Finding tumors in mammograms	Hong and Brady, 2003
	Retinal prostheses	Parick et al., 2010

Fig. 3. Some applications of visual attention modeling.

standpoint, many models of visual attention (at least those tested against first few seconds of eye movements in free-viewing) can be unified under the following general problem statement. Assume K subjects have viewed a set of N images $\mathcal{I} = \{\boldsymbol{I}_i\}_{i=1}^{N}$. Let $\boldsymbol{L}_i^k = \{\boldsymbol{p}_{ij}^k, t_{ij}^k\}_{j=1}^{n_i^k}$ be the vector of eye fixations (saccades) $\boldsymbol{p}_{ij}^k = (x_{ij}^k, y_{ij}^k)$ and their corresponding occurrence time t_{ij}^k for the k-th subject over image $\boldsymbol{I}_i$. Let the number of fixations of this subject over i-th image be n_i^k. The goal of attention modeling is to find a function (stimuli-saliency mapping) $f \in \mathcal{F}$ which minimizes the error on eye fixation prediction, i.e, $\sum_{k=1}^{K}\sum_{i=1}^{N} m(f(\boldsymbol{I}_i^k), \boldsymbol{L}_i^k)$, where $m \in \mathcal{M}$ is a distance measure (defined in section 2.7). An important point here is that the above definition better suits bottom-up models of overt visual attention, and may not necessarily cover some other aspects of visual attention (e.g., covert attention or top-down factors) that cannot be explained by eye movements.

Here we present a systematic review of major attention models that we apply to arbitrary images. In section 2, we first introduce several factors to categorize these models. In section 3, we then summarize and classify attention models according to these factors. Limitations and issues in attention modeling are then discussed in section 4 and are followed by conclusions in section 5.

2 Categorization Factors

We start by introducing 13 factors ($\mathtt{f}_{1..13}$) that will be used later for categorization of attention models. These factors have their roots in behavioral and computational studies of attention. Some factors describe models ($\mathtt{f}_{1,2,3}$, $\mathtt{f}_{8..11}$), others ($\mathtt{f}_{4..7}$, $\mathtt{f}_{12,13}$) are not directly related, but are just as important as they determine the scope of applicability of different models.

2.1 Bottom-up vs. Top-down Models

A major distinction among models is whether they rely on bottom-up influences ($\mathtt{f}_1$), top-down influences ($\mathtt{f}_2$), or a combination of both.

Bottom-up cues are mainly based on characteristics of a visual scene (stimulus-driven)[75], whereas top-down cues (goal-driven) are determined by cognitive phenomena like knowledge, expectations, reward, and current goals.

Regions of interest that attract our attention in a bottom-up manner must be sufficiently distinctive with respect to surrounding features. This attentional mechanism is also called exogenous, automatic, reflexive, or peripherally cued [78]. Bottom-up attention is fast, involuntary, and most likely feed-forward. A prototypical example of bottom-up attention is looking at a scene with only one horizontal bar among several vertical bars where attention is immediately drawn to the horizontal bar [81]. While many models fall in this category, they can only explain a small fraction of eye movements since the majority of fixations are driven by task [177].

On the other hand, top-down attention is slow, task-driven, voluntary, and closed-loop [77]. One of the most famous examples of top-down attention guidance is from Yarbus in 1967 [79], who showed that eye movements depend on the current task with the following experiment: subjects were asked to watch the same scene (a room with a family and an unexpected visitor entering the room) under different conditions (questions) such as "estimate the material circumstances of the family", "what are the ages of the people?", or simply to freely examine the scene. Eye movements differed considerably for each of these cases.

Models have explored three major sources of top-down influences in response to this question: How do we decide where to look?. Some models address visual search in which attention is drawn toward features of a target object we are looking for. Some other models investigate the role of scene context or gist to constrain locations that we look at. In some cases, it is hard to precisely say where or what we are looking at since a complex task governs eye fixations, for example in driving. While in principle, task demands on attention subsumes the other two factors, in practice models have been focusing on each of them separately. Scene layout has also been proposed as a source of top-down attention [80][93] and is here considered together with scene context.

1) **Object Features**. There is a considerable amount of evidence for target-driven attentional guidance in real-world search tasks [84][85][23][83]. In classical search tasks, target features are a ubiquitous source of attention guidance [81][82][83]. Consider a search over simple search arrays in which the target is a red item: attention is rapidly directed toward the red item in the scene. Compare this with a more complex target object, such as a pedestrian in a natural scene, where although it is difficult to define the target, there are still some features (e.g., upright form, round head, and straight body) to direct visual attention [87].

The guided search theory [82] proposes that attention can be biased toward targets of interest by modulating the relative gains through which different features contribute to attention. To return to our prior example, when looking for a red object, a higher gain would be assigned to red color. Navalpakkam *et al.* [51] derived the optimal integration of cues (channels of the BU saliency model [14]) for detection of a target in terms of maximizing the signal-to-noise ratio of

the target versus background. In [50], a weighting function based on a measure of object uniqueness was applied to each map before summing up the maps for locating an object. Butko *et al.* [161] modeled object search based on the same principles of visual search as stated by Najemnik *et al.* [20] in a partially observable framework for face detection and tracking, but they did not apply it to explain eye fixations while searching for a face. Borji *et al.* [89] used evolutionary algorithms to search in a space of basic saliency model parameters for finding the target. Elazary and Itti [90] proposed a model where top-down attention can tune both the preferred feature (e.g., a particular hue) and the tuning width of feature detectors, giving rise to more flexible top-down modulation compared to simply adjusting the gains of fixed feature detectors. Last but not least are studies such as [147][215][141] that derive a measure of saliency from formulating search for a target object.

The aforementioned studies on the role of object features in visual search are closely related to object detection methods in computer vision. Some object detection approaches (e.g., Deformable Part Model by Felzenszwalb *et al.* [206] and the Attentional Cascade of Viola and Jones [220]) have high detection accuracy for several objects such as cars, persons, and faces. In contrast to cognitive models, such approaches are often purely computational. Research on how these two areas are related will likely yield mutual benefits for both.

2) **Scene Context**. Following a brief presentation of an image ($\sim$80 ms or less), an observer is able to report essential characteristics of a scene [176][71]. This very rough representation of a scene, so called "gist", does not contain many details about individual objects but can provide sufficient information for coarse scene discrimination (e.g., indoor vs. outdoor). It is important to note that gist does not necessarily reveal the semantic category of a scene. Chun and Jiang [91] have shown that targets appearing in repeated configurations relative to some background (distractor) objects were detected more quickly [71]. Semantic associations among objects in a scene (e.g., a computer is often placed on top of a desk) or contextual cues have also been shown to play a significant role in the guidance of eye movements [199][84].

Several models for gist utilizing different types of low-level features have been presented. Oliva and Torralba [93], computed the magnitude spectrum of a Windowed Fourier Transform over non-overlapping windows in an image. They then applied principal component analysis (PCA) and independent component analysis (ICA) to reduce feature dimensions. Renninger and Malik [94] applied Gabor filters to an input image and then extracted 100 universal textons selected from a training set using K-means clustering. Their gist vector was a histogram of these universal textons. Siagian and Itti [95] used biological center-surround features from orientation, color, and intensity channels for modeling gist. Torralba [92] used wavelet decomposition tuned to 6 orientations and 4 scales. To extract gist, a vector is computed by averaging each filter output over a 4×4 grid. Similarly he applied PCA to the resultant 384D vectors to derive a 80D gist vector. For a comparison of gist models, please refer to [96][95].

Gist representations have become increasingly popular in computer vision since they provide rich global yet discriminative information useful for many applications such as search in the large-scale scene datasets available today [116], limiting the search to locations likely to contain an object of interest [92][87], scene completion [205], and modeling top-down attention [101][218]). It can thus be seen that research in this area has the potential to be very promising.

3) **Task Demands**. Task has a strong influence on deployment of attention [79]. It has been claimed that visual scenes are interpreted in a need-based manner to serve task demands [97]. Hayhoe *et al.* [99] showed that there is a strong relationship between visual cognition and eye movements when dealing with complex tasks. Subjects performing a visually-guided task were found to direct a majority of fixations toward task-relevant locations [99]. It is often possible to infer the algorithm a subject has in mind from the pattern of her eye movements. For example, in a "block-copying" task where subjects had to replicate an assemblage of elementary building blocks, the observers' algorithm for completing the task was revealed by patterns of eye-movements. Subjects first selected a target block in the model to verify the block's position, then fixated the workspace to place the new block in the corresponding location [216]. Other research has studied high-level accounts of gaze behavior in natural environments for tasks such as sandwich making, driving, playing cricket, and walking (see Henderson and Hollingworth [177], Rensink [178], Land and Hayhoe [135], and Bailensen and Yee [179]). Sodhi *et al.* [180] studied how distractors while driving such as adjusting the radio or answering a phone affect eye movements.

The prevailing view is that bottom-up and top-down attention are combined to direct our attentional behavior. An integration method should be able to explain when and how to attend to a top-down visual item or skip it for the sake of a bottom-up salient cue. Recently, [13] proposed a Bayesian approach that explains the optimal integration of reward as a top-down attentional cue, and contrast or orientation as a bottom-up cue in humans. Navalpakkam and Itti [80] proposed a cognitive model for task-driven attention constrained by the assumption that the algorithm for solving the task was already available. Peters and Itti [101] learned a top-down mapping from scene gist to eye fixations in video game playing. Integration was simply formulated as multiplication of BU and TD components.

2.2 Spatial vs. Spatio-temporal Models

In the real-world, we are faced with visual information that constantly changes due to egocentric movements or dynamics of the world. Visual selection is then dependent on both current scene saliency as well as the accumulated knowledge from previous time points. Therefore, an attention model should be able to capture scene regions that are important in a spatio-temporal manner.

To be detailed in section 3, almost all attention models include a spatial component. We can distinguish between two types of modeling temporal information in saliency modeling: 1) Some bottom-up models use the *motion* channel to capture human fixations drawn to moving stimuli [119]. More recently, several researchers have started modeling temporal effects on bottom-up saliency (e.g., [143][104][105]). 2) On the other hand, some models [109][218][26][25][102] aim to capture the spatio-temporal aspects of a task for example by learning sequences of attended objects or actions as the task progresses.

For instance, the Attention Gate Model (AGM) [183], emphasizes the temporal response properties of attention and quantitatively describes the order and timing for humans attending to sequential target stimuli. Previous information about images, eye fixations, image content at fixations, physical actions, as well as other sensory stimuli (e.g., auditory) can be exploited to predict the next eye movement. Adding a temporal dimension and the realism of natural interactive tasks brings a number of complications in predicting gaze targets within a computational model.

Suitable environments for modeling temporal aspects of visual attention are dynamic and interactive setups such as movies and games. Boiman and Irani [122] proposed an approach for irregularity detection from videos by comparing texture patches of actions with a learned dataset of irregular actions. Temporal information was limited to the stimulus level and did not include higher cognitive functions such as the sequence of items processed at the focus of attention or actions performed while playing the games. Some methods derive static and dynamic saliency maps and propose methods to fuse them (e.g., Jia Li *et al.* [133] and Marat *et al.* [49]). In [103], a spatio-temporal attention modeling approach for videos is presented by combining motion contrast derived from the homography between two images and spatial contrast calculated from color histograms. Virtual reality (VR) environments have also been used in [99][109][97]. Some other models dealing with the temporal dimension are [105][108][103]. We postpone the explanation of these approaches to section 3.

Factors f_3 indicates whether a model uses spatial only or spatio-temporal information for saliency estimation.

2.3 Overt vs. Covert attention

Attention can be differentiated based on its attribute as "overt" versus "covert". Overt attention is the process of directing the fovea towards a stimulus while covert attention is mentally focusing onto one of several possible sensory stimuli. An example of covert attention is staring at a person who is talking but being aware of visual space outside the central foveal vision. Another example is driving, where a driver keeps his eyes on the road while simultaneously covertly monitoring the status of signs and lights. The current belief is that covert attention is a mechanism for quickly scanning the field of view for an interesting location. This covert shift is linked to eye movement circuitry that sets up a saccade to that location (overt attention) [203]. However, this does not completely explain complex interactions between covert and overt attention. For instance, it is possible to attend to the right hand corner field of view and actively suppress eye movements to that location. Most of these models detect regions that attract eye fixations and few explain overt orientation of eyes along with head movements. Lack of computational frameworks for covert attention might be because behavioral mechanisms and functions of covert attention are still unknown. Further, it is not known yet how to measure covert attention.

Because of a great deal of overlap between overt and covert attention and since they are not exclusive concepts, saliency models could be considered as modeling both overt and covert mechanisms. However, in depth discussion of this topic goes beyond our scope and merits of this paper and demands special treatment elsewhere.

2.4 Space-based vs. Object-based Models

There is no unique agreement on the unit of attentional scale: Do we attend to spatial locations, to features, or to objects? The majority of psychophysical and neurobiological studies are about space-based attention (e.g., Posner's spatial cueing paradigm [98][111]). There is also strong evidence for feature-based attention (detecting an odd item in one feature dimension [81] or tuning curve adjustments of feature selective neurons [7]) and object-based attention (selectivity attending to one of two objects, e.g., face vs. vase illusion [112][113][84]). The current belief is that these theories are not mutually exclusive and visual attention can be deployed to each of these candidate units, implying there is no single unit of attention. Humans are capable of attending to multiple (between four and five) regions of interest simultaneously [114][115].

In the context of modeling, a majority of models are space-based (see Fig. 7). It is also viable to think that humans work and reason with objects (compared with rough pixel values) as main building blocks of top-down attentional effects [84]. Some object-based attentional models have previously been proposed, but they lack explanation for eye fixations (e.g., Sun and Fisher [117], Borji *et al.* [88]). This shortcoming makes verification of their plausibility difficult. For example, the limitation of the Sun and Fisher [117] model is the use of human segmentation of the images; it employs information that may not be available in the pre-attentive stage (before the objects in the image are recognized). Availability of object-tagged image and video datasets (e.g., LabelMe Image and Video [116][188]) has made conducting effective research in this direction possible. The link between object-based and space-based models remains to be addressed in the future. Feature-based models (e.g., [51][83]) adjust properties of some feature detectors in an attempt to make a target object more salient in a distracting background. Because of the close relationship between visual features and objects, in this paper we categorize feature-based models under object-based models as shown in Fig. 7.

The ninth factor f_9, indicates whether a model is space-based or object-based - meaning that it needs to work with objects instead of raw spatial locations.

2.5 Features

Traditionally, according to feature integration theory (FIT) and behavioral studies [81][82][118], three features have been used in computational models of attention: intensity (or intensity contrast, or luminance contrast), color, and orientation. Intensity is usually implemented as the average of three color channels (e.g., [14][117]) and processed by center-surround processes inspired by neural responses in lateral geniculate nucleus (LGN) [10] and V1 cortex. Color is implemented as red-green and blue-yellow channels inspired by color-opponent neurons in V1 cortex, or alternatively by using other color spaces such as HSV [50] or Lab [160]. Orientation is often implemented as a convolution with oriented Gabor filters or by the application of oriented masks. Motion was first used in [119] and was implemented by applying directional masks to the image (in the primate brain motion is derived by the neurons at MT and MST regions which are selective to direction of motion). Some studies have also added specific features for

directing attention like skin hue [120], face [167], horizontal line [93], wavelet [133], gist [92][93], center-bias [123], curvature [124], spatial resolution [125], optical flow [15][126], flicker [119], multiple superimposed orientations (crosses or corners) [127], entropy [129], ellipses [128], symmetry [136], texture contrast [131], above average saliency [131], depth [130], and local center-surround contrast [189]. While most models have used the features proposed by FIT [81], some approaches have incorporated other features like Difference of Gaussians (DOG) [144][141] and features derived from natural scenes by ICA and PCA algorithms [92][142]. For target search, some have employed the structural description of objects such as the histogram of local orientations [87][199]. For detailed information regarding important features in visual search and direction of attention, please refer to [118][81][82]. Factor $\mathtt{f}_{10}$, categorizes models based on features they use.

2.6 Stimuli and Task Type

Visual stimulus can be first distinguished as being either static (e.g., search arrays, still photographs; factor $\mathtt{f}_4$) or dynamic (e.g., videos, games; factor $\mathtt{f}_5$). Video games are interactive and highly dynamic since they do not generate the same stimuli each run and have nearly natural renderings, though they still lag behind the statistics of natural scenes and do not have the same noise distribution. The setups here are more complex, more controversial, and more computationally intensive. They also engage a large number of cognitive behaviors.

The second distinction is between synthetic stimuli (Gabor patches, search arrays, cartoons, virtual environments, games; factor $\mathtt{f}_6$) and natural stimuli (or approximations thereof, including photographs and videos of natural scenes; factor $\mathtt{f}_7$). Since humans live in a dynamic world, video and interactive environments provide a more faithful representation of the task facing the visual system than static images. Another interesting domain for studying attentional behavior, agents in virtual reality setups, can be seen in the work of Sprague and Ballard [109], who employed a realistic human agent in VR and used reinforcement learning (RL) to coordinate action selection and visual perception in a sidewalk navigation task involving avoiding obstacles, staying on the sidewalk, and collecting litter.

