Visual Attention

- Frintrop, Rome, Christensen, “Computational Visual Attention Systems and Their Cognitive Foundations: A Survey,” *ACM Transactions on Applied Perception*, Vol. 7, No. 1, Article 6, 2010. (**required**)
- Itti, L., Koch, C. “Computational modeling of visual attention,” *Nat. Rev. Neurosci.* 2, 3, 194–203, 2001. (**optional**)
- Liu et al., “Learning to detect a salient object,” *IEEE Trans PAMI*, 2010. (**optional**)

Computational Visual Attention Systems and Their Cognitive Foundations: A Survey

SIMONE FRINTROP
Rheinische Friedrich-Wilhelms-Universität
ERICH ROME
Fraunhofer Institute for Intelligent Analysis and Information Systems
and
HENRIK I. CHRISTENSEN
Georgia Tech

Based on concepts of the human visual system, computational visual attention systems aim to detect regions of interest in images. Psychologists, neurobiologists, and computer scientists have investigated visual attention thoroughly during the last decades and profited considerably from each other. However, the interdisciplinarity of the topic holds not only benefits but also difficulties: Concepts of other fields are usually hard to access due to differences in vocabulary and lack of knowledge of the relevant literature. This article aims to bridge this gap and bring together concepts and ideas from the different research areas. It provides an extensive survey of the grounding psychological and biological research on visual attention as well as the current state of the art of computational systems. Furthermore, it presents a broad range of applications of computational attention systems in fields like computer vision, cognitive systems, and mobile robotics. We conclude with a discussion on the limitations and open questions in the field.

Categories and Subject Descriptors: A.1 [**Introductory and Survey**]: I.2.10 [**Vision and Scene Understanding**]: I.4 [**Image Processing and Computer Vision**]: I.6.5 [**Model Development**]: I.2.9 [**Robotics**]:

General Terms: Algorithms, Design

Additional Key Words and Phrases: Visual attention, saliency, regions of interest, biologically motivated computer vision, robot vision

ACM Reference Format:

Frintrop, S., Rome, E., and Christensen, H. I. 2010. Computational visual attention systems and their cognitive foundations: A survey ACM Trans. Appl. Percept. 7, 1, Article 6 (January 2010), 39 pages.
DOI = 10.1145/1658349.1658355 http://doi.acm.org/10.1145/1658349.1658355

Authors' addresses: S. Frintrop, Institute of Computer Science III, Rheinische Friedrich-Wilhelms-Universität, Römerstr. 164, 53117 Bonn, Germany, email: frintrop@iai.uni-bonn.de. E. Rome, Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS), Schloss Birlinghoven, 53757 Sankt Augustin, Germany, email: erich.rome@iais.fraunhofer.de. H.I. Christensen, Georgia Tech, College of Computing, 85 Fifth Street, Atlanta, GA, 30308, USA, email: hic@cc.gatech.edu.

DOI 10.1145//1658349.1658355 http://doi.acm.org/10.1145//1658349.1658355

1. INTRODUCTION

Every stage director is aware of the concepts of human selective attention and knows how to exploit them to manipulate his audience: A sudden spotlight illuminating a person in the dark, a motionless character starting to move suddenly, a voice from a character hidden in the audience, these effects not only keep our interest alive, but they also guide our gaze, telling where the current action takes place. The mechanism in the brain that determines which part of the multitude of sensory data is currently of most interest is called *selective attention*. This concept exists for each of our senses; for example, the cocktail party effect is well known in the field of auditory attention. Although a room may be full of different voices and sounds, it is possible to voluntarily concentrate on the voice of a certain person [Cherry 1953]. Visual attention is sometimes compared with a spotlight in a dark room. The fovea—the center of the retina—has the highest resolution in the eye. Thus, directing the gaze to a certain region complies with directing a spotlight to a certain part of a dark room [Shulman et al. 1979]. By moving the spotlight around, one can obtain an impression of the contents of the room, while analogously, by scanning a scene with quick eye movements, one can obtain a detailed impression of it.

Evolution has favored the concepts of selective attention because of the human need to deal with a high amount of sensory input at each moment. This amount of data is, in general, too high to be completely processed in detail and the possible actions at one and the same time are restricted; the brain has to prioritize. The same problem is faced by many modern technical systems. Computer vision systems have to deal with thousands, sometimes millions, of pixel values from each frame and the computational complexity of many problems related to the interpretation of image data is very high [Tsotsos 1987]. The task becomes especially difficult if a system has to operate in real time. Application areas in which real-time performance is essential are cognitive systems and mobile robotics, since the systems have to react to their environment instantaneously.

For mobile autonomous robots, focusing on the relevant data is even more important than for pure vision systems. Many modules have to share resources on a robot. Usually, different modules share a visual sensor and each module has its own requirements. An obstacle-avoidance module requires access to peripheral data to generate a motion flow, whereas a recognition module requires high-resolution central data. Such a module might profit from zooming to the object, other modules might require gaze shifts. These resource conflicts depend on a selection mechanism, which controls and prioritizes possible actions. Furthermore, cameras are often used in combination with other sensors, and modules concerned with tasks such as navigation and manipulation of objects require additional computation power. And in contrast to early robotic systems applied in simple industrial conveyor belt tasks, current systems are supposed to drive and act autonomously in complex, previously unknown environments with challenges such as changing illuminations and people that walk around. Thus, for humans as well as for robots, limited resources require a selection mechanism that prioritizes the sensory input from "very important" to "not useful right now."

In order to cope with these requirements, people have investigated how the concepts of human selective attention can be exploited for computational systems. For many years, these investigations have been of mainly theoretical interest, since the computational demands were too high for practical applications. Only during the last 5 to 10 years, the computational power enabled implementations of computational attention systems that are useful in practical applications, causing an increasing interest in such mechanisms in fields like computer vision, cognitive systems, and mobile robotics. Example applications include object recognition, robot localization, or human–robot interaction.

In this article, we provide a survey of computational visual attention systems and their applications. This article is intended to bridge the gap between communities. For researchers from engineering sciences interested in computational attention systems, it provides the necessary psychophysical and

neuroscientific background knowledge about human visual attention. For psychologists and neurobiologists, it explains the techniques applied to build computational attention systems. And for all researchers concerned with visual attention, it provides an overview of the current state of the art and of applications in computer vision and robotics.

This work focuses on systems that are both biologically motivated and serve a technical purpose. Such systems aim to improve computational vision systems in speed and/or quality of detection and recognition. Other computational attention systems focus on the objective to basically simulate and understand the concepts of human visual attention. A brief overview is given in Section 2.3, but for a more thorough exposition, the authors point the interested reader to the following review papers. A review of computational attention systems with a psychological objective can be found in Heinke and Humphreys [2004], and a survey on computational attention models significantly inspired from neurobiology and psychophysics is presented by Rothenstein and Tsotsos [2006a]. Finally, a broad review on psychological attention models, in general, is found in Bundesen and Habekost [2005].

Since the term "attention" is not clearly defined, it is sometimes used in other contexts. In the broadest sense, any preprocessing method might be called attentional, because it focuses subsequent processing to parts of the data that seem to be relevant. For example, Viola and Jones [2004] present an object recognition technique which they call "attentional cascade," since it starts processing at a coarse level and intensifies processing only at interesting regions. In this article, we focus on approaches that are motivated by human visual attention (see Section 2.2 for a definition).

The structure of the article is as follows. In Section 2, we introduce the concepts of human visual attention and present the psychological theories and models that have been most influential for computational attention systems. Section 3 describes the general structure and characteristics of computational attention systems and provides an overview over the state of the art in this field. Applications of visual attention systems in computer vision and robotics are described in Section 4. A discussion on the limitations and open questions in the field concludes the article.