Factor $\mathtt{f}_8$ distinguishes among task types. The three most widely explored tasks to date in the context of attention modeling are: (1) Free viewing tasks, in which subjects are supposed to freely watch the stimuli (there is no task or question here, but many internal cognitive tasks are usually engaged), (2) Visual search tasks where subjects are asked to find an odd item or a specific object in a natural scene, and (3) Interactive tasks. In many real-world situations, tasks such as driving and playing soccer engage subjects tremendously. These complex tasks involve many subtasks such as visual search, object tracking, and focused and divided attention.

2.7 Evaluation Measures

So we have a model that outputs a saliency map S, and we have to quantitatively evaluate it by comparing it with eye movement data (or click positions) G. How do you compare these? We can think of them as probability distributions, and use Kullback-Leibler (KL) or Percentile metrics to measure distance between distributions. Or we can consider S as a binary classifier and use signal detection theory analysis (Area Under the ROC Curve (AUC) metric) to asses the performance of this classifier. We can also think of S and G as random variables and use Correlation Coefficient (CC) or Normalized Scanpath Saliency (NSS) to measure their statistical relationship. Another way is to think of G as a sequence of eye fixations (scanpath) and compare this sequence with the sequence of fixations chosen by a saliency model (string-edit distance).

While in principle any model might be evaluated using any measure, in Fig. 7 we list in factor $\mathtt{f}_{12}$ the measures which were used by the authors of each model. In the rest, when we use Estimated Saliency Maps (ESM S), we mean a saliency map of a model, and by Ground-truth Saliency Map (GSM G), we mean a map that is built by combining recorded eye fixations from all subjects or combining tagged salient regions by human subjects for each image.

From another perspective, evaluation measures for attention modeling can be classified into three categories: 1) point-based, 2) region-based, and 3) subjective evaluation. In point-based measures, salient points from ESMs are compared to GSMs made by combining eye fixations. Region-based measures are useful for evaluating attention models over regional saliency datasets by comparing the ESMs and labeled salient regions (GSM annotated by human subjects) [133]. In [103], subjective scores on estimated saliency maps were reported on three levels: "Good", "Acceptable", and "Failed". The problem with such subjective evaluation is that it is difficult to extend it to large-scale datasets.

In the following, we focus on explaining those metrics with more consensus from the literature and provide pointers for others (Percentile [134] and Fixation Saliency Method (FS) [131][182]) for reference.

Kullback-Leibler (KL) Divergence. The KL divergence is usually used to measure distance between two probability distributions. In the context of saliency, it is used to measure the distance between distributions of saliency values at human vs. random eye positions [145][77]. Let $t_i = 1 \cdots N$ be N human saccades in the experimental session. For a saliency model, ESM is sampled (or averaged in a small vicinity) at the human saccade $x_{i,human}$ and at a random point $x_{i,random}$. The saliency magnitude at the sampled locations is then normalized to the range [0,1]. The histogram of these values in q bins covering the range [0,1] across all saccades is then calculated. H_k and R_k are the fraction of points in bin k for salient and random points. Finally the difference between these histograms with the (symmetric) KL divergence (A.k.a relative entropy) is:

$$KL = \frac{1}{2}\sum_{k=1}^{q}\left(H_k log\frac{H_k}{R_k} + R_k log\frac{R_k}{H_k}\right) \quad (1)$$

Models that can better predict human fixations exhibit higher KL divergence, since observers typically gaze towards a minority of regions with the highest model responses while avoiding the majority of regions with low model responses. Advantages of KL divergence over other scoring schemes [212][131] are: 1) Other measures essentially calculate the rightward shift of H_k histogram relative to the R_k histogram, whereas KL is sensitive to any difference between the histograms, and 2) KL is invariant to reparameterizations, such that applying any continuous monotonic nonlinearity (e.g., $S^3, \sqrt{S}, e^S$) to ESM values

S does not affect scoring. One disadvantage of the KL divergence is that it does not have a well-defined upper bound — as the two histograms become completely non-overlapping, the KL divergence approaches infinity.

Normalized Scanpath Saliency (NSS). The normalized scanpath saliency [134][131] is defined as the response value at the human eye position, (x_h, y_h), in a model's ESM that has been normalized to have zero mean and unit standard deviation $NSS = \frac{1}{\sigma_s}(S(x_h, y_h) - \mu_S)$. Similar to the percentile measure, NSS is computed once for each saccade, and subsequently the mean and standard error are computed across the set of NSS scores. $NSS = 1$ indicates that the subjects' eye positions fall in a region whose predicted density is one standard deviation above average. Meanwhile $NSS \leq 0$ indicates that the model performs no better than picking a random position on the map. Unlike KL and percentile, NSS is not invariant to reparameterizations. Please see [134] for an illustration of NSS calculation.

Area Under Curve (AUC). AUC is the area under Receiver Operating Characteristic (ROC) [195] curve. As the most popular measure in the community, ROC is used for the evaluation of a binary classifier system with a variable threshold (usually used to classify between two methods like saliency vs. random). Using this measure, the model's ESM is treated as a binary classifier on every pixel in the image; pixels with larger saliency values than a threshold are classified as fixated while the rest of the pixels are classified as non-fixated [144][167]. Human fixations are then used as ground truth. By varying the threshold, the ROC curve is drawn as the *false positive rate* vs. *true positive rate*, and the area under this curve indicates how well the saliency map predicts actual human eye fixations. Perfect prediction corresponds to a score of 1. This measure has the desired characteristic of transformation invariance, in that area under the ROC curve does not change when applying any monotonically increasing function to the saliency measure. Please see [192] for an illustration of ROC calculation.

Linear Correlation Coefficient (CC). This measure is widely used to compare the relationship between two images for applications such as image registration, object recognition, and disparity measurement [196][197]. The linear correlation coefficient measures the strength of a linear relationship between two variables:

$$CC(G,S) = \frac{\sum_{x,y}(G(x,y) - \mu_G).(S(x,y) - \mu_S)}{\sqrt{\sigma_G^2.\sigma_S^2}} \quad (2)$$

where G and S represent the GSM (fixation map, a map with 1's at fixation locations, usually convolved with a Gaussian) and the ESM, respectively. μ and σ^2 are the mean and the variance of the values in these maps. An interesting advantage of CC is the capacity to compare two variables by providing a single scalar value between -1 and +1. When the correlation is close to $+1/-1$ there is almost a perfectly linear relationship between the two variables.

String Editing Distance. To compare the regions of interest selected by a saliency model (mROI) to human regions of interest (hROI) using this measure, saliency maps and human eye movements are first clustered to some regions. Then ROIs are ordered by the value assigned by the saliency algorithm or temporal ordering of human fixations in the scanpath. The results are strings of ordered points such as: $string_h =$ "$abcfeffgdc$" and $string_s =$ "$afbffdcdf$". The string editing similarity index S_s is then defined by an optimization algorithm with unit cost assigned to the three different operations: *deletion*, *insertion*, and *substitution*. Finally the sequential similarity between the two strings is defined as: $similarity = 1 - \frac{S_s}{|string_s|}$. For our example strings, above similarity is $1 - 6/9 = 0.34$ (see [198][127] for more information on string editing distance). Please see [127] for an illustration of this score.

2.8 Datasets

There are several eye movement datasets of still images (for studying static attention) and videos (for studying dynamic attention). In Fig. 7 we list as factor $\mathtt{f}_{13}$ some available datasets. Here we only mention those datasets that are mainly used for evaluation and comparison of attention models, though there are many other works that have gathered special-purpose data (e.g., for driving, sandwich making, and block copying [135]).

Figs. 4 and 5 show summaries of image and video eye movements datasets (For a few, labeled salient regions are available). Researchers have also used mouse tracking to estimate attention. Although this type of data is noisier, some early results show a reasonably good ground-truth approximation. For instance, Scheier and Egner [61] showed that mouse movement patterns are close to eye-tracking patterns. A web-based mouse tracking application was set up at the TCTS laboratory [110]. Other potentially useful datasets (which are not eye-movement datasets) are tagged-object datasets like PASCAL and Video LabelMe. Some attentional works have used this type of data (e.g., [116]).

3 Attention Models

In this section, models are explained based on their mechanism to obtain saliency. Some models fall into more than one category. In the rest of this review, we focus only on those models which have been implemented in software and can process arbitrary digital images and return corresponding saliency maps. Models are introduced in chronological order. Note that here we are more interested in models of saliency instead of those approaches that detect and segment the most salient region or object in a scene. While these models use a saliency operator at the initial stage, their main goal is not to explain attentional behavior. However, some methods have further inspired subsequent saliency models. Here, we reserve the term "saliency detection" to refer to such approaches.

3.1 Cognitive Models (C)

Almost all attentional models are directly or indirectly inspired by cognitive concepts. The ones that have more bindings to psychological or neurophysiological findings are described in this section.

Itti et al.'s basic model [14] uses three feature channels color, intensity, and orientation. This model has been the basis of later models and the standard benchmark for comparison. It has been shown to correlate with human eye movements in free-viewing tasks [131][184]. An input image is subsampled into a Gaussian pyramid and each pyramid level σ is decomposed into channels for Red (R), Green (G), Blue (B), Yellow (Y), Intensity (I), and local orientations

Study	Subjects	Dataset Size	Resolution	Viewing distance (cm)	Presentation time (s)	Description
Kienzle et al. [165]	14	200	1024 x 768	60	3	8-bit grayscale stimuli presented on a 19-inch Iiyama CRT at full screen size corresponding to 37° × 27° of visual angle.
Einhauser et al. [84]	7	54	640 x 480	50	-	Overall 32,225 fixations with average fixation duration as 370±293 ms and 11.9 fixations per image. The average distance of subsequent fixation points on the screen is 127 pixels(19°). Authors restricted their analysis to 76° ×55° regions which accounts for 92% (29,725) of all fixations. Stimuli was presented using NEC LT 157 projector at resolution 1024 × 768 at 60Hz on average spanned 133 x 100cm, corresponding to 37° × 27° of visual angle.
Ouerhani et al. [210]	6	-	640 x 480	70	5	Age range (24-34), with normal or corrected-to-normal acuity as well as normal color vision. Stimulus presented on a 19" monitor subtending 29° × 22°. Task was "just look at the image". Eyetracker: EyeLink, SenseoMotoric Instruments GmbH. Recording at 250Hz, accuracy 0.5° – 1° accuracy with a 3×3 point grid calibration sequence.
Bruce and Tsotsos [144]	20	120	681 x 511	75	4	Images (indoor and outdoor) were presented at random with 2 s gray mask in between on a 21-inch CRT monitor. The eye tracking apparatus consisted of an ERICA workstation including Hitachi CCD camera with an IR emitting LED. Stimuli were color images and task was free viewing. **Link:** www-sop.inria.fr/members/Neil.Bruce
Stark and Choi [211]	7	15	-	40	4	Bright Purkinje reflection captured by a video camera. Stimulus size was 15 × 20cm yielding to 21° × 29° with the 0.5-1 degree accuracy. Images were terrain photographs, landscapes and paintings. Task was free viewing.
Chikkerur et al. [154]	8	220	640 x 480	70	5	Scenes contained cars (4.6 ± 3.8) and pedestrians (2.1 ± 2.2)- visual angle: 16 × 12. Subjects were asked to count the number of cars or pedestrians. Using an ETL 400 ISCAN, table-mounted video-based eye tracker at 240 HZ and accuracy of 0.5°. (age:18-35). Images were 100 from x and 120 from LabelMe. **Link:** http://www.sharat.org/
Torralba et al. [92]	24	36	15.8 x 11.9	-	-	In people search task, 14 stimuli out of 36 contained no people and 22 included 1-6 people. The same set (36 indoor) images was used for painting search (17 images without any paintings and rest with 1-6 paintings) and for mug search (half without and half with 1-6 mugs). Eyetracking was performed by a Generation 5.5 SRI Dual Purkinje Image Eyetracker, sampling at 1000Hz. Color photos displayed on a NEC Multisync P750 monitor (143Hz refresh). Mean target size was 1.05%(1.24%) of the image size for people, 7.3%(7.6%) for painting and 0.5% (0.4%) for mugs. **Link:** http://people.csail.mit.edu/torralba/GlobalFeaturesAndAttention/
Judd et al. [166]	15	1003	Various	48	3	Images were collected from Flicker creative commons and LabelMe datasets. The longest dimension was 1024 with other ranging from 405 to 1024. There were 779 landscape images and 228 portrait images. Images were freely viewed with 1 sec gray screen between each two. Camera was recalibrated after every 50 images. First fixation was discarded. Age range: 18-35. **Link:** http://people.csail.mit.edu/tjudd/WherePeopleLook/index.html
Cerf et al. [167]	7	250	1024 x 768	80	-	Eye position of subjects were acquired at 1000Hz using an Eyelink 1000 (SR Research, Osgoode, Canada). The task had three phases: 1) free viewing, 2) searching for face, an object, banana, cell phone, toy car, etc shown by a probe image, and 3) 100 image recognition memory task where subjects had to answer with y/n whether they had seen the image before. Stimuli subtended 28° × 21° of visual angle. **Link:** http://www.fifadb.com/
Peters et al. [134]	12	100/ class	-	75	-	ISCAN Inc eye tracker was used to sample eye movements at 120Hz. Age range: 18-25; four did free-viewing over (outdoor photos, overhead satellite imagery, and fractals). Another 4 did free-viewing over involving Gabor snakes and Gabor arrays. Seven subjects did a contour detection task. Resolution was 1000 × 1000 to 1536 × 1024 subtending a visual angle of 15.8° × 15.8° to 16.2° × 25°. **Link:** http://ilab.usc.edu
Reinagel and Zador [212]	5	77	640 x 480	79	10	Images were 69 nature scenes, 38 man-made objects such as buildings, 17 animals or humans and 8 synthetics. An RK-416 infrared Pupil Tracking System and a 21-inch monitor was used. The whole image subtended 28° × 21° of visual angle. Subjects were instructed to "Study the images". Estimated tracking error was 0.5°. **Link:** http://zadorlab.cshl.edu/
Hwang and Pomplun [87]	30	160	1280 x 1024	-	10	Age range: 19-40. Stimuli were 160 photographs (1280 × 1024) real-world scenes including landscapes, home interiors, and city scenes and covered 20° × 20° of visual angle. An SR research EyeLink II system. Stimuli presented on 19-inch Del P992 monitor (85Hz refresh rate), the whole image subtended 28° × 21°. **Link:** http://www.cs.umb.edu/~marc/
Kootstra et al., [136]	31	99	1024 x 768	70	-	Eyelink head-mounted eye tracking (SR research) was used and was recalibrated before each session. Age range: 17-32. Task was free viewing. Stimuli were: 12 Animals, 12 Automan, 16 Buildings, 20 flowers, 41 natural scenes and were shown on a 18-inch CRT monitor (36 × 27 cm). **Link:** http://www.csc.kth.se/~kootstra/
Tatler [123]	14	48	800 x 600	60	-	Eyelink-I eye tracker was used. Subjetcs had normal or corrected to normal vision with age range 17-32. Image subtended 30° × 22° and were presented on a 17-inch SVGA color monitor (74 Hz refresh). Task was free viewing. **Link:** http://www.activevisionlab.org/
Engmann et al., [182]	8	90	1280 x 1024	85	-	Subjetcs had normal or corrected-to-normal vision and normal color vision with age range 20-27 (avg: 22.3). Stimuli were presented on a 19.7" EIZO FlexScan F77S CRT monitor (100 Hz refresh). Natural scenes selected from the Zurich natural image database (Einhauser et al. [99]) which only rarely contain isolated nameable objects or man-made artifacts at resolution 2048 × 1536. Image subtended 26° × 18° 17-inch SVGA color monitor. Task was free viewing. Eye tracker was Eyelink-2000 (SR Research Ltd. Canada) with 13 point calibration.
Engelke et al. [213]	30	7	512 x 512	60	8	Images were 4 human faces ("Barbara"), 1 "Glodhill" face (gurilla) and 1 "Peppers" images. Eye tracker was EyeTech TM3 and task was free viewing. Each image was presented for 8 sec with a gray screen with central fixation in between.
Le Meur et al. [41]	40	46	800 x 600	*	-	Stimuli were 46 degraded versions of 10 color images using spatial filtering. Task was free viewing. Eye tracker was made by Cambridge Research Corporation. Viewing distance was four times the TV monitor height. **Link:** http://www.irisa.fr/temics/staff/lemeur
Ehinger et al. [87]	14	912	800 x 600	75	15	Stimuli were color images (half with a pedestrian) with resolution 800 × 600 and were shown on a 21-inch CRT monitor with resolution 1024 × 768 and refresh rate 100Hz. A 240 Hz ISCAN RK-464 video-based eye tracker was used for recording. The task was to decide wether a pedestrain is in the scene or not. **Link:** http://cvcl.mit.edu/searchmodels/
Rajashekar et al. [174]	29	101	1042 x 768	134	-	Subjects were 18 males, 11 females with mean age of 27. Eye tracker was made by Image Systems Corp, MN. Grayscale images were shown on a 21-inch grayscale gamma corrected monitor with resolution 1024 × 768. The task was free viewing. **Link:** http://live.ece.utexas.edu/research/doves/

Fig. 4. Some benchmark eye movement datasets over still images often used to evaluate visual attention models.

(O_θ). From these channels, center-surround "feature maps" f_l for different features l are constructed and normalized. In each channel, maps are summed across scale and normalized again:

$$\begin{aligned} f_l &= \mathcal{N}\left(\sum_{c=2}^{4}\sum_{s=c+3}^{c+4} f_{l,c,s}\right), \forall l \in L_I \cup L_C \cup L_O \\ L_I &= \{I\}, L_C = \{RG, BY\}, L_O = \{0^\circ, 45^\circ, 90^\circ, 135^\circ\} \end{aligned} \quad (3)$$

These maps are linearly summed and normalized once more to yield the "conspicuity maps":

$$C_I = f_I, C_C = \mathcal{N}(\sum_{l\in L_C} f_l), C_O = \mathcal{N}(\sum_{l\in L_O} f_l) \quad (4)$$

Finally, conspicuity maps are linearly combined once more to generate the saliency map: $S = \frac{1}{3}\sum_{k\in\{I,C,O\}} C_k$.

There are at least four implementations of this model: iNVT by Itti [14], Saliency Toolbox (STB) by Walther [35], VOCUS by Frintrop [50], and a Matlab code by Harel [121]. In [119], this model was extended by adding motion and flicker contrasts to video domain. Zhaoping Li [170], introduced a neural implementation for saliency map in V1 area that can also account for search difficulty in pop-out and conjunction search tasks.

Le Meur et al. [41] proposed an approach for bottom-up saliency based on the structure of the human visual system (HVS). Contrast sensitivity functions, perceptual decompo-

Dataset	Features	Feature Value
CRCNS - ORIG [145]	C	50 clips (0:06-1:30 min each), ~25 min total, ~6GB for 46K frames
	S	8 (3 female, 5 male) subjects with normal corrected vision, Ages 23-32, From mixed ethnicities
	T	"Follow main actors and actions, try to understand overall what happens in each clip."
	ST	Complex video stimuli involving TV programs, outdoor sceces, video games Outdoor day & night, parks, crowds, rooftop bar. etc.
	D	ISCAN RK-464 eye tracker, 240 HZ recording, 9 point calibration after every 5 clips, 640× 480 resolution at 60.27HZ doublescan, 33.185ms/movie frame, (x,y) of each saccade
	L	http://crcns.org/data-sets/eye/eye-1
CRCNS - MTV [145]	C	50 video clips (4-7 subjects on each video clip)
	S	8 subjects different from subjects of CRCNS
	D	This dataset was created by cutting video clips of CRCNS into 1-3s "clipets" and reassembling those clippets in random order. Other aspects were the same as the original dataset.
	L	http://crcns.org/data-sets/eye/eye-1
Jia Li et al. [133]	C	431 videos with total length of 7.5 hours, 764,806 frames in total with 62,356 key frames
	S	23 (17 male and 4 female) subjects with age range between 21-37
	ST	6 genres: documentary, ad, cartoon, news, movie and surveillance
	D	10-23 subjects per each clip were assigned to manually label the salient regions with one or multiple rectangles from key frames. Drawback with this dataset is rectangular labeling but this may be resolved with segmentation, inefficiney to evaluate whatever
	L	http://www.jdl.ac.cn/user/jiali/
Peters and Itti [101]	C	24 game-play sessions, ~185 GB for 216K frames, 8,449 saccades of ampitude 2o or more
	S	5(3 male, 2 female) subjetcs with normal corrected vision
	T	"Play 4 or 5 five-minute segments of the Nintendo GameCube games"
	ST	Games include Mario Kart, Wave Race, Super Mario Sunshine, Hulk and Pac Man World.
	D	Subjects were seated viewing distance of 80 cm (28° × 21° usable field of view) Stimuli were presented on a 22" computer monitor (LaCie Corp; 640 × 480, 75 HZ refresh, mean screen luminance 30cd/m2, room luminance 4 cd/m2) ISCAN RK-464 eye tracker, 240 HZ recording, 9 point calibration after before game segment Frames were grabbed using a dual-CPU Linux computer with SCHED FIFO schedulingto ensure micorsecond accurate timing.
	L	http://ilab.usc.edu/rjpeters/
Shic and Scassellati [74]	C	2 clips, 10, young adults, normal and mildly mentally retarded
	T	"One minute long clips from back and white movie "Who's afraid of Virgnia Woolf"
	D	A head mounted eye-tracker (ISCAN Inc.) was used. The eye tracker employs dark pupil- corneal reflection video-occulography and had accuracy within ±0.3o over a horizontal and range of ±20o, with a sampling rate of 60 Hz. The subjects sat 63.5 cm from the 48.3 cm screen on which the movie was shown at a resolution of 640 × 480 pixels.
	L	http://sites.google.com/site/fredshic/home
Marat et al. [49]	C	53 short video clips (25 fps, 720 × 576 pixels), 1700 frames
	S	15 (3f,12m) subjects with age range 23-40 and had normal or corrected to normal vision
	ST	Each clip ~ 1-3sec long, 324 clip snippets. There was not a particular task or question. TV shows, TV news, animated movies, commercials, sport and music. Indoor, out-door, day-time, night-time)
	D	The clip snippets were strung to form 20 clips of 30 seconds (30.20 ± 0.61). Eye positions were recorded at 500 Hz (20 eye positions per frame for two eyes) using a Eyelink II (SR Research). Participants were positioned with their chin supported on a 21" color monitor (75 HZ) at a viewing distance of 57cm (40° × 30° usable field of view). A calibration was carried out at every five stimuli and a control drfit was done before each stimuli.
	L	http://start1g.ovh.net/~qgsmabaq/sophie/index.php
Le Meur et al. [138]	C	7 clips (25 Hz, 352 × 288 pixels), 2451 frames, Each clip ~ 4.5-33.8 sec long
	S	17-27 subject for different clips with normal or corrected to normal vision
	T	Free viewing
	ST	Faces, sporting events, audiencesm, landscape, logos, incrustations, low and high satiotemporal
	D	Dual-Purkinje eye tracker from Cambridge Research Corporation. Sampling frequency was 50Hz. CRT display 800 × 600 pixels, 25° × 27°. Distance to screen was 81 cm.
	L	http://www.irisa.fr/temics/staff/lemeur

C: Clips; **S**: Subjects; **T**: Task; **ST**: Stimuli Type; **D:** Description; **L:** Link

Fig. 5. Some benchmark eye movement datasets over video stimuli for evaluating visual attention prediction.

sition, visual masking, and center-surround interactions are some of the features implemented in this model. Later, Le Meur *et al.* [138] extended this model to spatio-temporal domain by fusing achromatic, chromatic and temporal information. In this new model, early visual features are extracted from the visual input into several separate parallel channels. A feature map is obtained for each channel, then a unique saliency map is built from the combination of those channels. The major novelty proposed here lies in the inclusion of the temporal dimension as well as the addition of a coherent normalization scheme.

Navalpakkam and Itti [51] modeled visual search as a top-down gain optimization problem by maximizing the signal-to-noise ratio (SNR) of the target vs. distractors instead of learning explicit fusion functions. That is, they learned linear weights for feature combination by maximizing the ratio between target saliency and distractor saliency.

Kootstra et al. [136] developed three symmetry-saliency operators and compared them with human eye tracking data. Their method is based on the isotropic symmetry and radial symmetry operators of Reisfeld *et al.* [137] and the color symmetry of Heidemann [64]. Kootstra *et al.* extended these operators to multi-scale symmetry-saliency models. The authors showed that their model performs significantly better on symmetric stimuli compared to the Itti *et al.* [14].

Marat et al. [104] proposed a bottom-up approach for spatio-temporal saliency prediction in video stimuli. This model extracts two signals from the video stream corresponding to parvocellular and magnocellular cells of the retina. From these signals, two static and dynamic saliency maps are derived and fused into a spatio-temporal map. Prediction results of this model were better for the first few frames of each clip snippet.

Murray et al. [200] introduced a model based on a low

level vision system in three steps: 1) visual stimuli are processed according to what is known about the early human visual pathway (color-opponent and luminance channels, followed by a multi-scale decomposition), 2) a simulation of the inhibition mechanisms present in cells of the visual cortex normalize their response to stimulus contrast, and 3) information is integrated at multiple scales by performing an inverse wavelet transform directly on weights computed from the non-linearization of the cortical outputs.

Cognitive models have the advantage of expanding our view of biological underpinnings of visual attention. This further helps understanding computational principles or neural mechanisms of this process as well as other complex dependent processes such as object recognition.

3.2 Bayesian Models (B)

Bayesian modeling is used for combining sensory evidence with prior constraints. In these models, prior knowledge (e.g., scene context or gist) and sensory information (e.g., target features) are probabilistically combined according to Bayes' rule (e.g., to detect an object of interest).

Torralba [92] and *Oliva* et al. [140] proposed a Bayesian framework for visual search tasks. Bottom-up saliency is derived from their formulation as $\frac{1}{p(f|f_G)}$ where f_G represents a global feature that summarizes the probability density of presence of the target object in the scene, based on analysis of the scene gist. Following the same direction, Ehinger *et al.* [87] linearly integrated three components (bottom-up saliency, gist, and object features) for explaining eye movements in looking for people in a database of about 900 natural scenes.

Itti and Baldi [145] defined surprising stimuli as those which significantly change beliefs of an observer. This is modeled in a Bayesian framework by computing the KL divergence between posterior and prior beliefs. This notion is applied both over space (surprise arises when observing image features at one visual location affects the observer's beliefs derived from neighboring locations) and time (surprise then arises when observing image features at one point in time affects beliefs established from previous observations).

Zhang et al. [141] proposed a definition of saliency, known as SUN: Saliency Using Natural statistics, by considering what the visual system is trying to optimize when directing attention. The resulting model is a Bayesian framework in which bottom-up saliency emerges naturally as the self-information of visual features, and overall saliency (incorporating top-down information with bottom-up saliency) emerges as the point-wise mutual information between local image features and the search target's features when searching for a target. Since this model provides a general framework for many models, we describe it in more detail.

SUN's formula for bottom-up saliency is similar to the work of Oliva *et al.* [140], Torralba [92], and Bruce and Tsotsos [144], in that they are all based on the notion of self-information (local information). However, differences between current image statistics and natural statistics lead to radically different kinds of self-information. Briefly, the motivating factor for using self-information with the statistics of the current image is that a foreground object is likely to have features that are distinct from those of the background. Since targets are observed less frequently than background during an organism's lifetime, rare features are more likely to indicate targets.

Let Z denote a pixel in the image, C whether or not a point belongs to a target class and L the location of a point (pixel coordinates). Also, let F be the visual features of a point. Having these, the saliency s_z of a point z is defined as $P(C = 1|F = f_z, L = l_z)$ where f_z and l_z are the feature and location of z. Using the Bayes rule and assuming that features and locations are independent and conditionally independent given $C = 1$, then saliency of a point is:

$$log\ s_z = -log\ P(F = f_z) + log\ P(F = f_z|C = 1) + log\ P(C = 1|L = l_z) \quad (5)$$

The first term at the right side is the self-information (bottom-up saliency) and it depends only on the visual features observed at the point Z. The second term on the right is the log-likelihood which favors feature values that are consistent with prior knowledge of the target (e.g., if the target is known to be green the log-likelihood will take larger values for a green point than for a blue point). The third term is the location prior which captures top-down knowledge of the target's location and is independent of visual features of the object. For example, this term may capture knowledge about some target being often found in the top-left quadrant of an image.

Zhang et al. [142] extended the SUN model to dynamic scenes by introducing temporal filters (Difference of Exponentials) and fitting a generalized Gaussian distribution to the estimated distribution for each filter response. This was implemented by first applying a bank of spatio-temporal filters to each video frame, then for any video, the model calculates its features and estimates the bottom-up saliency for each point. The filters were designed to be both efficient and similar to the human visual system. The probability distributions of these spatio-temporal features were learned from a set of videos from natural environments.

Jia Li et al. [133] presented a Bayesian multi-task learning framework for visual attention in video. Bottom-up saliency modeled by multi-scale wavelet decomposition was fused with different top-down components trained by a multi-task learning algorithm. The goal was to learn task-related "stimulus-to-saliency" functions, similar to [101]. This model also learns different strategies for fusing bottom-up and top-down maps to obtain the final attention map.

Boccignone [55] addressed joint segmentation and saliency computation in dynamic scenes, using a mixture of Dirichlet processes as a basis for object-based visual attention. He also proposed an approach for partitioning a video into shots based on a foveated representation of a video.

A key benefit of Bayesian models is their ability to learn from data and their ability to unify many factors in a principled manor. Bayesian models can, for example, take advantage of the statistics of natural scenes or other features that attract attention.

3.3 Decision Theoretic Models (D)

The decision-theoretic interpretation states that perceptual systems evolve to produce decisions about the states of the surrounding environment that are optimal in a decision theoretic sense (e.g., minimum probability of error). The overarching point is that visual attention should be driven by optimality with respect to the end task.

Gao and Vasconcelos [146] argued that for recognition, salient features are those that best distinguish a class of interest from all other visual classes. They then defined top-down attention as classification with minimal expected error. Specifically, given some set of features $F = \{F_1, \cdots, F_d\}$, a location l and a class label C with $C_l = 0$ corresponding to samples drawn from the surround region and $C_l = 1$ corresponding to samples drawn from a smaller central region centered at l, the judgment of saliency then corresponds to a measure of mutual information, computed as $I(F, C) = \sum_{i=1}^{d} I(F_i, C)$. They used DOG and Gabor filters, measuring the saliency of a point as the KL divergence between the histogram of filter responses at the point and the histogram of filter responses in the surrounding region. In [185], the same authors used this framework for bottom-up saliency by combining it with center-surround image processing. They also incorporated motion features (optical flow) between pairs of consecutive images to their model to account for dynamic stimuli. They adopted a dynamic texture model using a Kalman filter in order to capture the motion patterns in dynamic scenes.

Here we show the Bayesian computation of (5) is a special case of the Decision theoretic model. Saliency computation in the entire decision theoretic approach boils down to calculating the target posterior probability $P(C = 1|F = f_z)$ (the output of their simple cells [215]). By applying Bayesian rule, we have:

$$P(C_l = 1|F_l = f_z) = \sigma\Big(log\frac{P(F_l = f_z|C_l = 1)P(C_l = 1)}{P(F_l = f_z|C_l = 0)P(C_l = 0)}\Big) \quad (6)$$

where $\sigma(x) = (1 + e^{-x})^{-1}$ is the sigmoid function. The log likelihood ratio inside the sigmoid can be trivially written (using the independence assumptions of [141]) as:

$$-logP(F = f_z|C = 0) + logP(F = f_z|C = 1) + \frac{P(C = 1|L = l_z)}{P(C = 0|L = l_z)} \quad (7)$$

which is the same as (5) under the following assumptions: 1) $P(F = f_z|C = 0) = P(F = f_z)$ and 2) $P(C = 0|L = l_z) = K$, for some constant K. Assumption 1 states that the feature distribution in the absence of the target is the same as the feature distribution for the set of natural images. Since the overwhelming majority of natural images do not have the target, this is really not much of an assumption. The two distributions are virtually identical. Assumption 2 simply states that the absence of the target is equally likely in all image locations. This, again, seems like a very mild assumption.

Because of above connections, both Decision theoretic and Bayesian approaches have a biologically plausible implementation, which has been extensively discussed by Vasconcelos and colleagues [223][147][215]. The Bayesian methods can be mapped to a network with a layer of simple cells and the decision theoretic models to a network with a layer of simple and a layer of complex cells. The simple cell layer in fact can also implement AIM [144] and Rosenholtz [191] models in Section 3.4, Elazary and Itti [90], and probably some more. So, while these models have not been directly derived from biology, they can be implemented as cognitive models.

Gao and Vasconcelos [147] used discriminant saliency model for visual recognition and showed good performance on PASCAL 2006 dataset.

Mahadevan and Vasconcelos [105] presented an unsupervised algorithm for spatio-temporal saliency based on biological mechanisms of motion-based perceptual grouping. It is an extension of the discriminant saliency model [146]. Combining center-surround saliency with the power of dynamic textures made their model applicable to highly dynamic backgrounds and moving cameras.