2. HUMAN VISUAL ATTENTION

This section introduces background knowledge on human visual attention that researchers should have when dealing with computational visual attention. We start by briefly sketching the human visual system in Section 2.1. After that, Section 2.2 introduces the concepts of visual attention. Finally, in Section 2.3, we present the most important psychological theories and models of visual attention that form the basis for most current computational systems.

2.1 The Human Visual System

Here, we start with providing a very rough overview of the human visual system (see Figure 1). Further literature on this topic can be found in Palmer [1999], Kandel et al. [1996], and Zeki [1993].

The light that arrives at the eye is projected onto the retina and then the visual information is transmitted via the optic nerve to the optic chiasm. From there, two pathways go to each brain hemisphere: the collicular pathway leading to the superior colliculus (SC) and, more important, the retino-geniculate pathway, which transmits about 90% of the visual information and leads to the lateral geniculate nucleus (LGN). From the LGN, the information is transferred to the primary visual cortex (V1). Up to here, the processing stream is also called the *primary visual pathway*. Many simple feature computations take part during this pathway. Already in the retina, there are cells responding to color contrasts and orientations. Up through the pathway, cells become more complex and combine results obtained from many previous cell outputs.

From V1, the information is transmitted to the "higher" brain areas V2 through V4, infero temporal cortex (IT), the middle temporal area (MT or V5), and the posterior parietal cortex (PP). Although there

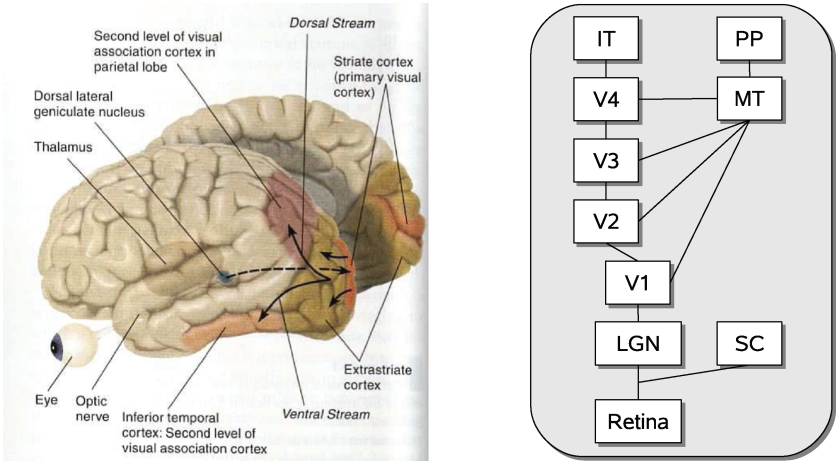


Fig. 1. Left: visual areas and pathways in the human brain (Figure from http://philosophy.hku.hk/courses/cogsci/ncc.php). Right: Some of the known connections of visual areas in the cortex (Figure adapted from Palmer [1999]).

are still many open questions concerning V1 [Olshausen and Field 2005; 2006], even less is known on the extrastriate areas. One of the most important findings during the last decade was that the processing of the visual information is not serial but highly parallel. Many authors have claimed that the extrastriate areas are functionally separated [Kandel et al. 1996; Zeki 1993; Livingstone and Hubel 1987; Palmer 1999]. Some of the areas process mainly color, some form, and some motion.

The processing leads to mainly two different locations in the brain: First, the color and form processing leads to IT, the area where the recognition of objects takes place. Since IT is concerned with the question of "what" is in a scene, this pathway is called the *what pathway*. Other names are the *P pathway* or *ventral stream* because of its location on the ventral part of the body. Second, the motion and depth processing leads to PP. Since this area is mainly concerned with the question of "where" something is in a scene, this pathway is also called the *where pathway*. Other names are the *M pathway* or *dorsal stream* because it lies dorsally.

Newer findings propose that there is much less segregation of feature computations. It is, for example, indicated that luminance and color are not separated, but there is a continuum of cells, varying from cells that respond only to luminance to a few cells that do not respond to luminance at all [Gegenfurtner 2003]. Additionally, the form processing is not clearly segregated from color processing, since most cells that respond to oriented edges respond also to color contrasts.

2.2 Visual Attention

In this section, we discuss several concepts of visual attention. More detailed information can be found in some books on this topic, for example, Pashler [1997], Styles [1997], and Johnson and Proctor [2003]. Here, we start with a definition of visual attention, and introduce the concepts of covert and overt attention, the units of attention, bottom-up saliency, and top-down guidance. Then, we elaborate on visual search, its efficiency, pop-out effects, and search asymmetries. Finally, we discuss the neurobiological correlates of attention.

2.2.1 *What is Visual Attention?* The concept of selective attention refers to a fact already mentioned by Aristotle: "it is impossible to perceive two objects coinstantaneously in the same sensory act." Although we usually have the impression to retain a rich representation of our visual world and that large changes to our environment will attract our attention, various experiments reveal that our ability

to detect changes is usually highly overestimated. Only a small region of the scene is analyzed in detail at each moment: The region that is currently attended. This is usually but not always the same region that is fixated by the eyes. That other regions than the attended one are usually largely ignored is shown, for example, in experiments on change blindness [Simons and Levin 1997; Rensink et al. 1997]. In these experiments, a significant change in a scene remains unnoticed, which means the observer is "blind" for this change.

The reason why people are nevertheless effective in everyday life, is that they are usually able to automatically attend to regions of interest in their surrounding and to scan a scene by rapidly changing the focus of attention. The order in which a scene is investigated is determined by the mechanisms of selective attention. For example, a definition is given in Corbetta [1990]: "Attention defines the mental ability to select stimuli, responses, memories, or thoughts that are behaviorally relevant among the many others that are behaviorally irrelevant." Although the term attention is also often used to refer to other psychological phenomena (e.g., the ability to remain alert for long periods of time), in this work, attention refers exclusively to perceptual selectivity.

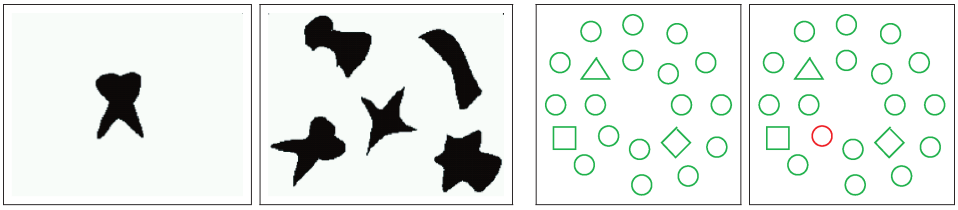
2.2.2 *Covert versus Overt Attention.* Usually, directing the focus of attention to a region of interest is associated with eye movements (*overt attention*). However, this is only half of the story. We are also able to attend to peripheral locations of interest without moving our eyes, a phenomenon which is called *covert attention*. This phenomenon was already described in the 19th century by von Helmholtz [1896]: "I found myself able to choose in advance which part of the dark field off to the side of the constantly fixated pinhole I wanted to perceive by indirect vision" (English translation from M. Mackeben in [Nakayama and Mackeben 1989]). This mechanism should be well known to each of us when we detect peripheral motion or suddenly spot our name in a list.