In *Gu* et al. [148], an activation map was first computed by extracting primary visual features and detecting meaningful objects from the scene. An adaptable retinal filter was applied to this map to generate "regions of interest" (ROIs whose locations correspond to these activation peaks and whose sizes were estimated by an iterative adjustment algorithm). The focus of attention was moved serially over the detected ROIs by a decision theoretic mechanism. The generated sequence of eye fixations was determined from a perceptual benefit function based on perceptual costs and rewards, while the time distribution of different ROIs was estimated by memory learning and decaying.

Decision theoretic models have been very successful in computer vision applications such as classification while achieving high accuracy in fixation prediction.

3.4 Information Theoretic Models (I)

These models are based on the premise that localized saliency computation serves to maximize information sampled from one's environment. They deal with selecting the most informative parts of a scene and discarding the rest.

Rosenholtz [191][193] designed a model of visual search which could also be used for saliency prediction over an image in free-viewing. First, features of each point, p_i, are derived in an appropriate uniform feature space (e.g., uniform color space). Then, from the distribution of the features, mean, μ, and covariance, $\sum$, of distractor features are computed. The model then defines target saliency as the Mahalanobis distance, Δ, between the target feature vector, T, and the mean of the distractor distribution, where $\Delta^2 = (T - \mu)^{'} \sum^{-1} (T - \mu)$. This model is similar to [92][141][160] in the sense that it estimates $1/P(x)$ (rarity of a feature or self-information) for each image location x. This model also underlies a clutter measure of natural scenes (same authors [189]). An online version of this model is available at [194].

Bruce and Tsotsos [144] proposed the AIM model (Attention based on Information Maximization) which uses Shannon's self-information measure for calculating saliency of image regions. Saliency of a local image region is the information that region conveys relative to its surroundings. Information of a visual feature X is $I(X) = -log\ p(X)$, which is inversely proportional to the likelihood of observing X (i.e., $p(X)$). To estimate $I(X)$, the probability density function $p(X)$ must be estimated. Over RGB images, considering a local patch of size $M \times N$, X has the high dimensionality of $3 \times M \times N$. To make the estimation of $p(X)$ feasible, they used ICA to reduce the dimensionality of the problem to estimating $3 \times M \times N$ 1D probability density functions. To find the bases of ICA, they used a large sample of RGB patches drawn from natural scenes. For a given image, the 1D pdf for each ICA basis vector is first computed using non-parametric density estimation. Then, at each image location, the probability of observing the RGB values in a local image patch is the product of the corresponding ICA basis likelihoods for that patch.

Hou and Zhang [151] introduced the Incremental Coding Length (ICL) approach to measure the respective entropy gain of each feature. The goal was to maximize the entropy of the sample visual features. By selecting features with large coding length increments, the computational system can achieve attention selectivity in both dynamic and static scenes. They proposed ICL as a principle by which energy is distributed in the attention system. In this principle, the salient visual cues correspond to unexpected features. According to the definition of ICL, these features may elicit entropy gain in the perception state and are therefore assigned high energy.

Mancas [152] hypothesized that attention is attracted by minority features in an image. The basic operation is to count similar image areas by analyzing histograms which makes this approach closely related to Shannon's self-information measure. Instead of comparing only isolated pixels it takes into account the spatial relationships of areas surrounding each pixel (e.g., mean and variance). Two types of rarity models are introduced: Global and Local. While global rarity considers uniqueness of features over entire image, some image details may still appear salient due to local contrast or rarity. Similar to the center-surround ideas of [14], they used a multi-scale approach for the computation of local contrast.

Seo and Milanfar [108] proposed the Saliency prediction by Self-Resemblance (SDSR) approach. First a local image structure at each pixel is represented by a matrix of local descriptors (local regression kernels), which are robust in the presence of noise and image distortions. Then, matrix cosine similarity (a generalization of cosine similarity) is employed to measure the resemblance of each pixel to its surroundings. For each pixel, the resulting saliency map represents the statistical likelihood of its feature matrix F_i given the feature matrices F_j of the surrounding pixels:

$$s_i = \frac{1}{\sum_{j=1}^{N} exp(\frac{-1+\rho(F_i,F_j)}{\sigma^2})} \quad (8)$$

where $\rho(F_i, F_j)$ is the matrix cosine similarity between two feature maps F_i and F_j, and σ is a local weighting parameter. The columns of local feature matrices represent the output of local steering kernels which are modeled as:

$$K(x_l - x_i) = \frac{\sqrt{det(C_i)}}{h^2} exp\left\{\frac{(x_l - x_i)^T C_l (x_l - x_i)}{-2h^2}\right\} \quad (9)$$

where $l = 1, ..., P$, P is the number of the pixels in a local window, h is a global smoothing parameter, and the matrix C_l is a covariance matrix estimated from a collection of spatial gradient vectors within the local analysis window around a sampling position $x_l = [x_1, x_2]_l^T$.

Yin Li et al. [171] proposed a visual saliency model based on conditional entropy for both image and video. Saliency was defined as the minimum uncertainty of a local region given its surrounding area (namely the minimum conditional entropy), when perceptional distortion is considered. They approximated the conditional entropy by the lossy coding length of multivariate Gaussian data. The final saliency map was accumulated by pixels and further segmented to detect the proto-objects. Yan *et al.* [186] proposed a newer version of this model by adding a multi resolution scheme to it.

Wang et al. [201], introduced a model to simulate human saccadic scanpaths on natural images by integrating three related factors guiding eye movements sequentially: 1) reference sensory responses, 2) fovea-periphery resolution discrepancy, and 3) visual working memory. They compute three multi-band filter response maps for each eye movement which are then combined into multi-band residual filter response maps. Finally, they compute residual perceptual information (RPI) at each location. The next fixation is selected as the location with the maximal RPI value.

3.5 Graphical Models (G)

A graphical model is a probabilistic framework in which a graph denotes the conditional independence structure between random variables. Attention models in this category treat eye movements as a time series. Since there are hidden variables influencing the generation of eye movements, approaches like Hidden Markov Models (HMM), Dynamic Bayesian Networks (DBN), and Conditional Random Fields (CRF) have been incorporated.

Salah et al. [52] proposed an approach for attention and applied it to handwritten digit and face recognition. In the first step (Attentive level), a bottom-up saliency map is constructed using simple features. In the intermediate level "what" and "where" information is extracted by dividing the image space into uniform regions and training a single-layer perceptron over each region in a supervised manner. Eventually this information is combined at the associative level with a discrete Observable Markov Model (OMM). Regions visited by a fovea are treated as states of the OMM. An inhibition of return allows the fovea to focus on the other positions in the image.

Liu et al. [43] proposed a set of novel features and adopted a Conditional Random Field to combine these features for salient object detection on their regional saliency dataset. Later, they extended this approach to detect salient object sequences in videos [48]. They presented a supervised approach for salient object detection, formulated as an image segmentation problem using a set of local, regional and global salient object features. A CRF was trained and evaluated on a large image database containing 20,000 labeled images by multiple users.

Harel et al. [121] introduced Graph-Based Visual Saliency (GBVS). They extract feature maps at multiple spatial scales. A scale-space pyramid is first derived from image features: intensity, color, and orientation (similar to Itti *et al.* [14]). Then, a fully-connected graph over all grid locations of each feature map is built. Weights between two nodes are assigned proportional to the similarity of feature values and their spatial distance. The dissimilarity between two positions (i, j) and (p, q) in the feature map, with respective feature values $M(i, j)$ and $M(p, q)$, is defined as:

$$d((i,j) \parallel (p,q)) = |log\frac{M(i,j)}{M(p,q)}| \quad (10)$$

The directed edge from node (i, j) to node (p, q) is then assigned a weight proportional to their dissimilarity and their distance on lattice M:

$$w((i,j),(p,q)) = d((i,j) \parallel (p,q)).F(i-p, j-q)$$
$$\text{where } F(a,b) = exp\Big(-\frac{a^2+b^2}{2\sigma^2}\Big) \quad (11)$$

The resulting graphs are treated as Markov chains by normalizing the weights of the outbound edges of each node

to 1 and by defining an equivalence relation between nodes and states, as well as between edge weights and transition probabilities. Their equilibrium distribution is adopted as the activation and saliency maps. In the equilibrium distribution, nodes that are highly dissimilar to surrounding nodes will be assigned large values. The activation maps are finally normalized to emphasize conspicuous detail, and then combined into a single overall map.

Avraham et al. [153] introduced the E-saliency (Extended saliency) model by utilizing a graphical model approximation to extend their static saliency model based on self similarities. The algorithm is essentially a method for estimating the probability that a candidate is a target. The E-Saliency algorithm is as follows: 1) Candidates are selected using some segmentation process, 2) The preference for a small number of expected targets (and possibly other preferences) is used to set the initial (prior) probability for each candidate to be a target, 3) The visual similarity is measured between every two candidates to infer the correlations between the corresponding labels, 4) Label dependencies are represented using a Bayesian network, 5) The N most likely joint label assignments are found, and 6) Saliency of each candidate is deduced by marginalization.

Pang et al. [102] presented a stochastic model of visual attention based on the signal detection theory account of visual search and attention [155]. Human visual attention is not deterministic and people may attend to different locations on the same visual input at the same time. They proposed a dynamic Bayesian network to predict where humans typically focus in a video scene. Their model consists of four layers. In the first layer, a saliency map (Itti's) is derived that shows the average saliency response in each location in a video frame. Then in the second layer, a stochastic saliency map converts the saliency map into natural human responses through a Gaussian state space model. As to the third layer, an eye movement pattern controls the degree of overt shifts of attention through a Hidden Markov Model and finally an eye focusing density map predicts positions that people likely pay attention to based on the stochastic saliency map and eye movement patterns. They reported a significant improvement in eye fixation detection over previous efforts at the cost of decreased speed.

Chikkerur et al. [154] proposed a model similar to the model of Rao *et al.* [217] based on assumptions that the goal of the visual system is to know what is where and that visual processing happens sequentially. In this model, attention emerges as the inference in a Bayesian graphical model which implements interactions between ventral and dorsal areas. This model is able to explain some physiological data (neural responses in ventral stream (V4 and PIT) and dorsal stream (LIP and FEF)) as well as psychophysical data (human fixations in free viewing and search tasks).

Graphical models could be seen as a generalized version of Bayesian models. This allows them to model more complex attention mechanisms over space and time which results in good prediction power (e.g., [121]). The drawbacks lie in model complexity, especially when it comes to training and readability.

3.6 Spectral Analysis Models (S)

Instead of processing an image in the spatial domain, models in this category derive saliency in the frequency domain.

Hou and Zhang [150] developed the spectral residual saliency model based on the idea that similarities imply redundancies. They propose that statistical singularities in the spectrum may be responsible for anomalous regions in the image, where proto-objects become conspicuous. Given an input image $I(x)$, amplitude $\mathcal{A}(f)$ and phase $\mathcal{P}(f)$ are derived. Then, the log spectrum $\mathcal{L}(f)$ is computed from the down-sampled image. From $\mathcal{L}(f)$, the spectral residual $\mathcal{R}(f)$ can be 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. Using the inverse Fourier transform, they construct the saliency map in the spatial domain. The value of each point in the saliency map is then squared to indicate the estimation error. Finally, they smooth the saliency map with a Gaussian filter $g(x)$ for better visual effect. The entire process is summarized below:

$$\begin{aligned} \mathcal{A}(f) &= \mathcal{R}\Big(\mathcal{F}[I(x)]\Big), \\ \mathcal{P}(f) &= \varphi\Big(\mathcal{F}[I(x)]\Big), \\ \mathcal{L}(f) &= log\Big(\mathcal{A}(f)\Big), \\ \mathcal{R}(f) &= \mathcal{L}(f) - h_n(f) * \mathcal{L}(f), \\ \mathcal{S}(x) &= g(x) * \mathcal{F}^{-1}\Big[exp\Big(\mathcal{R}(f) + \mathcal{P}(f)\Big)\Big]^2 \end{aligned} \tag{12}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the Fourier and Inverse Fourier Transforms, respectively. $\mathcal{P}$ denotes the phase spectrum of the image, and is preserved during the process. Using a threshold they find salient regions called proto objects for fixation prediction. As a testament to its conceptual clarity, residual saliency could be computed in 5 lines of Matlab code [187]. But note that these lines exploit complex functions that has long implementations (e.g., $\mathcal{F}$ and $\mathcal{F}^{-1}$).

Guo et al. [156] showed that incorporating the phase spectrum of the Fourier transform instead of the amplitude transform leads to better saliency predictions. Later, Guo *et al.* [157] proposed a quaternion representation of an image combining intensity, color, and motion features. They called this method "phase spectrum of quaternion Fourier transform (PQFT)" for computing spatio-temporal saliency and applied it to videos. Taking advantage of the multi-resolution representation of the wavelet, they also proposed a foveation approach to improve coding efficiency in video compression.

Achanta et al. [158] implemented a frequency-tuned approach to salient region detection using low-level features of color and luminance. First, the input RGB image I is transformed to CIE Lab color space. Then, the scalar saliency map S for image I is computed as: $S(x,y) = \|I_\mu - I_{\omega_{hc}}\|$ where I_μ is the arithmetic mean image feature vector, $I_{\omega_{hc}}$ is a Gaussian blurred version of the original image using a 5×5 separable binomial kernel, $\|.\|$ is the L_2 norm (Euclidean distance), and x, y are the pixel coordinates.

Bian and Zhang [159] proposed the Spectral Whitening (SW) model based on the idea that visual system bypasses the redundant (frequently occurring, non-informative) features while responding to rare (informative) features. They used spectral whitening as a normalization procedure in the construction of a map that only represents salient features and localized motion while effectively suppressing redundant (non-informative) background information and ego-motion. First, a grayscale input image $I(x,y)$ is low-pass fil-

tered and subsampled. Next, a windowed Fourier transform of the image is calculated as: $f(u,v) = F[w(I(x,y))]$, where F denotes the Fourier transform and w is a windowing function. The normalized (flattened or whitened) spectral response $\left(n(u,v) = f(u,v)/\|f(u,v)\|\right)$ is transformed into the spatial domain through the inverse Fourier transform (F^{-1}) squared to emphasize salient regions. Finally it is convolved with a Gaussian low-pass filter $g(u,v)$ to model the spatial pooling operation of complex cells: $S(x,y) = g(u,v) * \|F^{-1}[n(u,v)]\|^2$.

Spectral analysis models are simple to explain and implement. While still very successful, biological plausibility of these models is not very clear.

3.7 Pattern Classification Models (P)

Machine learning approaches have also been used in modeling visual attention by learning models from recorded eye-fixations or labeled salient regions. Typically, attention control works as a "stimuli-saliency" function to select, reweight, and integrate the input visual stimuli. Note that these models may not be purely bottom-up since they use features that guide top-down attention (e.g., faces or text).

Kienzle et al. [165] introduced a non-parametric bottom-up approach for learning attention directly from human eye tracking data. The model consists of a nonlinear mapping from an image patch to a real value, trained to yield positive outputs on fixations, and negative outputs on randomly selected image patches. The saliency function is determined by its maximization of prediction performance on the observed data. A support vector machine (SVM) was trained to determine the saliency using the local intensities. For videos, they proposed to learn a set of temporal filters from eye-fixations to find the interesting locations.

The advantage of this approach is that it does not need a priori assumptions about features that contribute to salience or how these features are combined to a single salience map. Also this method produces center-surround operators analogous to receptive fields of neurons in early visual areas (LGN and V1).

Peters and Itti [101] trained a simple regression classifier to capture the task-dependent association between a given scene (summarized by its gist) and preferred locations to gaze at while human subjects were playing video games. During testing of the model, the gist of a new scene is computed for each video frame, and is used to compute the top-down map. They showed that a point-wise multiplication of bottom-up saliency with the top-down map learned in this way results in higher prediction performance.

Judd et al. [166], similar to Kienzle *et al.* [165], trained a linear SVM from human fixation data using a set of low, mid, and high-level image features to define salient locations. Feature vectors from fixated locations and random locations, were assigned $+1$ and -1 class labels, respectively. Their results over a dataset of 1003 images observed by 15 subjects (gathered by the same authors) show that combining all aforementioned features plus distance from image center produces the best eye fixation prediction performance.

As available eye movement data increases and with wider spread of eye tracking devices supporting gathering mass data, these models are becoming popular. This however, makes models data-dependent thus influencing fair model comparison, slow, and to some extent, black-box.

3.8 Other Models (O)

Some other attention models that do not fit into our categorization are discussed below.

Ramstrom and Christiansen [168] introduced a saliency measure using multiple cues based on game theory concepts inspired by the selective tuning approach of Tsotsos *et al.* [15]. Feature maps are integrated using a scale pyramid where the nodes are subject to trading on a market and the outcome of the trading represents the saliency. They use the spot-light mechanism for finding regions of interest.

Rao et al. [23] proposed a template matching type of model by sliding a template of the desired target to every location in the image and at each location compute salience as some similarity measure between template and local image patch.

Ma et al. [33] proposed a user attention model to video contents by incorporating top-down factors into the classical bottom-up framework by extracting semantic cues (e.g., face, speech, and camera motion). First, the video sequence is decomposed into primary elements of basic channels. Next, a set of attention modeling methods generate attention maps separately. Finally, fusion schemes are employed to obtain a comprehensive attention map which may be used as importance ranking or the index of video content. They applied this model to video summarization.

Rosin [169] proposed an edge-based scheme (EDS) for saliency detection over grayscale images. First, a Sobel edge detector is applied to the input image. Second, the graylevel edge image is thresholded at multiple levels to produce a set of binary edge images. Third, a distance transform is applied to each of the binary edge images to propagate the edge information. Finally, the gray-level distance transforms are summed to obtain the overall saliency map. This approach has not been successful over color images.

Garcia-Diaz et al. [160] introduced the Adaptive Whitening Saliency (AWS) model by adopting the variability in local energy as a measure of saliency estimation. The input image is transformed to *Lab* color space. The luminance (L) channel is decomposed into multi-oriented multi-resolution representation by means of Gabor-like bank of filters. The opponent color components a and b undergo a multi-scale decomposition. By decorrelating the multi-scale responses, extracting from them a local measure of variability, and further performing a local averaging they obtained a unified and efficient measure of saliency. Decorrelation is achieved by applying PCA over a set of multi-scale low level features. Distinctiveness is measured using the Hoteling's T^2 statistic.

Goferman et al. [46] proposed a context-aware saliency detection model. Salient image regions are detected based on four principles of human attention: 1) Local low-level considerations such as color and contrast, 2) Global considerations which suppress frequently occurring features while maintaining features that deviate from the norm, 3) Visual organization rules which state that visual forms may possess one or several centers of gravity about which the form is organized, and 4) High-level factors, such as human faces. They applied their saliency method to two applications: retargeting and summarization.

Aside from the models discussed so far, there are several other attention models that are relevant to the topic of this review, though they do not explicitly generate saliency maps. Here we mention them briefly.

To overcome the problem of designing the state-space for a complex task, an approach proposed by Sprague and Ballard [109] decomposes a complex temporally-extended task to simple behaviors (also called micro-behaviors), one of which is to attend to obstacles or other objects in the world. This behavior-based approach learns each micro-behavior and uses arbitration to compose these behaviors and solve complex tasks. This complete agent architecture is of interest as it studies the role of attention while it interacts and shares limited resources with other behaviors.

Based on the idea that vision serves action, Jodogne *et al.* [162] introduced an approach for learning action-based image classification known as Reinforcement Learning of Visual Classes (RLVC). RLVC consists of two interleaved learning processes: An RL unit which learns image to action mappings and an image classifier which incrementally learns to distinguish visual classes. RLVC is a feature-based approach in which the entire image is processed to find out whether a specific visual feature exists or not in order to move in a binary decision tree. Inspired by RLVC and U-TREE [163], Borji *et al.* [88] proposed a three-layered approach for interactive object-based attention. Each time the object that is most important to disambiguate appears, a partially unknown state is attended by the biased bottom-up saliency model and recognized. Then the appropriate action for the scene is performed. Some other models in this category are: Triesch *et al.* [97], Mirian *et al.* [100], and Paletta *et al.* [164].

Walker *et al.* [21] built a model based on the idea that humans fixate at those informative points in an image which reduce our overall uncertainty about the visual stimulus - similar to another approach by Lee and Yu [149]. This model is a sequential information maximization approach whereby each fixation is aimed at the most informative image location given the knowledge acquired at each point. A foveated representation is incorporated with reducing resolution as distance increases from the center. Shape histogram edges are used as features.

Lee and Yu [149] proposed that mutual information among the cortical representations of the retinal image, the priors constructed from our long-term visual experience, and a dynamic short-term internal representation constructed from recent saccades, all provide a map for guiding eye navigations. By directing the eyes to locations of maximum complexity in neuronal ensemble responses at each step, the automatic saccadic eye movement system greedily collects information about the external world while modifying the neural representations in the process. This model is close to Najemnik & Geisler's work [20].

To recap, here, we offer a unification of several saliency models from a statistical viewpoint. The first class measures bottom-up saliency as $1/P(x)$ or $logP(x)$ or $E_X[-logP(x)]$ which is the entropy. This includes Torralba and Oliva [92][93], SUN [141], AIM [144], Hou and Zhang [151], and probably Yin Li [171]. Some other methods are equivalent to this but with specific assumptions for $P(x)$. For example, Rosenholtz [191] assume a Gaussian, and Seo and Milanfar [108] assumes that $P(x)$ is a kernel density estimate (with the kernel that appears inside the summation on the denominator of (7)). Next, there is a class of top-down models with the same saliency measure. For example, Elazary and Itti [90] use $logP(x|Y=1)$ (where $Y=1$ means target presence) and assume a Gaussian

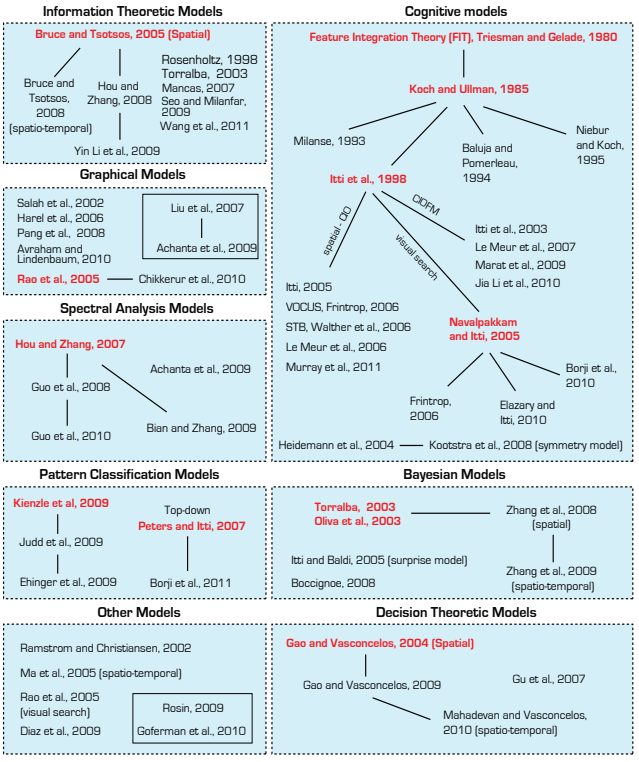

Fig. 6. A hierarchical illustration of described models. Solid rectangles show salient region detection methods.

for $P(x|Y=1)$. SUN can also be seen like this, if you call the first term of (5) a bottom-up component. But, as discussed next, it is probably better to just consider it an approximation to the methods in the third class. The third class includes models that compute posterior probabilities $P(Y=1|X)$ or likelihood ratios $log[P(x|Y=1)/P(x|Y=0)]$. This is the case of discriminant saliency [146][147][215] but also appears in Harel *et al.* [121] (e.g. equation 10) and in Liu *et al.* [43] (if you set the interaction potentials of a CRF to zero, you end up with a computation of the posterior $P(Y=1|X)$ at each location). All these methods model the saliency of each location independently of the others. The final class, graphical models, introduces connections between spatial neighbors. These could be clique potentials in CRFs, edge weights in Harel *et al.* [121], etc.

Fig. 6 shows a hierarchical illustration of models. A summary of attention models and their categorization according to factors mentioned in section 2 is presented in Fig. 7.

4 Discussion

There are a number of outstanding issues with attention models that we discuss next.

A big challenge is the degree to which a model agrees with biological findings. Why is such an agreement important? How can we judge whether a model is indeed biologically plausible? While there is no clear answer to these questions in the literature, here we give some hints at their answer. In the context of attention, biologically inspired models have resulted in higher accuracies in some cases. In support of this statement, the Decision theoretic [147][223] and (later) AWS model [160] (and perhaps some other models) are good examples because they explains

No	Model	Year	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13
Bottom-up (saliency models)															
1	Itti et al. [14]	1998	+	-	-	+	-	-	+	f	+	CIO	C	-	-
2	Privitera & Stark [127]	2000	+	-	-	+	-	-	+	f	+	-	O	-	Stark and Choi
3	Salah et al. [52]	2002	+	+	-	+	-	-	+	-	+	O	G	DR	Digit & Face
4	Itti et al. [119]	2003	+	-	+	+	+	+	+	f	+	CIOFM	C	-	-
5	Torralba [92]	2003	-	+	-	+	-	-	+	s	+	CI	B	DR	Torralba et al.
6	Sun & Fisher [117]	2003	+	-	-	+	-	-	+	-	-	CIO	G	-	-
7	Gao & Vasconcelos [146]	2004	-	+	-	+	-	-	+	s	-	DCT	D	DR	Brodatz, Caltech
8	Ouerhani et al. [210]	2004	+	-	-	+	-	-	+	f	+	CIO+Corner	C	CC	Ouerhani
9	Boccignone & Ferraro [175]	2004	+	-	+	-	+	-	+	f	-	Optical Flow	B	-	BEHAVE
10	Frintrop [50]	2005	+	+	+	+	+	+	+	f/s	+/-	CIOM	C	-	-
11	Itti & Baldi [145]	2005	+	-	+	+	+	+	-	f	+	CIOFM	B	KL, AUC	ORIG-MTV
12	Ma et al. [33]	2005	+	-	+	+	-	-	+	f	+	M*	O	-	-
13	Bruce & Tsotsos [144]	2006	+	-	-	+	-	+	+	f	+	DOG, ICA	I	KL, ROC	Bruce and Tsotsos
14	Navalpakkam & Itti [51]	2006	-	+	-	+	-	+	+	s	+	CIO	C	-	-
15	Zhai & Shah [103]	2006	+	-	+	+	+	-	+	f	+	SIFT	O	-	-
16	Harel et al. [121]	2006	+	-	-	+	-	-	+	f	+	IO	G	AUC	Bruce and Tsotsos
17	Le Meur et al. [41]	2006	+	-	-	+	-	-	+	f	+	LM*	C	CC, KL	Le Meur et al.
18	Walther & Koch [35]	2006	+	-	-	+	-	+	+	f	+/-	CIO	C	-	-
19	Peters & Itti [101]	2007	+	+	+	+	+	-	+	i	+	CIOFM	P	KL, NSS	Peters and Itti
20	Liu et al. [43]	2007	+	-	-	+	-	-	+	f	-	Liu*	G	F-measure	Regional
21	Shic & Scassellati [74]	2007	+	-	+	+	+	-	+	f	+	CIOM	C	ROC	Shic and Scassellati
22	Hou & Zhang [150]	2007	+	-	-	+	-	+	+	f	+	FFT, DCT	S	NSS	DB of Hou and Zhang, 2007
23	Cerf et al. [167]	2007	+	+	-	+	-	+	+	f/s	+	CIO :)	C	AUC	Cerf et al.
24	Le Meur et al. [138]	2007	+	-	+	+	+	-	+	f	+	LM*	C	CC, KL	Le Meur et al.
25	Mancas [152]	2007	+	-	+	+	+	+	+	f	+	CI	I	CC	Le Meur et al.
26	Guo et al. [156]	2008	+	-	-	+	-	-	+	f	+	CIO	D	CC	Self data
27	Zhang et al. [141]	2008	+	-	-	+	-	+	+	f	+	DOG, ICA	B	KL, AUC	Bruce and Tsotsos
28	Hou & Zhang [151]	2008	+	-	+	+	+	-	+	f	+	ICA	I	AUC, KL	Bruce and Tsotsos, ORIG
29	Pang et al. [102]	2008	+	+	+	+	+	-	+	f	+	CIOM	G	NSS	ORIG, Self data
30	Kootstra et al. [136]	2008	+	-	-	+	-	-	+	f	+	Symmetry	C	CC	Kootstra et al.
31	Ban et al. [172]	2008	+	-	+	+	+	-	+	f	+	CIO+SYM	I	-	-
32	Rajashekar et al. [174]	2008	+	-	-	+	-	-	+	f	+	R*	S	CC	Rajashekar et al.
33	Kienzle et al. [165]	2009	+	-	-	+	-	-	+	f	+	I	P	K*	Kienzle et al.
34	Marat et al. [49]	2009	+	-	+	+	+	-	+	f	+	SM*	C	NSS	Marat et al.
35	Judd et al. [166]	2009	+	-	-	+	-	-	+	f	+	J*	P	AUC	Judd et al.
36	Seo & Milanfar [108]	2009	+	-	+	+	+	+	+	f	+	LSK	I	AUC, KL	Bruce and Tsotsos, ORIG
37	Rosin [169]	2009	+	-	-	+	-	-	+	f	+	C+ Edge	O	PR, F-measure	DB of Liu et al, 2007
38	Yin Li et al. [171]	2009	-	+	+	+	+	+	+	s	+	RGB	S	DR	DB of Hou and Zhang, 2007
39	Bian & Zhang [159]	2009	+	-	+	+	+	+	+	f	+	FFT	S	AUC	Bruce and Tsotsos
40	Diaz et al. [160]	2009	+	-	-	+	-	+	+	f	+	CIO	O	AUC	Bruce and Tsotsos
41	Zhang et al. [142]	2009	+	-	+	-	+	-	+	f	+	DOG, ICA	B	KL, AUC	Bruce and Tsotsos
42	Achanta et al. [158]	2009	+	-	-	+	-	-	+	f	+	DOG	S	PR	DB of Liu et al, 2007
43	Gao et al. [147]	2009	+	-	+	+	+	+	+	f	+	CIO	D	AUC	Bruce and Tsotsos
44	Chikkerur et al. [154]	2010	+	+	-	+	-	+	+	f/s	+/-	CIO	B	AUC	Bruce and Tsotsos, Chikkerur
45	Mahadaven & Vasconcelos [106]	2010	+	-	+	-	+	-	+	-	+	I	D	DR, AUC	SVCL background data
46	Avraham & Lindenbaum [153]	2010	+	+	-	+	-	+	+	f/s	+/-	CIO	G	DR, CC	UWGT, Ouerhani et al.
47	Jia Li et al. [133]	2010	-	+	+	+	+	-	+	f	+	CIO	B	AUC	RSD, MTV, ORIG, Peters and Itti
48	Guo et al. [157]	2010	+	-	+	+	+	+	+	f/s	+/-	FFT	S	DR	Self data
49	Borji et al. [89]	2010	-	+	-	+	-	+	+	s	+/-	CIO	O	DR	-
50	Goeferman et al. [46]	2010	+	-	-	+	-	-	+	-	+	C :)	O	AUC	DB of Hou and Zhang, 2007
51	Murray et al. [200]	2011	+	-	-	+	-	-	+	f	+	CIO	C	AUC, KL	Bruce and Tsotsos, Judd et al.
52	Wang et al. [201]	2011	+	-	-	+	-	-	+	f	+	ICA	I	AUC	Self data
Top-down (general attention models)															
53	McCallum [163]	1995	-	+	-	+	-	+	-	i	+	-	R	-	Self data
54	Rao et al. [23]	1995	-	+	-	+	-	-	+	s	+	CIO	O	-	Self data
55	Ramstrom & Christiansen [168]	2002	-	+	-	+	-	-	+	-	+	CI	O	-	-
56	Sprague & Ballard [109]	2003	-	+	+	-	+	+	+	i	-	S*	R	-	-
57	Renninger et al. [94]	2004	-	+	-	+	-	+	-	s	-	Edgelet	I	DR	Self data
58	Navalpakkam & Itti [80]	2005	-	+	-	+	-	+	+	-	+	CIO	C	-	Self data
59	Paletta et al. [164]	2005	-	+	-	+	-	-	+	-	-	SIFT	R	DR	COIL-20, TSG-20
60	Jodogne & Piater [162]	2007	-	+	-	+	-	-	+	i	-	SIFT	R	-	-
61	Butko & Movellan [161]	2009	-	+	+	+	+	+	+	s	-	-	R	-	-
62	Verma & McOwan [214]	2009	+	-	-	+	-	+	-	s	-	CIO	O	-	-
63	Borji et al. [89]	2010	-	+	-	+	-	-	+	i	-	CIO	R	-	-

Fig. 7. Summary of visual attention models. 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: 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.