There is evidence that simple manipulation tasks can be performed without overt attention [Johansson et al. 2001]. On the other hand, there are cases in which an eye movement is not preceeded by covert attention: Findlay and Gilchrist [2001] found that in tasks like reading and complex object search, saccades (quick, simultaneous movements of both eyes in the same direction [Cassin and Solomon 1990]) were made with such frequency that covert attention could not have scanned the scene first. Even though, covert attention and saccadic eye movements usually work together: The focus of attention is directed to a region of interest followed by a saccade that fixates the region and enables the perception at a higher resolution. That covert and overt attention are not independent was shown by Deubel and Schneider [1996]: It is not possible to attend to one location while directing the eyes to a different one.

2.2.3 *The Unit of Attention.* During the last decades, there has been a long debate about the units of attention, which means about the target our attentional focus is directed to. Do we attend to spatial locations, to features, or to objects?

The majority of studies, both from psychophysics and from neurobiology, is about space-based attention (also referred to as location-based attention) [Posner 1980; Eriksen and St. James 1986; Yantis et al. 2002; Bisley and Goldberg 2003]. However, there is also strong evidence for feature-based attention [Treisman and Gelade 1980; Giesbrecht et al. 2003; Liu et al. 2003] and for object-based attention [Duncan 1984; Driver and Baylis 1998; Scholl 2001; Ben-Shahar et al. 2007; Einhäuser et al. 2008]. Today, most researchers believe that these theories are not mutually exclusive but that visual attention can be deployed to each of these candidate units [Vecera and Farah 1994; Fink et al. 1997; Yantis and Serences 2003]. A broad introduction and overview over the different approaches and studies can be found in Yantis [2000].

Finally, it is worth mentioning that there is often not just a single unit of attention. Humans are able to attend simultaneously to multiple regions of interest, usually between four and five regions. This



(a) Cueing experiment (b) Attentional capture

Fig. 2. (a) Cueing experiment: A cue (left) is presented for 200ms. Then, human subjects have to search for the cued shape in a search array (right) (Figure reprinted with permission from Vickery et al. [2005] © 2005 The Association for Research in Vision and Ophthalmology (ARVO)). (b) Attentional capture: in both displays, human subjects had to search for the diamond. Although they knew that color was unimportant in this search task, the red circle in the right display slowed down the search about 65ms (885 vs. 950ms) [Theeuwes 2004]. That means, the color pop-out "captures" the attention independent of the task (Figure adapted from Theeuwes [2004]).

has been shown in psychological [Pylyshyn and Storm 1988; Pylyshyn 2003; Awh and Pashler 2000] as well as neurobiological experiments [McMains and Somers 2004].

2.2.4 *Bottom-up versus Top-down Attention.* There are two major categories of factors that drive attention: *bottom-up factors* and *top-down factors* [Desimone and Duncan 1995]. Bottom-up factors are derived solely from the visual scene [Nothdurft 2005]. Regions of interest that attract our attention in a bottom-up way are called *salient* and the responsible feature for this reaction must be sufficiently discriminative with respect to surrounding features. Besides bottom-up attention, this attentional mechanism is also called exogenous, automatic, reflexive, or peripherally cued [Egeth and Yantis 1997].

On the other hand, top-down attention is driven by cognitive factors such as knowledge, expectations, and current goals [Corbetta and Shulman 2002]. Other terms for top-down attention are endogenous [Posner 1980], voluntary [Jonides 1981], or *centrally cued* attention. There are many intuitive examples of this process. Car drivers are more likely to see the petrol stations in a street and cyclists notice cycle tracks. If you are looking for a yellow highlighter on your desk, yellow regions will attract the gaze more readily than other regions.

Yarbus [1967] has already early shown that eye movements depend on the current task: For the same scene ("an unexpected visitor" that shows a room with a family and a person entering the room), subjects got different instructions 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. Visual context, such as the *gist* (semantic category) or the spatial layout of objects, also influence visual attention in a top-down manner. For example, Chun and Jiang [1998] have shown that targets appearing in learned configurations were detected more quickly.

In psychophysics, top-down influences are often investigated by so called *cueing experiments*. In these experiments, a "cue" directs the attention to the target. Cues may have different characteristics: They may indicate *where* the target will be, for example, by a central arrow that points into the direction of the target [Posner 1980; Styles 1997], or what the target will be, for example, the cue is a (similar or exact) picture of the target or a word (or sentence) that describes the target ("search for the black, vertical line") [Vickery et al. 2005; Wolfe et al. 2004] (see Figure 2(a)).

The performance in detecting a target is typically better in trials in which the target is present at the cued location than in trials in which the target appears at an uncued location; this was called the *Posner cueing paradigm* [Posner 1980]. A cue speeds up the search if it matches the target exactly and slows down the search if it is invalid. Deviations from the exact match slowdown search speed, although they

lead to faster speed compared with a neutral cue or a semantic cue [Vickery et al. 2005; Wolfe et al. 2004]. Recent physiological evidence from monkey experiments support these findings: Neurons give enhanced responses when a stimulus in their receptive field matches a feature of the target [Bichot et al. 2005].

Evidence from neurophysiological studies indicates that two independent but interacting brain areas are associated with the two attentional mechanisms [Corbetta and Shulman 2002]. During normal human perception, both mechanisms interact. As per Theeuwes [2004], the bottom-up influence is not voluntary suppressible: A highly salient region "captures" the focus of attention regardless of the task. For example, if there is an emergency bell, you will probably stop reading this article, regardless of how engrossed in the text you were. This effect is called attentional capture (see Figure 2(b)). Neural evidence from monkey experiments support these findings: Ogawa and Komatsu [2004] show that even if monkeys searched for a target of one dimension (shape or color), singletons (pop-out elements) from the other dimension (color or shape) induced high activation in some neurons. However, although attentional capture is definitely a strong effect that occurs frequently, there is also evidence that in some cases the bottom-up effects can be overridden completely [Bacon and Egeth 1994]. These difficulties are discussed in more detail in Connor et al. [2004]; a review of different studies on attentional capture can be found in Rauschenberger [2003].

Bottom-up attention mechanisms have been more thoroughly investigated than top-down mechanisms. One reason is that data-driven stimuli are easier to control than cognitive factors, such as knowledge and expectations. Even less is known on the interaction between the two processes.

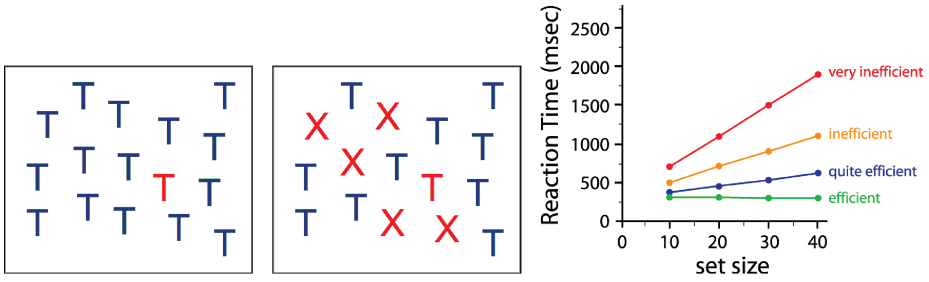
2.2.5 *Visual Search and Pop-out Effect.* An important tool in research on visual attention is *visual search* [Neisser 1967; Styles 1997; Wolfe 1998a]. The general question of visual search is: Given a target and a test image, is there an instance of the target in the test image? We perform visual search all the time in everyday life. For example, finding a friend in a crowd is such a visual search task. Tsotsos [1987; 1990] has proven that the problem of unbounded visual search is so complex that it in practice is unsolvable in acceptable time.[1] In contrast, *bounded visual search* (the target is explicitly known in advance) can be performed in linear time. Also, psychological experiments on visual search with known targets report that the search time complexity is linear and not exponential, thus the computational nature of the problem strongly suggests that attentional top-down influences play an important role during the search.