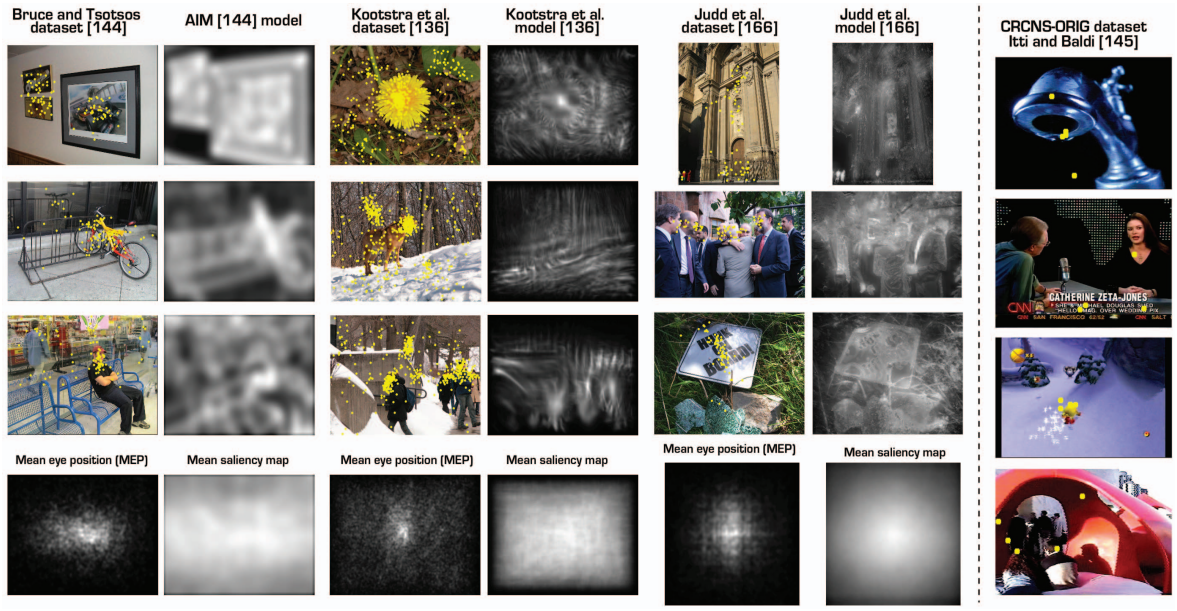

Fig. 8. Sample images from image and video datasets along with eye fixations and predicted attention maps. As could be seen, human and animal body and face, symmetry, and text attract human attention. Fourth row shows that these datasets are highly center-biased mainly because there are some interesting objects at the image center (MEP map). Less center-bias at mean saliency map of models indicates that a Gaussian in average works better than many models.

some basic behavioral data (e.g., nonlinearity against orientation contrast, efficient (parallel) and inefficient (serial) search, orientation and presence-absence asymmetries, and Weber's law [75]) well that has been less explored by other models. These models are among the best in predicting fixations over images and videos [160]. Hence, biological plausibility could be rewarding. We believe that creating a standard set of experiments for judging biological plausibility of models would be a promising direction to take. For some models, prediction of fixations is more important than agreement with biology (e.g., pattern classification vs. cognitive models). These models usually feed features to some classifier - but what type of features or classifiers fall under the realm of biologically inspired techniques? The answer lies in the behavioral validity of each individual feature as well as the classifier (e.g., faces or text, SVM vs. Neural Networks). Note that these problems are not specific to attention modeling and are applicable to other fields in computer vision (e.g., object detection and recognition).

Regarding fair model comparison, results often disagree when using different evaluation metrics. Therefore, a unified comparison framework is required - one that standardizes measures and datasets. We should also discuss the treatment of image borders and its influence on results. For example, KL and NSS measures are corrupted by an edge effect due to variations in handling invalid filter responses at the image borders. Zhang *et al.* [141] studied the impact of varying amounts of edge effects on ROC score over a dummy saliency map (consisting of all ones) and showed that as the border increases, AUC and KL measures increase as well. The dummy saliency map gave an ROC value of 0.5, a four-pixel black border gave 0.62, and an eight-pixel black border map gave 0.73. The same 3 border sizes would yield KL scores of 0, 0.12, and 0.25. Another challenge is handling the center-bias that results from a high density of eye fixations at the image center. Because of this, a trivial Gaussian blob model scores higher than almost all saliency models (see [166]). This can be partially verified from the average eye fixation maps of three popular datasets shown in Fig. 8. Comparing the mean saliency map of models and the fixation distributions, it could be seen that Judd et al. [166] model has higher center-bias due to explicitly using the center feature, which leads to higher eye movement prediction for this model as well. To eliminate the border and center-bias effects, Zhang *et al.* [141] defined an unshuffled AUC metric instead of the uniform AUC metric: for an image, the positive sample set is composed of the fixations of all subjects on that image and the negative set is composed of the union of all fixations across all images - except for the positive samples.

As shown by Figs. 4 and 5 many different eye movement datasets are available, each one recorded in different experimental conditions with different stimuli and tasks. Yet more datasets are needed because the available ones suffer from several drawbacks. Consider that current datasets do not tell us about covert attention mechanisms at all and can only tell us about overt attention (eye tracking). One approximation can compare overt attention shifts to verbal or other reports, whereby reported objects that were not fixated might have been covertly attended to. There is also a lack of multimodal datasets in interactive environments. In this regard, a promising new effort is to create tagged object datasets similar to video LabelMe [188]. Bruce and Tsotsos [144] and ORIG [184] are respectively the most widely used image and video datasets though they are highly center-biased (see Fig. 8). Thus there is a need for standard benchmark datasets as well as rigorous performance measures for attention modeling. Similar efforts have already been started amongst other research communities, such as object recognition (PASCAL challenge), text information retrieval (TREC datasets), and face recognition (e.g., FERET).

The majority of models are bottom-up though it is known

that top-down factors play a major role in directing attention [177]. However, the field of attention modeling lacks principled ways to model top-down attention components as well as the interaction of bottom-up and top-down factors. Feed-forward bottom-up models are general, easy to apply, do not need training, and yield reasonable performance making them good heuristics. On the other hand, top-down definitions usually use feedback and employ learning mechanisms to adapt themselves to specific tasks/environments and stimuli, making them more powerful but more complex to deploy and test (e.g., need to train on large datasets).

Some models need many parameters to be tuned while some others need fewer (e.g., spectral saliency models). Methods such as Gao *et al.* [147], Itti *et al.* [14], Oliva *et al.* [140], and Zhang *et al.* [142]) are based on Gabor or DOG filters and require many design parameters such as the number and type of filters, choice of non-linearities, and normalization schemes. Properly tuning the parameters is important in performance of these types of models.

Fig. 9 presents sample saliency maps of some models discussed in this paper.

5 Summary and Conclusion

In this paper, we discussed recent advances in modeling visual attention with an emphasis on bottom-up saliency models. A large body of past research was reviewed and organized in a unified context by qualitatively comparing models over 15 experimental criteria. Advancement in this field could greatly help solving other challenging vision problems such as cluttered scene interpretation and object recognition. In addition, there are many technological applications that can benefit from it. Several factors influencing bottom-up visual attention have been discovered by behavioral researchers and have further inspired the modeling community. However, there are several other factors remaining to be discovered and investigated. Incorporating those additional factors may help to bridge the gap between human inter-observer (a map built from fixations of other subjects over the same stimulus) and prediction accuracy of computational models. With the recent rapid progress, there is hope this may be accessible in the near future.

Most of the previous modeling research has been focused on the bottom-up component of visual attention. While previous efforts are appreciated, the field of visual attention still lacks computational principles for task-driven attention. A promising direction for future research is the development of models that take into account time varying task demands, especially in interactive, complex, and dynamic environments. In addition, there is not yet a principled computational understanding of covert and overt visual attention, which should be clarified in the future. The solutions are beyond the scope of computer vision and require collaboration from the machine learning community.

Acknowledgments

This work was supported by Defense Advanced Research Projects Agency (government contract no. HR0011-10-C-0034), the National Science Foundation (CRCNS grant number BCS-0827764), the General Motors Corporation, and the Army Research Office (grant number W911NF-08-1-0360).

The authors would like to thank reviewers for their helpful comments on the paper.

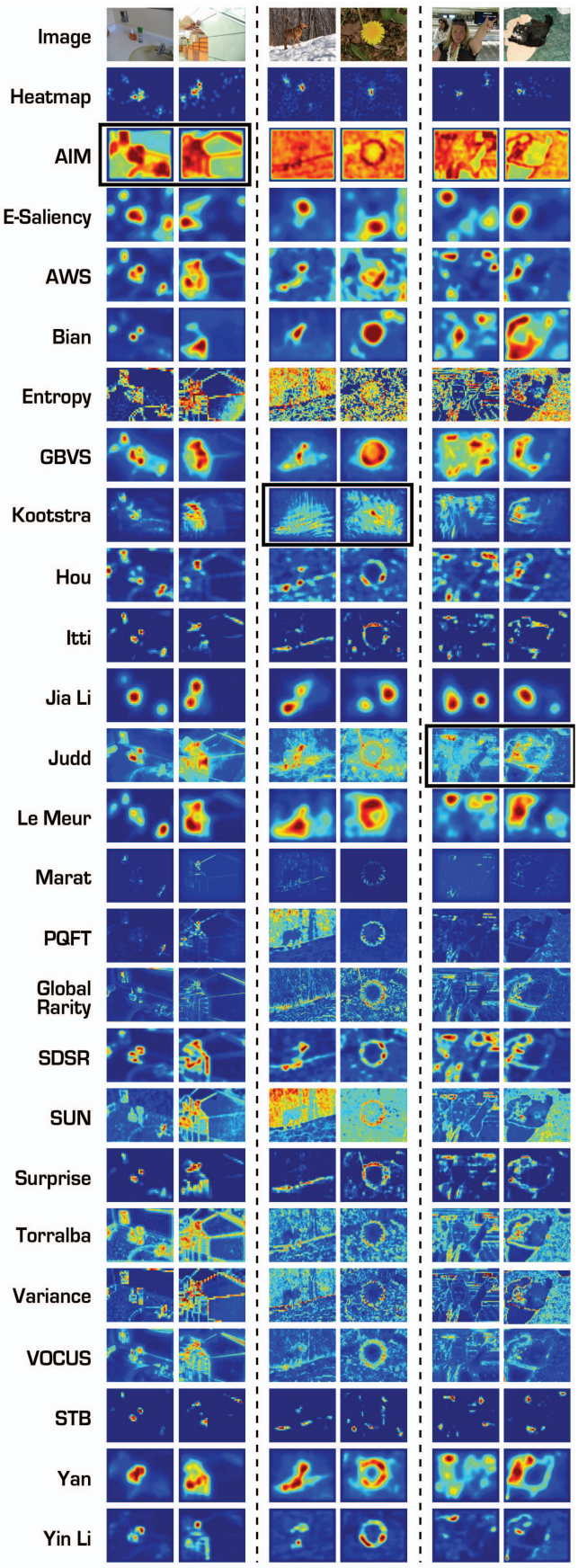

Fig. 9. Sample saliency maps of models over Bruce and Tsotsos (left), Kootstra *et al.* (middle), and Judd et al. datasets. Black rectangles means dataset was first used by that model.

REFERENCES

[1] K. Koch, J. McLean, R. Segev, M.A. Freed, M.J. Berry, V. Balasubramanian, and P. Sterling, "How Much the Eye Tells the Brain," *Current Biology*, vol. 25, no. 16(14), pp. 1428-34, 2006.

[2] L. Itti, Models of Bottom-Up and Top-Down Visual Attention, California Institute of Technology, PhD. Thesis, 2000.

[3] D.J. Simons and D.T. Levin, "Failure to Detect Changes to Attended Objects," *Investigative Ophthalmology & Visual Science*, vol. 38, no. 4, pp. 3273, 1997.

[4] R. A. Rensink, "How Much of a Scene is Seen - the Role of Attention in Scene Perception", *Investigative Ophthalmology & Visual Science*, vol. 38, 1997.

[5] D.J. Simons and C.F. Chabris, "Gorillas in Our Midst: Sustained Inattentional Blindness for Dynamic Events," *Perception*, vol. 28, no. 9, pp. 1059-1074, 1999.

[6] J.E. Raymond, K.L. Shapiro, and K.M. Arnell, "Temporary Suppression of Visual Processing in an RSVP Task: An Attentional Blink?" *Journal of exp. psych.*, vol. 18, no 3, pp. 849-60, 1992.

[7] S. Treue and J.H.R. Maunsell, "Attentional Modulation of Visual Motion Processing in Cortical Areas MT and MST," *Nature*, vol 382, pp. 539-541, 1996.

[8] S. Frintrop, E. Rome, and H.I. Christensen, "Computational Visual Attention Systems and Their Cognitive Foundations: A Survey," *ACM Trans. Appl. Percept*, vol.7, no. 1. 2010.

[9] A. Rothenstein and J. Tsotsos, "Attention Links Sensing to Recognition," *Image Vision Comput*, vol. 26, pp. 114-126, 2006.

[10] R. Desimone and J. Duncan, "Neural Mechanisms of Selective Visual Attention," *Annu Rev Neurosci*, vol. 18, pp. 193-222, 1995,

[11] S.J. Luck, L. Chelazzi, S.A. Hillyard, and R. Desimone, "Neural Mechanisms of Spatial Selective Attention in Areas V1, V2, and V4 of Macaque Visual Cortex," *J. Neurophysiol*, vol. 77, 1997.

[12] C. Bundesen and T. Habekost, "Attention," *In Handbook of Cognition, K. Lamberts and R. Goldstone, Eds.*, 2005.

[13] V. Navalpakkam, C. Koch, A. Rangel, and P. Perona, "Optimal Reward Harvesting in Complex Perceptual Environments," *PNAS*, vol. 107, no. 11, pp. 5232-5237, 2010.

[14] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Transactions on PAMI*, vol. 20, no. 11, pp. 1254-1259, 1998.

[15] J.K. Tsotsos, S.M. Culhane, W.Y.K. Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling Visual Attention via Selective Tuning," *Artif. Intell.*, vol. 78, no. 1-2, pp. 507-545, 1995.

[16] R. Milanese, Detecting Salient Regions in an Image: From Biological Evidence to Computer Implementation, Ph.D. thesis, University of Geneva, Switzerland. 1993.

[17] S. Baluja and D. Pomerleau, "Using a Saliency Map for Active Spatial Selective Attention: Implementation & Initial Results," *NIPS*, pp. 451-458, 1994.

[18] C. Koch and S. Ullman, "Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry," *Human Neurobiology*, vol. 4, no. 4, pp. 219-227, 1985.

[19] K. Rayner, "Eye Movements in Reading and Information Processing: 20 Years of Research," *Psychological Bulletin*, 1998.

[20] J. Najemnik and W.S. Geisler, "Optimal Eye Movement Strategies in Visual Search," *Nature*, no. 434, pp. 387-391, 2005.

[21] L.W. Renninger, J.M. Coughlan, P. Verghese, and J. Malik, "An Information Maximization Model of Eye Movements," *NIPS*, vol. 17, pp. 1121-1128, 2005.

[22] U. Rutishauser and C. Koch, "Probabilistic Modeling of Eye Movement Data During Conjunction Search via Feature-based Attention," *Journal of Vision*, vol. 7, no. 6, 2007.

[23] R. Rao, G. Zelinsky, M. Hayhoe, and D. Ballard, "Eye Movements in Iconic Visual Search," *Vision Res.*, vol. 42, 2002.

[24] A.T. Duchowski, "A Breadth-first Survey of Eye-tracking Applications," *Behav. Res. Methods Instrum Comput*, 2002.

[25] G.E. Legge, T.S. Klitz, and B. Tjan, "Mr. Chips: An Ideal-Observer Model of Reading," *Psychological Review*, 1997.

[26] R.D. Rimey and C.M. Brown, "Controlling Eye Movements with Hidden Markov Models," *IJCV*, vol. 7, no. 1, pp. 47-65, 1991.

[27] S. Treue, "Neural Correlates of Attention in Primate Visual Cortex," *Trends in Neurosciences*, vol. 24, no. 5, pp. 295-300, 2001.

[28] S. Kastner and L.G. Ungerleider, "Mechanisms of Visual Attention in the Human Cortex," *Annu. Rev. Neurosci.*, vol. 23, pp. 315-341, 2000.

[29] E.T. Rolls and G. Deco, "Attention in Natural Scenes: Neurophysiological and Computational Bases," *Neural Networks*, vol. 19, no. 9, pp. 1383-1394, 2006.

[30] G.A. Carpenter and S. Grossberg, "A Massively Parallel Architecture for a Self-organizing Neural Pattern Recognition Machine," *Computer Vision, Graphics, and Image Processing*, vol. 37, no. 1, pp. 54-115, 1987.

[31] N. Ouerhani and H. Hügli, "Real-time Visual Attention on a Massively Parallel SIMD Architecture," *Real-Time Imaging*, vol. 9, no. 3, pp. 189-196, 2003.

[32] Q. Ma, L. Zhang, and B. Wang, "New Strategy for Image and Video Quality Assessment," *J. Electronic Imaging*, vol. 19, 2010.

[33] Y. Ma, X. Hua, L. Lu, and H. Zhang, "A Generic Framework of User Attention Model and Its Application in Video Summarization," *IEEE transactions on multimedia*, vol. 7, no. 5, 2005.

[34] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Does Where you Gaze on an Image Affect your Perception of Quality? Applying Visual Attention to Image Quality Metric," *ICIP*, vol. 2, pp. 169-172, 2007.

[35] D. Walther and C. Koch, "Modeling Attention to Salient Proto-objects," *Neural Networks*, vol. 19, no. 9, pp. 1395-1407, 2006.

[36] C. Siagian and L. Itti, "Biologically Inspired Mobile Robot Vision Localization," *IEEE Transactions on Robotics*, 2009.

[37] S. Frintrop and P. Jensfelt, "Attentional Landmarks and Active Gaze Control for Visual SLAM," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1054-1065 , 2008.

[38] D. DeCarlo and A. Santella, "Stylization and Abstraction of Photographs," *ACM Trans. on Graphics*, vol. 21, no. 3, 2002.

[39] L. Itti, "Automatic Foveation for Video Compression Using a Neurobiological Model of Visual Attention," *IEEE Trans. Image Process.*, vol. 13, no. 10, 2004.

[40] L. Marchesotti, C. Cifarelli, and G. Csurka, "A Framework for Visual Saliency Detection with Applications to Image Thumbnailing," *ICCV*, 2009.

[41] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A Coherent Computational Approach to Model Bottom-Up Visual Attention," *IEEE PAMI*, vol. 28, no. 5, pp. 802-817, 2006.

[42] G. Fritz, C. Seifert, L. Paletta and H. Bischof, "Attentive Object Detection Using an Information Theoretic Saliency Measure," *LNCS*, vol. 3368, pp. 29-41, 2005.

[43] T. Liu, J. Sun, N.N Zheng, and H.Y Shum, "Learning to Detect a Salient Object," *CVPR*, 2007.

[44] V. Setlur, R. Raskar, S. Takagi, M. Gleicher, and B. Gooch, "Automatic Image Retargeting, In Mobile and Ubiquitous Multimedia (MUM)," *ACM*, 2005.

[45] C. Chamaret and O. Le Meur, "Attention-based Video reframing: Validation Using Eye-tracking," *ICPR*, 2008.

[46] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-Aware Saliency Detection," *CVPR*, 2010.

[47] N. Sadaka and L.J. Karam, "Efficient Perceptual Attentive Super-resolution," *IEEE Image Processing (ICIP)*, 2009.

[48] H. Liu, S. Jiang, Q. Huang, and C. Xu, "A Generic Virtual Content Insertion System Based on Visual Attention Analysis," *ACM international conference on Multimedia*, pp. 379-388, 2008.

[49] S. Marat, M. Guironnet, and D. Pellerin, "Video Summarization Using a Visual Attention Model," *EUSIPCO*, 2007.

[50] S. Frintrop, VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search, Springer 2006.

[51] V. Navalpakkam and L. Itti, "An Integrated Model of Top-down and Bottom-up Attention for Optimizing Detection Speed," *CVPR*, 2006.

[52] A. Salah, E. Alpaydin, and L. Akrun, "A Selective Attention-based Method for Visual Pattern Recognition with Application to Handwritten Digit Recognition and Face Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 420-425, 2002.

[53] S. Frintrop, "General Object Tracking with a Component-based Target Descriptor," *ICRA*, pp. 4531-4536, 2010.

[54] M.S. El-Nasr, T. Vasilakos, C. Rao, and J. Zupko, "Dynamic Intelligent Lighting for Directing Visual Attention in Interactive 3D Scenes," *IEEE Trans. on Comp. Intell. and AI in Games*, 2009.

[55] G. Boccignone, "Nonparametric Bayesian Attentive Video Analysis," *ICPR*, 2008.

[56] G. Boccignone, A. Chianese, V. Moscato, A. Picariello, "Foveated Shot Detection for Video Segmentation," *IEEE Trans. Circuits Syst. Video Techn.* vol. 15, no. 3, pp. 365-377, 2005.

[57] B., Mertsching, M., Bollmann, R., Hoischen, and S. Schmalz, "The neural active vision system." In Handbook of Computer Vision and Applications, Academic Press, 1999.
[58] A. Dankers, N. Barnes, and A. Zelinsky, "A Reactive Vision System: Active-dynamic Saliency," *ICVS*, 2007.
[59] N. Ouerhani, A. Bur, and H. Hügli, "Visual Attention-based Robot Self-localization," *ECMR*, pp. 813, 2005.
[60] S. Baluja, and D. Pomerleau, "Expectation-based Selective Attention for Visual Monitoring and Control of a Robot Vehicle," *Rob. Auton. Syst.*, vol. 22, no. 3-4, pp. 329-344, 1997.
[61] C. Scheier and S. Egner, "Visual Attention in A Mobile Robot," *International Symposium on Industrial Electronics*, pp. 48-53, 1997.
[62] C. Breazeal, "A Context-dependent Attention System for a Social Robot," *IJCAI*, pp. 1146-1151, 1999.
[63] G. Heidemann, R. Rae, H. Bekel, I. Bax, and H. Ritter, "Integrating Context-free and Context-dependent Attentional Mechanisms for Gestural Object Reference," *Mach. Vision Appl.*, vol. 16, no. 1, pp. 64-73, 2004.
[64] G. Heidemann, "Focus-of-attention from Local Color Symmetries," *IEEE Trans PAMI*, vol. 26, no. 7, pp. 817-830, 2004.
[65] A. Belardinelli, Salience Features Selection: Deriving a Model from Human Evidence, Ph.D. thesis, Italy, 2008.
[66] Y. Nagai, "From Bottom-up Visual Attention to Robot Action Learning," *ICDL*, 2009.
[67] C., Muhl, Y., Nagai, and G. Sagerer, "On constructing a communicative space in HRI." *Proceedings of the 30th German Conference on Artificial Intelligence, Springer*, 2007.
[68] T. Liu, S.D. Slotnick, J.T. Serences, and S. Yantis, "Cortical Mechanisms of Feature-based Intentional Control," *Cerebral Cortex*, vol. 13, no. 12, 2003.
[69] B.W. Hong and M. Brady, "A Topographic Representation for Mammogram Segmentation," *LNCS*, vol. 2879, 2003.
[70] N. Parikh, L. Itti, and J. Weiland, "Saliency-based Image Processing for Retinal Prostheses," *Journal of Neural Engineering*, vol 7, no 1, 2010.
[71] O.R. Joubert, D. Fize, G.A. Rousselet, and M. Fabre-Thorpe, "Early Interference of Context Congruence on Object Processing in Rapid Visual Categorization of Natural Scenes," *Journal of Vision*, vol. 8, no. 13, 2008.
[72] H. Li and K.N. Ngan, "Saliency Model-based Face Segmentation and Tracking in Head-and-shoulder Video Sequences," *J. Vis. Commun. Image R.*, vol. 19, pp. 320-333, 2008.
[73] N. Courty and E. Marchand, "Visual Perception based on Salient Features," *IROS*, 2003.
[74] F. Shic and B. Scassellati, "A Behavioral Analysis of Computational Models of Visual Attention," *IJCV*, vol. 73, 2007.
[75] H.C. Nothdurft, "Salience of Feature Contrast," *In Neurobiology of Attention, L. Itti, G. Rees, and J. K. Tsotsos, Eds.*, 2005.
[76] M. Corbetta, and G. L. Shulman, "Control of Goal-directed and Stimulus-driven Attention in the Brain," *Nat. Rev.*, vol. 3, no. 3, pp. 201-215, 2002.
[77] L. Itti and C. Koch, "Computational Modeling of Visual Attention," *Nat. Rev. Neurosci.*, vol. 2, no. 3, pp. 194-203, 2001.
[78] H.E. Egeth and S. Yantis, "Visual Attention: Control, Representation, and Time Course," *Ann. Rev. Psych.*, vol. 48, 1997.
[79] A.L. Yarbus, Eye-Movements and Vision, Plenum Press, New York, 1967.
[80] V., Navalpakkam and L. Itti, "Modeling the Influence of Task on Attention." *Vision Res.*, vol. 45, no. 2, pp. 205-231, 2005.
[81] A.M. Treisman and G. Gelade, "A Feature Integration Theory of Attention," *Cognitive Psych.*, vol. 12, pp. 97-136, 1980.
[82] J.M. Wolfe, "Guided Search 4.0: Current Progress with a Model of Visual Search," *In Integrated Models of Cognitive Systems, W. D. Gray, Ed. Oxford University Press, Oxford, UK*, 2007.
[83] G.J. Zelinsky, "A Theory of Eye Movements During Target Acquisition," *Psychological Review*, vol. 115, no. 4, 787-835, 2008.
[84] W. Einhauser, M. Spain, and P. Perona, "Objects Predict Fixations Better Than Early Saliency," *Journal of Vision*, 2008.
[85] M. Pomplun, "Saccadic Selectivity in Complex Visual Search Displays," *Vision Research*, vol. 46, pp. 1886-1900, 2006.
[86] A. Hwang and M. Pomplun, "A Model of Top-down Control of Attention During Visual Search in Real-world Scenes," *Journal of vision*, vol. 8, no. 6, pp. 2008.
[87] K. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva, "Modeling Search for People in 900 Scenes: A Combined Source Model of Eye Guidance," *Visual Cognition*, vol. 17, 2009.
[88] A. Borji, M.N. Ahmadabadi, B. N. Araabi, and M. Hamidi, "Online Learning of Task-driven Object-based Visual Attention Control," *Image. Vision Comput.*, vol. 28, pp. 1130-1145, 2010.
[89] A. Borji, M.N. Ahmadabadi, and B. N. Araabi, "Cost-sensitive Learning of Top-down Modulation for Attentional Control," *Machine Vision and Applications*, vol. 22, 2011.
[90] L. Elazary and L. Itti, "A Bayesian Model for Efficient Visual Search and Recognition," *Vision Research*, vol. 50, 2010.
[91] M.M. Chun and Y. Jiang, "Contextual Cueing: Implicit Learning and Memory of Visual Context Guides Spatial Attention," *Cognitive Psychology*, vol. 36, pp. 28-71, 1998.
[92] A. Torralba, "Modeling Global Scene Factors in Attention," *Journal of Optical Society of America*, vol. 20, no. 7, 2003.
[93] A. Oliva and A. Torralba, "Modeling the Shape of the Scene: a Holistic Representation of the Spatial Envelope," *International Journal in Computer Vision*, vol. 42, pp. 145-175, 2001.
[94] L.W. Renninger and J. Malik, "When is Scene Recognition Just Texture Recognition?" *Vis. Res.*, vol. 44, pp. 2301-2311, 2004.
[95] C. Siagian and L. Itti, "Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention," *IEEE PAMI*, vol. 29, no. 2, pp. 300-312, 2007.
[96] M. Viswanathan, C. Siagian, and L. Itti, Vision Science Symposium (VSS), 2007.
[97] J. Triesch, D.H. Ballard, M.M. Hayhoe, and B.T. Sullivan, "What You See Is What You Need," *Journal of Vision*, 2003.
[98] M.I. Posner, Orienting of Attention, Q. J. Exp. Psych. vol. 32, pp. 3-25, 1980.
[99] M. Hayhoe and D. Ballard, "Eye Movements in Natural Behavior," *Trends in Cognitive Sciences*, vol. 9, pp. 188-194, 2005.
[100] M.S. Mirian, M. N. Ahmadabadi, B.N. Araabi, R. R. Siegwart, "Learning Active Fusion of Multiple Experts' Decisions: An Attention-based Approach," Neural Computation, 2011.
[101] R.J. Peters and L. Itti, "Beyond Bottom-up: Incorporating Task-dependent Influences Into a Computational Model of Spatial Attention," *CVPR*, 2007.
[102] D. Pang, A. Kimura, T. Takeuchi, J. Yamato, and K. Kashino, "A Stochastic Model of Selective Visual Attention with a Dynamic Bayesian Network," *ICME*, 2008.
[103] Y. Zhai and M. Shah, "Visual Attention Detection in Video Sequences Using Spatiotemporal Cues',' *ACM International Conference on Multimedia*, 2006.
[104] S. Marat, T. Ho-Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué, "Modeling Spatio-temporal Saliency to Predict Gaze Direction for Short Videos," *IJCV*, 2009.
[105] V. Mahadevan and N. Vasconcelos, "Spatiotemporal Saliency in Dynamic Scenes." *IEEE PAMI*, vol. 32, no. 1, 2010.
[106] V. Mahadevan and N. Vasconcelos, "Saliency Based Discriminant Tracking." *In IEEE (CVPR)*, 2009.
[107] N. Jacobson, Y-L. Lee, V. Mahadevan, N. Vasconcelos and T.Q. Nguyen, "A Novel Approach to FRUC using Discriminant Saliency and Frame Segmentation." *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2924-2934, 2010.
[108] H.J. Seo and P. Milanfar, "Static and Space-time Visual Saliency Detection by Self-Resemblance," *Journal of Vision*, vol. 9, no. 12, pp. 1-27, 2009.
[109] N. Sprague and D.H. Ballard, "Eye Movements for Reward Maximization," *NIPS*, 2003.
[110] http://tcts.fpms.ac.be/ mousetrack/
[111] J. Bisley and M. Goldberg, "Neuronal Activity in The Lateral Intraparietal Area and Spatial Attention," *Science*, 2003.
[112] J. Duncan, "Selective Attention and The Organization of Visual Information," *J. Exp. Psych.*, vol. 113, pp. 501-517, 1984.
[113] B.J. Scholl, "Objects and Attention: The State of The Art," *Cognition*, vol. 80, pp. 1-46. 2001.
[114] Z. W. Pylyshyn and R. W. Storm, "Tracking Multiple Independent Targets: Evidence for a Parallel Tracking Mechanism," *Spatial Vision*, vol. 3, pp. 179-197, 1988.
[115] E. Awh and H. Pashler, "Evidence For Split Attentional Foci," *J. Exp.Psych. Hum. Percept. Perform.*, vol. 26, pp. 834-846, 2000.
[116] B.C. Russell, A. Torralba, K.P. Murphy, and W.T. Freeman, "LabelMe: A Database and Web-based Tool for Image Annotation," *IJCV*, vol. 77, no. 1-3, pp. 157-173, 2008.
[117] Y. Sun and R. Fisher, "Object-based Visual Attention for Computer Vision," *Artif. Intell.*, vol. 146, no. 1, pp. 77-123, 2003.
[118] J.M. Wolfe and T.S. Horowitz, "What Attributes Guide the Deployment of Visual Attention and How Do They Do It?" *Nat. Rev. Neurosci.*, vol. 5, pp. 1-7, 2004.

[119] L. Itti, N. Dhavale, and F. Pighin, "Realistic Avatar Eye and Head Animation Using a Neurobiological Model of Visual Attention," *SPIE*, vol. 5200, pp. 64-78, 2003.

[120] R. Rae, Gestikbasierte Mensch-Maschine-Kommunikation auf der Grundlage visueller Aufmerksamkeit und Adaptivität. Ph.D. thesis, Universität Bielefeld, Germany, 2000.

[121] J. Harel, C. Koch, and P. Perona, "Graph-based Visual Saliency," *NIPS*, vol. 19, pp. 545-552, 2006.

[122] O. Boiman and M. Irani, "Detecting Irregularities in Images and in Video," *ICCV*, 2005.

[123] B.W. Tatler, "The Central Fixation Bias in Scene Viewing: Selecting an Optimal Viewing Position Independently of Motor Bases and Image Feature Distributions," *J. Vision*, 2007.

[124] R. Milanese, Detecting Salient Regions in an Image: From Biological Evidence to Computer Implementation, Ph.D. thesis, University of Geneva, Switzerland, 1993.

[125] F.H. Hamker, "The Emergence of Attention by Poulation-based Inference and Its Role in Distributed Processing and Cognitive Control of Vision," *J. Comput. Vision Image Understanding*, vol. 100, no. 1-2, pp. 64106. 2005.

[126] S. Vijayakumar, J. Conradt, T. Shibata, and S. Schaal, "Overt Visual Attention For a Humanoid Robot," *IROS*, 2001.

[127] C.M. Privitera and L.W. Stark, "Algorithms for Defining Visual Regions-of-Interest: Comparison with Eye Fixations," *IEEE PAMI*, vol. 22, no. 9, pp. 970-982, 2000.

[128] K. Lee, H. Buxton, and J. Feng, "Selective Attention for Cue-guided Search Using a Spiking Neural Network," *WAPCV*, pp. 5562, 2003.

[129] T. Kadir and M. Brady, "Saliency, Scale and Image Description," *Int. J. Comput. Vision*, vol. 45, no. 2, pp. 83-105, 2001.

[130] A. Maki, P. Nordlund, and J.O. Eklundh, "Attentional Scene Segmentation: Integrating Depth and Motion," *Comput. Vision Image Understanding*, vol. 78, no. 3, pp. 351-373, 2000.

[131] D. Parkhurst, K. Law, and E. Niebur, "Modeling the Role of Salience in The Allocation of Overt Visual Attention," *Vision Res*. vol. 42, no. 1, pp. 107-123, 2002.