In psychophysical experiments, the efficiency of visual search is measured by the *reaction time* (also *response time*) (RT) that a subject needs to detect a target among a certain number of distractors (the elements that differ from the target) or by the *search accuracy*.

To measure the RT, a subject has to report a detail of the target or has to press one button if the target was detected and another if it is not present in the scene. The RT is represented as a function of set size (the number of elements in the display). The search efficiency is determined by the slopes and the intercepts of these RT $\times$ set size functions (see Figure 3(c)).

The searches vary in their efficiency: The smaller the slope of the function and the lower the value on the y-axis, the more efficient the search. Two extremes hereby are *serial* and *parallel* search. In serial search, the reaction time increases with the number of distractors, whereas in parallel search, the slope is near zero, that is, there is no significant variation in reaction time if the number of distractors grows; here, a target is found immediately without the need to perform several shifts of attention. Experiments by Wolfe [1998b] indicate that the studies of visual search should not be classified into the

[1]The problem is *NP-complete*, that is, it belongs to the hardest problems in computer science. No polynomial algorithm is known for this class of problems and they are expected to require exponential time in the worst-case [Garey and Johnson 1979].



(a) Feature search (b) Conjunction search (c) The continuum of search slopes

Fig. 3. (a) Feature search: The target (red T) differs from the distractors (blue T's) by a unique visual feature (pop-out effect). (b) Conjunction search: The target (red T) differs from the distractors (red Xs and blue Ts) by a conjunction of features. (c) The reaction time (RT) of a visual search task is a function of set size. The efficiency is measured by the intercept and slopes of the functions (Figure adapted from Wolfe [1998a]).

distinct groups "parallel" and "serial," since the increase in reaction time is a continuum. He suggests instead to describe them as "efficient" versus "inefficient." This allows one to use expressions like "more efficient than", "quite efficient", or "very inefficient" (see Figure 3(c)).

The concept of efficient search was discovered a long time ago. Already in the 11th century, Ibn Al-Haytham (English translation: [Sabra 1989]) found that "some of the particular properties of which the forms of visible objects are composed appear at the moment when sight glances at the object, while others appear only after scrutiny and contemplation." This effect is referred to as *pop-out effect*, according to the subjective impression that the target leaps out of the display to grab attention (see Figure 3(a)). Scenes with pop-outs are sometimes also referred to as *odd-man-out* scenes. Efficient search is often but not always accompanied by pop-out [Wolfe 1994]. Usually, pop-out effects only occur when the distractors are homogeneous, for example, the target is red and the distractors are green. Instead, if the distractors are green and yellow, search is efficient but there is no pop-out effect.

In *conjunction search tasks* (also *conjunctive search*), in which the target is defined by several features, the search is usually less efficient (see Figure 3(b)). However, the steepness of the slope depends on the experiment; there are also search tasks in which conjunction search is quite efficient [Wolfe 1998a; 1998b].

While experimentally simple to perform, RT measures are not sufficient to answer all questions concerning visual search. It documents only the completion of search and not the search process itself. Thus, neither spatial information (where is the subject looking during search and how many saccades are performed) nor temporal information (how long is each part fixated) can be measured. According to Zelinsky and Sheinberg [1997], measuring eye movements is more suitable to provide such information.

Another method to determine search efficiency is by measuring accuracy. A search stimulus is presented only briefly and followed by a mask that terminates the search. The time between the onset of the stimulus and that of the mask is called *stimulus onset asynchrony (SOA)*. The SOA is varied and accuracy is plotted as a function of SOA [Wolfe 1998a]. Easy search tasks can be performed efficiently even with short SOAs, whereas harder search tasks require longer SOAs. A single-stage Signal Detection Theory (SDT) model can predict these accuracy results in terms of the probability of correctly detecting the presence or absence of the target [Verghese 2001; Cameron et al. 2004] (see Section 2.3.3).

Finally, it is worth mentioning the *eccentricity effect*: The physical layout of the retina, with high resolution in the center and low resolution in the periphery, makes targets at peripheral locations more difficult to detect. Both reaction times and errors increase with increasing distance from the center [Carrasco et al. 1995].

There has been a multitude of experiments on visual search and many settings have been designed to discover which features enable efficient search and which do not. Some interesting examples are the search for numbers among letters, for mirrored letters among normal ones, for the silhouette of a "dead" elephant (legs to the top) among normal elephants [Wolfe 2001a], and for the face of another race among faces of the same race as the test subject [Levin 1996].

One purpose of these experiments is to study the *basic features* (also *primitive features* or *attributes*) of human perception, which means the features that are early and preattentively processed in the human brain and guide visual search. Testing the efficiency of visual search helps to investigate this, since efficient search is said to take place if the target is defined by a single basic feature and the distractors are homogeneous [Treisman and Gormican 1988]. Thus, finding out that a red blob pops out among green ones indicates that color is a basic feature. Opinions on what are basic features are still controversial. Some features are doubtless basic, others are guessed to be basic but there is limited data or dissenting opinions. A listing of the current opinion is presented by Wolfe and Horowitz [2004]. According to them, undoubted basic features are color, motion, orientation, and size (including length and spatial frequency). The role of luminance (intensity) is still unclear. In some studies, luminance behaves like colors, whereas in others, it acts more independently [Wolfe 1998a]. Probable basic features are luminance onset (flicker), luminance polarity, Vernier offset (a small lateral break in a line), stereoscopic depth and tilt, pictorial depth cues, shape, line termination, closure, topological status, and curvature. Features which are possibly basic, but have even less confidence, are lighting direction (shading), glossiness (luster), expansion, number, and aspect ratio. Features which are unconvincing but still possible are novelty, letter identity, and alphanumeric category. Finally, features which are probably not basic are intersection, optic flow, color change, three-dimensional volumes, faces, your name, and semantic categories as "animal" or "scary." While this listing does not claim to be exhaustive, it gives a good overview about the current state of research.

An interesting effect in visual search tasks are *search asymmetries*, which means the effect that a search for stimulus "A" among distractors "B" produces different results from a search for "B" among "A"s. An example is that finding a tilted line among vertical distractors is easier than vice versa (see Figure 4). An explanation is proposed by Treisman and Gormican [1988]: The authors claim that it is easier to find deviations among canonical stimuli than vice versa. Given that vertical is a canonical stimulus, the tilted line is a deviation and may be detected fast. Therefore, by investigating search asymmetries, it is possible to determine the canonical stimuli of visual processing, which might be identical to feature detectors. For example, Treisman suggests that for color, the canonical stimuli are red, green, blue, and yellow; for orientation, they are vertical, horizontal, and left and right diagonal; and for luminance there exist separate detectors for darker and lighter contrasts [Treisman 1993]. Especially when building a computational model of visual attention, this is of significant interest: If it is clear what feature detectors exist in the human brain, it might be adequate to focus on the computation of these features. However, one should be careful to accept evidence about search asymmetries. Findings by Rosenholtz [2001] indicate that the asymmetries in many of the studies are due to built-in design asymmetries instead of to an underlying asymmetry in the search mechanism. A comprehensive overview about search asymmetries is provided by Wolfe [2001a], more papers can be found in the same issue of Perception & Psychophysics, 63 (3), 2001.