[132] T.S. Horowitz, and J.M. Wolfe, "Visual Search Has No Memory," *Nature*, vol. 394, pp. 575-577, 1998.

[133] J. Li, Y. Tian, T. Huang, and W. Gao, "Probabilistic Multi-Task Learning for Visual Saliency Estimation in Video," *IJCV*, 2010.

[134] R. Peters, A. Iyer, L. Itti, and C. Koch, "Components of Bottom-up Gaze Allocation in Natural Images," *Vis. Res.*, 2005.

[135] M. Land and M. Hayhoe, "In What Ways Do Eye Movements Contribute to Everyday Activities?" *Vis. Res.*, vol. 41, 2001.

[136] G. Kootstra, A. Nederveen, and B. de Boer, "Paying Attention to Symmetry," *BMVC*, pp. 1115-1125, 2008.

[137] D. Reisfeld, H. Wolfson, and Y. Yeshurun, "Context-free Attentional Operators: The Generalized Symmetry Transform," *Int. Journal of Computer Vision*, vol. 14, no. 2, pp. 119-130, 1995.

[138] O. Le Meur, P. Le Callet and D. Barba, "Predicting Visual Fixations on Video Based on Low-level Visual Features," *Vision Research*, vol. 47/19, pp. 2483-2498, 2007.

[139] D.D. Salvucci, "An Integrated Model of Eye Movements and Visual Encoding," *Cognitive Systems Research*, vol. 1, 2001.

[140] A. Oliva, A. Torralba, M.S. Castelhano, and J.M. Henderson, "Top-down Control of Visual Attention in Object Detection," *ICIP*, pp. 253-256, 2003.

[141] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian Framework for Saliency Using Natural Statistics," *J. of Vision*, vol. 8(7), no. 32, pp. 1-20, 2008.

[142] L. Zhang, M.H. Tong, and G.W. Cottrell, "SUNDAy: Saliency Using Natural Statistics for Dynamic Analysis of Scenes," *In Thirty-first Annual Cognitive Science Society Conference*, 2009.

[143] N.D.B. Bruce and J.K. Tsotsos, "Spatiotemporal Saliency: Towards a Hierarchical Representation of Visual Saliency," WAPCV, 2008.

[144] N.D.B. Bruce, J.K. Tsotsos, "Saliency Based on Information Maximization," *NIPS*, 2005.

[145] L. Itti and P. Baldi, "Bayesian Surprise Attracts Human Attention," *NIPS*, 2005.

[146] D. Gao and N. Vasconcelos, "Discriminant Saliency for Visual Recognition from Cluttered Scenes," *NIPS*, 2004.

[147] D. Gao, S. Han and N. Vasconcelos, "Discriminant Saliency, the Detection of Suspicious Coincidences, and Applications to Visual Recognition." IEEE Trans. PAMI. vol. 31, no. 6, 2009.

[148] E. Gu, J. Wang, and N.I. Badler, "Generating Sequence of Eye Fixations Using Decision-Theoretic Attention Model," *WAPCV*, pp. 277-29, 2007.

[149] T.S. Lee and S. Yu, "An Information-theoretic Framework for Understanding Saccadic Behaviors," *NIPS*, 2000.

[150] X. Hou and L. Zhang. "Saliency Detection: A Spectral Residual Approach," *CVPR*, 2007.

[151] X. Hou and L. Zhang. "Dynamic Visual Attention: Searching for Coding Length Increments," *NIPS*, 2008.

[152] M. Mancas, Computational Attention: Modelisation and Application to Audio and Image Processing, PhD. thesis, 2007.

[153] T. Avraham, M. Lindenbaum, "Esaliency (Extended Saliency): Meaningful Attention Using Stochastic Image Modeling," *IEEE PAMI*, vol. 32, no. 4, pp. 693-708, 2010.

[154] S. Chikkerur, T. Serre, C. Tan, and T. Poggio, "What and Where: A Bayesian Inference Theory of Visual Attention," Vision Research, 2010.

[155] P. Verghese, "Visual Search and Attention: A Signal Detection Theory Approach," *Neuron*, vol. 31, pp. 523-535, 2001.

[156] C. Guo, Q. Ma, and L. Zhang, "Spatio-Temporal Saliency Detection Using Phase Spectrum of Quaternion Fourier Transform," *CVPR*, 2008.

[157] C. Guo and L. Zhang, "A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 185-198, 2010.

[158] R. Achanta, S.S. Hemami, F.J. Estrada, and S. Süsstrunk, "Frequency-tuned Salient Region Detection," *CVPR*, 2009.

[159] P. Bian and L. Zhang, "Biological Plausibility of Spectral Domain Approach for Spatiotemporal Visual Saliency," *LNCS*, vol. 5506, pp. 251-258, 2009.

[160] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil, "Decorrelation and Distinctiveness Provide With Human-Like Saliency," *ACIVS*, vol. 5807, 2009.

[161] N.J. Butko and J.R. Movellan, "Optimal Scanning for Faster Object Detection," *CVPR*, 2009.

[162] S. Jodogne and J. Piater, "Closed-Loop Learning of Visual Control Policies," *Journal of Artificial Intelligence Research*, vol. 28, pp. 349-391, 2007.

[163] R. McCallum, Reinforcement Learning with Selective Perception and Hidden State, Ph.D thesis, 1996.

[164] L. Paletta, G. Fritz, and C. Seifert, "Q-learning of Sequential Attention for Visual Object Recognition From Informative Local Descriptors," *ICML*, pp. 649-656, 2005.

[165] W. Kienzle, M.O. Franz, B. Schölkopf, and F.A. Wichmann, "Center-surround Patterns Emerge as Optimal Predictors for Human Saccade Targets," *Journal of Vision*, vol. 9, 2009.

[166] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to Predict Where Humans Look," *ICCV*, 2009.

[167] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting Human Gaze Using Low-level Saliency Combined With Face Detection," *NIPS*, 2007.

[168] O. Ramström and H.I. Christensen, "Visual Attention Using Game Theory," *Biologically Motivated Computer Vision Conference*, pp. 462-471, 2002.

[169] P.L. Rosin, "A Simple Method for Detecting Salient Regions," *Pattern Recognition*, vol. 42, no. 11, pp. 2363-2371, 2009.

[170] Z. Li, "A Saliency Map in Primary Visual Cortex," *Trends in Cognitive Sciences*, vol. 6, no. 1, pp. 9-16, 2002.

[171] Y. Li, Y. Zhou, J. Yan, and J. Yang, "Visual Saliency Based on Conditional Entropy," *ACCV*, 2009.

[172] S.W Ban, I. Lee, and M. Lee, "Dynamic Visual Selective Attention Model," *Neurocomputing*, vol. 71, no. 4-6, 2008.

[173] M. T. López, M. A. Fernńdez, A. Fernández-Caballero, J. Mira, A.E. Delgado, "Dynamic Visual Attention Model in Image Sequences," *Image and Vision Computing*, vol. 25, 2007.

[174] U. Rajashekar, I. van der Linde, A.C. Bovik, and L. K., Cormack, "GAFFE: A Gaze-Attentive Fixation Finding Engine," *IEEE Trans. on Image Processing*, vol. 17, no. 4, pp. 564-573, 2008.

[175] G. Boccignone and M. Ferraro, "Modeling gaze shift as a constrained random walk", Physica A, vol. 331, 2004.

[176] M. C. potter, "Meaning in Visual Scenes," *Science*, vol. 187, pp. 965-966, 1975.

[177] J. M. Henderson and A. Hollingworth, "High-level Scene Perception," *Ann. Rev. of Psychology*, vol. 50, pp. 243-271, 1999.

[178] R.A. Rensink, "The Dynamic Representation of Scenes," *Visual Cognition*, vol. 7, pp. 17-42, 2000.

[179] J. Bailenson and N. Yee, "Digital Chameleons: Automatic Assimilation of Nonverbal Gestures in Immersive Virtual Environments," *Psychological Science*, vol. 16, pp. 814-819, 2005.
[180] M. Sodhi , B. Reimer, J.L. Cohen, E. Vastenburg, R. Kaars, and S. Kirschenbaum, "On-road Driver Eye Movement Tracking Using Head-mounted Devices," *Proceedings of the Symposium on Eye tracking Research & Applications*, 2002.
[181] J.H. Reynolds and D.J. Heeger, "The Normalization Model of Attention," *Neuron*, vol. 61, no. 2, pp. 168-185, 2009.
[182] S. Engmann, B.M. Hart, T. Sieren, S. Onat, P. König, and W. Einhäuser, "Saliency on a Natural Scene Background: Effects of Color and Luminance Contrast Add Linearly," *Attention, Perception & Psychophysics*, vol. 71, no. 6, pp. 1337-1352, 2009.
[183] A. Reeves and G. Sperling, "Attention Gating in Short-term Visual Memory," *Psych. Review*, vol. 93, no. 2, pp. 180-206, 1986.
[184] L. Itti, "Quantifying the Contribution of Low-Level Saliency to Human Eye Movements in Dynamic Scenes," *Visual Cognition*, vol. 12, no. 6, pp. 1093-1123, 2005.
[185] D. Gao, V. Mahadevan and N. Vasconcelos, "On the Plausibility of the Discriminant Center-surround Hypothesis for Visual Saliency," *Journal of Vision*, vol. 8, no. (7):13, pp. 1-18, 2008.
[186] J. Yan, J. Liu, Y. Li, and Y. Liu, "Visual Saliency via Sparsity Rank Decomposition," *ICIP*, 2010.
[187] http://www.its.caltech.edu/∼xhou/
[188] J. Yuen, B.C. Russell, C. Liu, and A. Torralba, "LabelMe Video: Building a Video Database with Human Annotations," *ICCV*, 2009.
[189] R. Rosenholtz, Y. Li, and L. Nakano, Measuring Visual Clutter. Journal of Vision, vol. 7(2), no. 17, pp. 1-22, 2007.
[190] R. Rosenholtz, A. Dorai, and R. Freeman, "Do Predictions of Visual Perception Aid Design?" *ACM Transactions on Applied Perception (TAP)*, vol. 8, no. 2, 2011.
[191] R. Rosenholtz, "A Simple Saliency Model Predicts a Number of Motion Popout Phenomena," *Vis. Res.*, vol. 39, 1999.
[192] X. Hou, J. Harel, and Christof Koch, "Image Signature: Highlighting sparse salient regions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.
[193] R. Rosenholtz, A.L. Nagy, and N. R. Bell, "The Effect of Background Color on Asymmetries in Color Search", *Journal of Vision*, vol. 4, no. 3, pp. 224-240, 2004.
[194] http://alpern.mit.edu/saliency/
[195] D. Green and J. Swets, Signal Detection Theory and Psychophysics, New York: John Wiley, 1966.
[196] T. Jost, N. Ouerhani, R. von Wartburg, R. Mäuri, and H. Häugli, "Assessing the Contribution of Color in Visual Attention," *Computer Vision and Image Understanding*, vol. 100, 2005.
[197] U. Rajashekar, A.C. Bovik, and L.K. Cormack, "Visual Search in Noise: Revealing the Influence of Structural Cues by Gaze-contingent Classification Image Analysis," *J. Vision*, 2006.
[198] S.A. Brandt and L.W. Stark, Spontaneous Eye Movements during Visual Imagery Reflect the Content of the Visual Scene," *Journal of Cognitive Neuroscience*, vol. 9, no. 27-38,1997.
[199] A.D. Hwang, H.C. Wang, and M. Pomplun, "Semantic Guidance of Eye Movements in Real-world Scenes," *Vis. Res*, 2011.
[200] N. Murray, M. Vanrell, X. Otazu, and C. Alejandro Parraga, "Saliency Estimation Using a Non-Parametric Low-Level Vision Model," *CVPR*, 2011.
[201] W. Wang, C. Chen, Y. Wang, T. Jiang, F. Fang, and Y. Yao, "Simulating Human Saccadic Scanpaths on Natural Images," *CVPR*, 2011.
[202] R.L Canosa, "Real-World Vision: Selective Perception and Task," *ACM Transactions on Applied Perception*, vol. 6, no. 2, 2009.
[203] M.S. Peterson, A.F. Kramer, and D.E. Irwin, "Covert Shifts of Attention Precede Involuntary Eye Movements," Perception & Psychophysics, vol. 66, pp. 398-405, 2004.
[204] F. Baluch and L. Itti, "Mechanisms of Top-Down Attention," *Trends in Neuroscience*, vol. 34, no. 4, 2011.
[205] J. Hayes and A. Efros, "Scene Completion Using Millions of Photographs", *SIGGRAPH*, 2007.
[206] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part Based Models," *IEEE Trans. PAMI*, vol. 32, no. 9, 2010.
[207] A.K. Mishra and Y. Aloimonos, "Active Segmentation," *International Journal of Humanoid Robotics*, vol 6., pp 361-386, 2009.
[208] B. Suh, H. Lingm, B.B. Bederson, and D.W. Jacobs. "Automatic Thumbnail Cropping and Its Effectiveness," In *UIST*, pp. 95-104, 2003
[209] S. Mitri, S. Frintrop, K. Pervolz, H. Surmann, and A. Nuchter, "Robust Object Detection at Regions of Interest with an Application in Ball Recognition," *ICRA*, pp. 126-131, 2005.
[210] N. Ouerhani, R. von Wartburg, H. Hugli, and R.M. Muri, "Empirical Validation of Saliency-based Model of Visual Attention," *Electronic Letters on Computer Vision and Image Analysis*, vol. 3, no. 1, pp. 13-24, 2003.
[211] L.W. Stark and Y. Choi, "Experimental Metaphysics: The Scanpath as an Epistemological Mechanism," *Visual Attention and Cognition.* pp. 3-69, 1996.
[212] P. Reinagel and A. Zador, "Natural Scenes at the Center of Gaze," *Network*, vol. 10, pp. 341-50, 1999.
[213] U. Engelke, H.J. Zepernick, and A. Maeder, "Visual Attention Modeling: Region-of-interest Versus Fixation Patterns", *Picture Coding Symposium (PCS)*, 2009.
[214] M. Verma and P.W. McOwana, "Generating Customised Experimental Stimuli for Visual Search Using Genetic Algorithms Shows Evidence For a Continuum of Search Efficiency," *Vision Research*, vol. 49, no. 3, pp. 374-382, 2009.
[215] S. Han and N. Vasconcelos, "Biologically Plausible Saliency Mechanisms Improve Feedforward Object Recognition", *Vision Research*, vol. 50, no. 22, pp. 2295-2307, 2010.
[216] D. Ballard, M. Hayhoe, J. Pelz, "Memory Representations in Natural Tasks," *J. of Cognitive Neuroscience*, vol. 7, no. 1, 1995.
[217] R. Rao, "Bayesian Inference and Attentional Modulation in the Visual Cortex," *NeuroReport*, vol. 16, no. 16, 2005.
[218] A.Borji, D.N. Sihite, and L. Itti, "Computational Modeling of Top-down Visual Attention in Interactive Environments," *BMVC*, 2011.
[219] E. Niebur and C.Koch, "Control of Selective Visual Attention: Modeling the Where Pathway," *NIPS*, pp. 802-808, 1995.
[220] P. Viola and M.J. Jones, "Robust Real-Time Face Detection," *IJCV*, vol. 57, no. 2, pp. 137-154, 2004.
[221] W. Kienzle, B. Schø"lkopf, F.A. Wichmann, M.O. Franz, "How to Find Interesting Locations in Video: A Spatiotemporal Interest Point Detector Learned from Human Eye Movements," *DAGM-Symposium*, pp. 405-414, 2007.
[222] J. Wang, J. Sun, L. Quan, X. Tang, and H.Y Shum, "Picture Collage," CVPR, 2006.
[223] D. Gao and N. Vasconcelos, "Decision-theoretic saliency: computational principles, biological plausibility, and implications for neurophysiology and psychophysics," *Neural Computation*, vol. 21, pp. 239-271, 2009.
[224] M. Carrasco, "Visual attention: The past 25 years," *Vision Research*, vol. 51, pp. 1484-1525, 2011.

Ali Borji received the BS and MS degrees in computer engineering from Petroleum University of Technology, Tehran, Iran, 2001 and Shiraz University, Shiraz, Iran, 2004, respectively. He did his Ph.D. in cognitive neurosciences at Institute for Studies in Fundamental Sciences (IPM) in Tehran, Iran, 2009. He is currently a postdoctoral scholar at iLab, University of Southern California, Los Angeles, CA. His research interests include: visual attention, visual search, machine learning, robotics, neurosciences and biologically plausible vision models.

Laurent Itti received his M.S. degree in Image Processing from the Ecole Nationale Superiere des Télecommunications in Paris in 1994 and his Ph.D. in Computation and Neural Systems from Caltech in 2000. He is now and associate professor of Computer Science, Psychology and Neurosciences at the University of Southern California. Dr Itti's research interests are in biologically-inspired computational vision, in particular in the domains of visual attention, gist, saliency, and surprise, with technological applications to video compression, target detection, and robotics.