2.2.6 *Neurobiological Correlates of Visual Attention.* The mechanisms of selective attention in the human brain still belong to the open problems in the field of research on perception. Perhaps the most prominent outcome of neurophysiological findings on visual attention is that there is no single brain area guiding the attention, but neural correlates of visual selection appear to be reflected in nearly all brain areas associated with visual processing [Maunsell 1995]. Additionally, new findings indicate

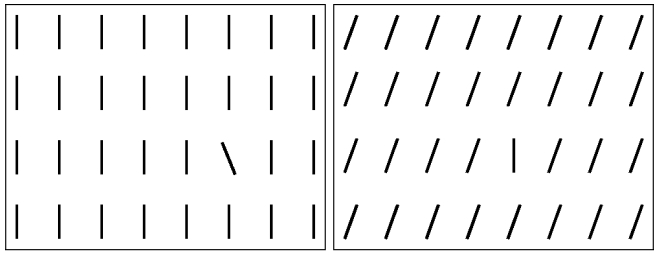


Fig. 4. Search asymmetries: It is easier to detect a tilted line among vertical distractors (a) than vice versa (b)

that many brain areas share the processing of information from different senses and there is growing evidence that large parts of the cortex are multisensory [Ghazanfar and Schroeder 2006].

Attentional mechanisms are carried out by a network of anatomical areas [Corbetta and Shulman 2002]. Important areas of this network are the Posterior Parietal cortex (PP), the Superior Colliculus (SC), the lateral intraparietal area (LIP), the frontal eye field (FEF), and the pulvinar. Regarding the question of which area fulfills which task, the opinions diverge. We review several findings here.

Posner and Petersen [1990] describe three major functions concerning attention: orienting of attention, target detection, and alertness. They claim that the first function, the orienting of attention to a salient stimulus, is carried out by the interaction of three areas: the PP, the SC, and the pulvinar. The PP is responsible for disengaging the focus of attention from its present location (inhibition of return), the SC shifts the attention to a new location, and the pulvinar is specialized in reading out the data from the indexed location. Posner and Petersen call this combination of systems the *posterior attention system*. The second attentional function, the detection of a target, is carried out by what the authors call the *anterior attention system*. They claim that the anterior cingulate gyrus in the frontal part of the brain is involved in this task. Finally, they state that the alertness to high-priority signals is dependent on activity in the norepinephrine system (NE) arising in the locus coeruleus.

Brain areas involved in guiding eye movements are the FEF and the SC. Furthermore, Bichot [2001] claims that the FEF is the place where a kind of saliency map is located, which derives information from bottom-up as well as from top-down influences. Other groups locate the saliency map at different areas, for example, at LIP [Gottlieb et al. 1998], at SC [Findlay and Walker 1999], at V1 [Li 2005], or at V4 [Mazer and Gallant 2003].

There has been evidence that the source of top-down biasing signals may derive from a network of areas in parietal and frontal cortex. According to Kastner and Ungerleider [2001], these areas include the superior parietal lobule (SPL), the FEF and the supplementary eye field (SEF), and, less consistently, areas in the inferior parietal lobule (IPL), the lateral prefrontal cortex in the region of the middle frontal gyrus (MFG), and the anterior cingulate cortex. Corbetta and Shulman [2002] find transient responses to a cue in the occipital lobe (fusiform and MT+) and more sustained responses in the dorsal posterior parietal cortex along the intraparietal sulcus (IPs) and in the frontal cortex at or near the putative human homologue of the FEFs. According to Ogawa and Komatsu [2004], the interaction of bottom-up and top-down cues takes place in V4.

In summary, at the current time, it is known that there is not a single brain area that controls attention but a network of areas. Several areas have been verified to be involved in attentional processes, but the accurate task and behavior of each area as well as the interplay among them still remain open questions.

2.3 Psychophysical Theories and Models of Attention

In the field of psychology, a wide variety of theories and models on visual attention exist. Their objective is to explain and better understand human perception. Here, we introduce the theories and models that have been most influential for computational attention systems. More on psychological attention models can be found in the review of Bundesen and Habekost [2005].

2.3.1 *Feature Integration Theory.* The Feature Integration Theory (FIT) of Treisman [1980] has been one of the most influential theories in the field of visual attention. The theory was first introduced in 1980, but it was steadily modified and adapted to current research findings. One has to be careful when referring to FIT, since some of the older findings concerning a dichotomy between serial and parallel search are no longer believed to be valid (see Section 2.2.5). An overview of the theory is found in Treisman [1993].

The theory claims that "different features are registered early, automatically and in parallel across the visual field, while objects are identified separately and only at a later stage, which requires focused attention" [Treisman and Gelade 1980]. Information from the resulting *feature maps*—topographical maps that highlight conspicuities according to the respective feature—is collected in a *master map of location*. This map specifies where in the display things are, but not what they are. Scanning serially through this map focuses the attention on the selected scene regions and provides this data for higher perception tasks (see Figure 5).

Treisman mentions that the search for a target is easier the more features differentiate the target from the distractors. If the target has no unique features but differs from the distractors only in how its features are combined, the search is more difficult and often requires focused attention (conjunctive search). This usually results in longer search times. However, if the features of the target are known in advance, conjunction search can sometimes be accomplished rapidly. She proposes that this is done by inhibiting the feature maps, which code nontarget features.

Additionally, Treisman introduced so called *object files* as "temporary episodic representations of objects." An object file "collects the sensory information that has so far been received about the object. This information can be matched to stored descriptions to identify or classify the object" [Kahneman and Treisman 1992].

2.3.2 *Guided Search Model.* Besides FIT, the Guided Search Model of Wolfe is among the most influential work for computational visual attention systems. Originally, the model was created as an answer to some criticism on early versions of the FIT. During the years, a competition arose between Treisman's and Wolfe's work, resulting in continuously improved versions of the models.

The basic goal of the model is to explain and predict the results of visual search experiments. There has also been a computer simulation of the model [Cave and Wolfe 1990; Wolfe 1994]. As Treisman's work, the model has been continuously developed further over the years. Mimicking the convention of numbered software upgrades, Wolfe has denoted successive versions of his model as Guided Search 1.0 [Wolfe et al. 1989], Guided Search 2.0 [Wolfe 1994], Guided Search 3.0 [Wolfe and Gancarz 1996], and Guided Search 4.0 [Wolfe 2001b; 2007]. Here, we focus on Guided Search 2.0, since this is the best elaborated description of the model. Versions 3.0 and 4.0 contain changes, which are of minor importance here. For example, in 3.0, eye movements are included into the model, and in 4.0, the implementation of memory for previously visited items and locations is improved.

The architecture of the model is depicted in Figure 6. It shares many concepts with the FIT but is more detailed in several aspects that are necessary for computer implementations. An interesting point is that in addition to bottom-up saliency, the model also considers the influence of top-down information by selecting the feature type, which distinguishes the target best from its distractors.

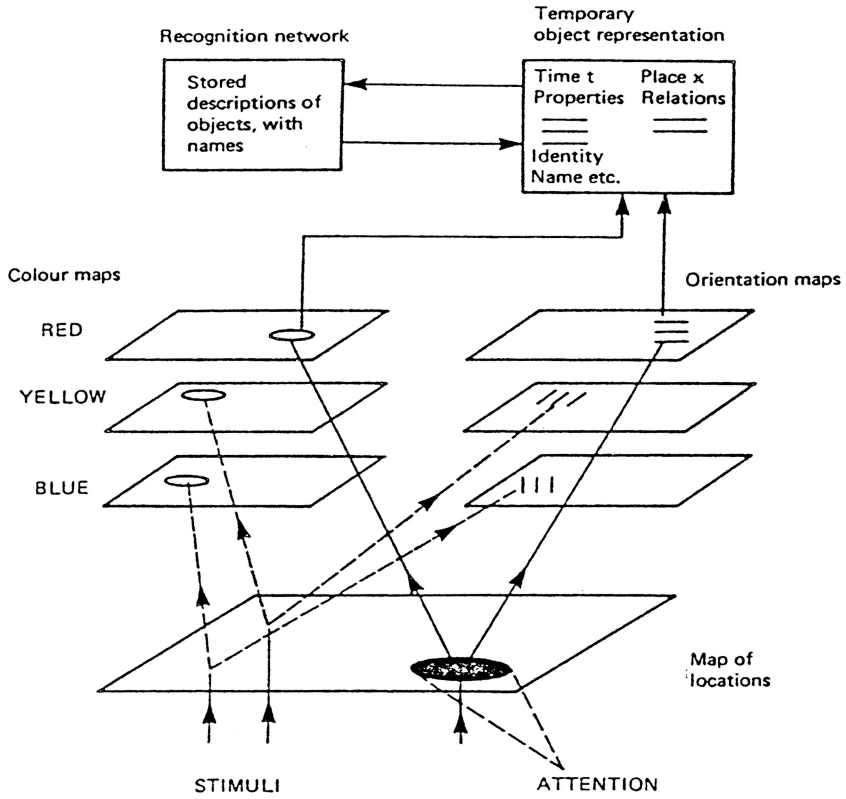


Fig. 5. Model of the Feature Integration Theory (FIT) (Figure reprinted with permission from Treisman and Gormican [1988] © 1988 American Psychological Association (APA)).

2.3.3 *Other Theories and Models.* Other than these approaches, there is a wide variety of psychophysical models on visual attention. Eriksen and St. James [1986] have introduced the zoom lens model. In this model, the spatial extent of the attentional focus can be manipulated by precueing. In this model, the scene is investigated by a spotlight with varying size. Many attention models fall into the category of connectionist models, which means models based on neural networks. They are composed of a large number of processing units connected by inhibitory or excitatory links. Examples are the dynamic routing circuit [Olshausen et al. 1993] and the models MORSEL [Mozer 1987], SLAM (selective attention model) [Phaf et al. 1990], SERR (search via recursive rejection) [Humphreys and Müller 1993], and SAIM (selective attention for identification model) [Heinke and Humphreys 2003].

A formal mathematical model is presented by Logan [1996]: the CODE theory of Visual Attention (CTVA). It integrates the COntour DEtector (CODE) theory for perceptual grouping [van Oeffelen and Vos 1982] with the Theory of Visual Attention (TVA) [Bundesen 1990]. The theory is based on a race model of selection. In these models, a scene is processed in parallel and the element that first finishes processing is selected (the winner of the race). That means a target is processed faster than the distractors in a scene. Newer work concerning CTVA can be found in Bundesen [1998].

Another type of psychological model is based on the *signal detection theory (SDT)*, a method to measure the search accuracy by quantifying the ability to distinguish between signal and noise [Green and Swets 1966; Abdi 2007]. The distractors in a search task are considered to be noise and the target is signal plus noise. In an SDT experiment, one or several search displays are presented briefly and masked afterward.

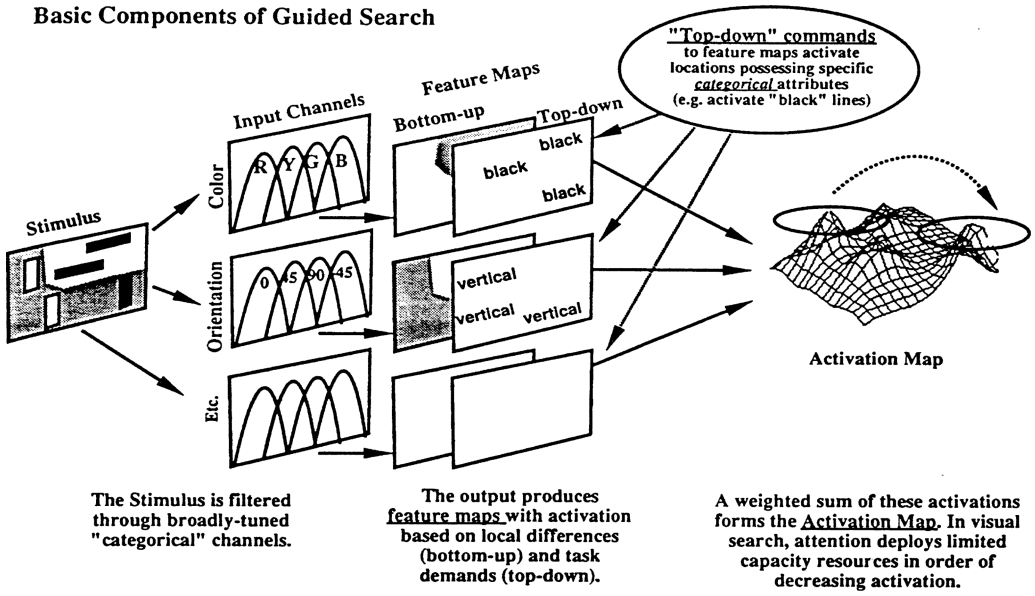


Fig. 6. The Guided Search model of Wolfe (Figure reprinted with permission from Wolfe [1994] ©1994 Psychonomic Society).

In yes/no designs, one display is presented and the observer has to decide whether the target was present or not; in an M-AFC (alternative forced-choice) design, M displays are shown and the observer has to identify the display containing the target. The order of presentation is varied randomly in different trials. Performance is measured by determining how well the target can be distinguished from the distractors, and the SDT model is used to calculate the performance degradation with increasing set size. SDT models that have been used to predict human performance for detection and localization of targets have been presented in Palmer et al. [1993], Verghese [2001], and Eckstein et al. [2000].

An interesting theoretical model has been introduced by Rensink [2000]. His triadic architecture consists of three parts: First, a low-level vision system produces proto-objects rapidly and in parallel. Second, a limited-capacity attentional system forms these structures into stable object representations. Finally, a nonattentional system provides setting information, for example, on the *gist*—the abstract meaning of a scene, for example, beach scene, city scene, and so on—and on the *layout*—the spatial arrangement of the objects in a scene. This information influences the attentional system, for example, by restricting the search for a person on the sand region of a beach scene and ignoring the sky region.

3. COMPUTATIONAL ATTENTION SYSTEMS

In computer vision and robotics, there is increasing interest in a selection mechanism, which determines the most relevant parts within the large amount of visual data. Visual attention is such a selection mechanism; therefore, many computational attention systems have been built during the last three decades (mainly during the last 5 to 10 years). The systems that are considered here have in common that they are built on the psychological and neurobiological concepts and theories that have been presented in the previous section. In contrast to the models described in Section 2.3, we focus here on computational systems with an engineering objective. The objective of these systems is less in understanding human perception but more in improving existing vision systems. Usually, they are able to cope not only with synthetical images but also with natural scenes. The systems vary in detail, but most of them have a similar structure.

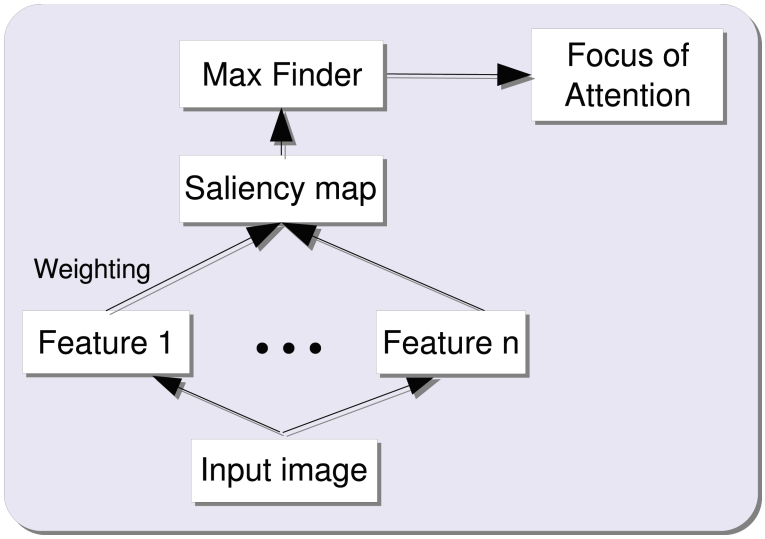


Fig. 7. General structure of most bottom-up attention systems.

We start with a description of the general structure of typical computational attention systems (Section 3.1). Then, we continue with a more detailed investigation of the characteristics of different approaches. Connectionist versus filter models are distinguished (Section 3.2), the choice of different feature channels is discussed (Section 3.3), and the integration of top-down cues is introduced (Section 3.4). Finally, we provide a chronological overview of important computational attention systems (Section 3.5).

3.1 General Structure

Most computational attention systems have a very similar structure, which is depicted in Figure 7. This structure is originally adapted from psychological theories such as the feature integration theory [Treisman and Gormican 1988] and the Guided Search model [Wolfe 1994]. The main idea is to compute several features in parallel and to fuse their saliencies in a representation, which is usually called a *saliency map*. Detailed information on how to implement such a system is presented, for example, in Itti et al. [1998] or Frintrop [2005]. The necessary background knowledge on computer vision methods is summed up in the appendix of Frintrop [2005]. An overview of the techniques follows.

In filter-based models (see Section 3.2), usually the first step is to compute one or several image pyramids from the input image, to enable the computation of features on different scales [Itti et al. 1998]. This saves computation time, since it avoids explicitly applying large filters to the image. The following computations are performed on several of the layers of the pyramid, usually ignoring the first, finest layers to reduce the influence of noise. An alternative approach is to use integral images for a fast computation of features on different scales [Frintrop et al. 2007].

An interesting approach is to exchange this standard uniform sampling scheme with a more biologically plausible space-variant sampling, according to the space-variant arrangement of photoreceptors in the retina. However, Vincent et al. [2007] have found that this causes feature coding unreliability and that there is "only a very weak relation between target eccentricity and discrimination performance." Interesting in this context would be a replacement of the normal camera with a retina-like sensor to achieve space-variant sampling [Sandini and Metta 2002].

Next, several features are computed in parallel, and feature-dependent saliencies are computed for each feature channel. The information for different features is collected in maps. These might be represented as gray-scale images, in which the brightness of a pixel is proportional to its saliency (see Figure 8), or as collections of nodes of an artificial neural network.

Commonly used features are intensity, color, and orientation; a detailed investigation of the choice of features is presented in Section 3.3. Usually, the computation of these feature dimensions is subdivided into the computation of several feature types; for example, for the feature dimension color, the feature types red, green, blue, and yellow may be computed. The feature types are usually displayed in feature maps and summed up to feature dependent saliency maps, which are often called *conspicuity maps*, a term first used by Milanese [1993]. The conspicuity maps are finally fused to a single saliency map [Koch and Ullman 1985], a term that is widely used and corresponds to Treisman's master map of location.

The feature maps collect the local within-map contrast. This is usually computed by *center-surround mechanisms*, also called *center-surround differences* [Marr 1982]. This operation compares the average value of a center region to the average value of a surrounding region, inspired from the ganglion cells in the visual receptive fields of the human visual system [Palmer 1999]. In most implementations, the feature detectors are based on rectangular regions, which makes them less accurate than a circular filter but much easier to implement and faster to compute.

A very important aspect of attentional systems, maybe even the most important one, is the way different maps are fused, that is, how the between-map interaction takes place. How is it accomplished that the important information is not lost in the large collection of maps? How is it achieved that the red ball on green grass pops out, although this saliency only shows up strongly in one of the maps, namely the red-green map? It is not yet completely clear how this task is solved in the brain nor is an optimal solution known how to solve this problem in a computational system. Usually, a weighting function, we call it *uniqueness weight* [Frintrop 2005], is applied to each map before summing up the maps. This weighting function determines the uniqueness of features: If there is only a single bright region in a map, its uniqueness weight is high; if there are several equally bright regions, it is lower. One simple solution to compute this is to determine the number of local maxima m in each map and divide each pixel by the square root of m [Frintrop 2005]. Other solutions are presented, for example, in Itti et al. [1998], Itti and Koch [2001b], and Harel et al. [2007]. An evaluation of different weighting approaches has, to our knowledge, not yet been done. However, even if it is not clear what the optimal weighting looks like, all these approaches are able to reproduce the human pop-out effect and detect outliers in images from psychophysical experiments, such as the one in Figure 3(a). An example of applying such a weighting function to real-world images is shown in Figure 8. Note that this weighting by uniqueness covers only the bottom-up aspect of visual attention. In humans, visual attention almost always top-down effects participate and guide our attention according to the current situation. These effects will be discussed in Section 3.4.

Before the weighted maps are summed up, they are usually normalized. This is done to weed out the differences between a priori not comparable modalities with different extraction mechanisms. Additionally, it prevents the higher weighting of channels that have more feature maps than others. Most straightforward is to normalize all maps to a fixed range [Itti et al. 1998]. This results in problems if one channel is more important than another, since information about the magnitude of the maps is removed. A method that keeps this information is to determine the maximum M of all maps, which shall be summed up and normalize each map to the range $[0..M]$ [Frintrop et al. 2005]. An alternative that scales each conspicuity map with respect to a long-term estimate of its maximum is presented in Ouerhani et al. [2006].

After weighing and normalizing, the maps are summed up to the saliency map. This saliency map might already be regarded as an output of the system, since it shows the saliency for each region of

a scene. But usually, the output of the system is a trajectory of image regions—mimicking human saccades—which starts with the highest saliency value. The selected image regions are local maxima in the saliency map. They might be determined by a winner-take-all (WTA) network which was introduced by Koch and Ullman [1985]. It shows how the selection of a maximum is implementable by neural networks, which means by single units that are only locally connected. This approach is strongly biologically motivated and shows how such a mechanism might work in the human brain. A simpler, more technically motivated alternative to the WTA with the same result is to straightforwardly determine the pixel with the largest intensity value in the image. This method requires fewer operations to compute the most salient region, but note that the WTA might be a good solution if implemented on a parallel architecture such as a GPU.

Since the Focus of Attention (FOA) is usually not on a single point but on a region (we call it *most salient region* (MSR)), the next step is to determine this region. The simplest approach is to determine a fixed-sized circular region around the most salient point [Itti et al. 1998]. More sophisticated approaches integrate segmentation approaches on feature [Walther 2006] or saliency maps [Frintrop 2005] to determine a irregularly shaped attention region.

After the FOA has been computed, some systems determine a feature vector, which describes how much each feature contributes to the region. Usually, the local or global surrounding of the region is also considered [Navalpakkam et al. 2005; Frintrop et al. 2005]. The vector can be used to match the region to previously seen regions, for example, to search for similar regions in a top-down guided visual search task [Frintrop et al. 2005] or to track a region over subsequent frames [Frintrop and Kessel 2009]. Such a feature vector resembles the psychological concept of object files as temporary episodic representations of objects, which were introduced by Treisman (see Section 2.3.1).

The most common method to obtain a trajectory of image regions that mimics a human search trajectory is called Inhibition of Return (IOR). It refers to the observation that in human vision, the speed and accuracy with which a target is detected is impaired after the target was attended. It was first described by Posner and Cohen [1984] and prevents that the FOA stays at the most salient region. In computational systems, IOR is implemented by inhibiting (reseting) the surrounding region in the saliency map. The surrounding region can be a fixed region around the FOA (spatial inhibition) or the MSR (feature-based inhibition), or a combination, as in Aziz and Mertsching [2007]. Interesting in this context is that Horowitz and Wolfe [2003] discovered that human visual search has no complete memory, that is, not all items in a search display are marked after they have been considered. That means IOR probably works only for a few items at a time. A possible implementation inhibits each distractor for a short time, dependent on a probabilistic value. In Wolfe [2007], this results, on average, in about three inhibited items at a time. An alternative that is simple to implement and obtains good results is to determine all peaks in the saliency map, sort them by their saliency values, and direct the FOA attention subsequently to each salient region [Frintrop and Cremers 2007]. IOR is not necessary in this approach. We found that this method yielded better results than the IOR method, since it avoids "border effects" in which the FOA returns to the border of the inhibited region. More difficult is IOR in dynamic scenes, since not only the currently focused region must be tracked over time but also every inhibited region [Backer et al. 2001].

The structure described so far was purely bottom-up. Including prior knowledge and target information to the system in a top-down manner is described in Section 3.4.

3.2 Connectionist versus Filter Models

A basic difference between models concerns the underlying structure, which is either based on neural networks (connectionist models) or on a collection of gray-scale maps (filter models). Usually, the connectionist models claim to be more biologically plausible than the filter models, since they have single

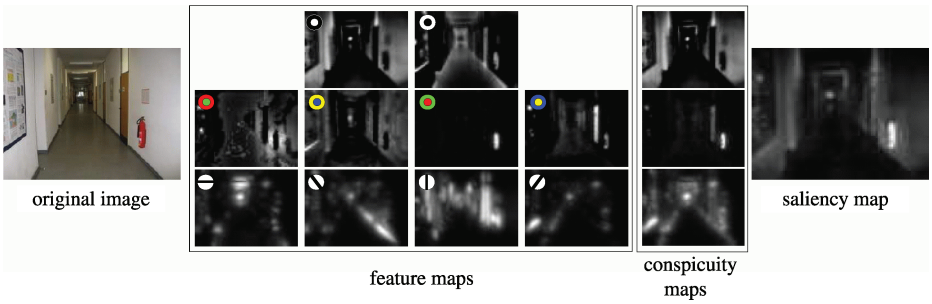


Fig. 8. Feature, conspicuity, and saliency map(s) for an example image computed with the attention system VOCUS [Frintrop 2005]. 1st row: intensity maps (on-off and off-on). 2nd row: color maps (green, blue, red, yellow). 3rd row: orientation maps (0°, 45°, 90°, 135°). The feature map 'red' is weighted highest, since the red fire extinguisher is unique in the scene. This results in a strong peak in the conspicuity color map and finally in a strong saliency in the saliency map.

units corresponding to neurons in the human brain, but it has to be noted that they are still a high abstraction from the processes in the brain. Examples of connectionist systems of visual attention are presented in Olshausen et al. [1993], Postma [1994], Tsotsos et al. [1995], Baluja and Pomerleau [1997], and Cave [1999]. Many psychophysical models fall into this category, too, for example, Mozer [1987], Phaf et al. [1990], Humphreys and Müller [1993], Heinke and Humphreys [2003]. An advantage of connectionist models is that they are—at least theoretically—able to show a different behavior for each neuron, whereas in filter models, usually each pixel in a map is treated equally. In practice, treating each unit differently is usually too costly, so a group of units shows the same behavior.

Advantages of filter models are that they can profit from approved techniques in computer vision and that they are especially well suited for the application to real-world images. Examples of linear filter systems of visual attention are presented in Milanese [1993], Itti et al. [1998], Backer et al. [2001], Sun and Fisher [2003], Heidemann et al. [2004], Hamker [2005], and Frintrop [2005].

3.3 The Choice of Features

Many computational attention systems focus on the computation of mainly three features: intensity, color, and orientation [Itti et al. 1998; Draper and Lionelle 2005; Sun and Fisher 2003; Ramström and Christensen 2004]. Reasons for this choice are that these features belong to the basic features proposed in psychological and biological work [Treisman 1993; Palmer 1999; Wolfe 1994; Wolfe and Horowitz 2004] and that they are relatively easy to compute. A special case of color computation is the separate computation of skin color [Rae 2000; Heidemann et al. 2004; Lee et al. 2003]. This is often useful if faces or hand gestures have to be detected. Other features that are considered are, for example, curvature [Milanese 1993], spatial resolution [Hamker 2005], optical flow [Tsotsos et al. 1995; Vijayakumar et al. 2001], flicker [Itti et al. 2003], or corners [Fraundorfer and Bischof 2003; Heidemann et al. 2004; Ouerhani et al. 2005]. Several systems compute also more complex features that use approved techniques of computer vision to extract image information. Examples for such features are entropy [Kadir and Brady 2001; Heidemann et al. 2004], Shannon's self-information measure [Bruce and Tsotsos 2005b], ellipses [Lee et al. 2003], eccentricity [Backer et al. 2001], or symmetry [Backer et al. 2001; Heidemann et al. 2004; Lee et al. 2003].

A very important feature in human perception is motion. Some systems that consider motion as a feature are presented in Maki et al. [2000], Ouerhani [2003], Itti et al. [2003], and Rae [2000]. These approaches implement a simple kind of motion detection: Usually, two subsequent images in a video stream are subtracted and the difference codes the feature conspicuity. Note that these approaches

require a static camera and are not applicable on a mobile system as a robot. A sophisticated approach concerning motion was proposed by Tsotsos et al. [2005]. This approach considers the direction of movements and processes motion on several levels similar to the processing in the brain regions V1, MT, and MST. In the previously described approaches, motion and static features are combined in a competitive scheme: They all contribute to a saliency map and the strongest cue wins. Bur et al. [2007] propose instead a motion priority scheme in which motion is prioritized by suppressing the static features in presence of motion.

Another important but rarely considered aspect in human perception is depth. From the psychological literature, it is not clear whether depth is simply a feature or something else; definitely, it has some unusual properties distinguishing it from other features. If one of the dimensions in a conjunctive search is depth, a second feature can be searched in parallel [Nakayama and Silverman 1986], a property that does not exist for the other features. Computing depth for an attention system is usually solved with stereo vision [Maki et al. 2000; Bruce and Tsotsos 2005a; Björkman and Eklundh 2007]. Another approach is to use special sensors to obtain depth data (e.g., 3D laser scanners), which provide dense and precise depth information and may provide additionally reflection data [Frintrop et al. 2005], or 3D cameras [Ouerhani and Hügli 2000].

Finally, it may be noted that although considering more features usually results in more accurate and biologically plausible detection results, it also reduces the processing speed, since the parallel models are usually implemented sequentially. Therefore, a trade-off has to be found between accuracy and speed. Using three to four feature channels seems to be a useful compromise for most systems.

3.4 Top-Down Cues

As outlined in Section 2.2.4, top-down cues play an important role in human perception. For a computational attention system, they are equally important: Most systems shall not only detect bottom-up salient regions, but there are also goals to achieve and targets to detect. Despite the well-known significance of top-down cues, most systems consider only bottom-up computations.

In human perception, different kinds of top-down influences exsist. They have in common that they represent information on the world or the state of the subject (or system). This includes aspects like current tasks and prior knowledge about the target, the scene, or the objects that might occur in the environment, as well as emotions, desires, and motivations. In the following, we discuss these different kinds of top-down information.

Emotions, desires, and motivations are hard to conceptualize and are not realized in any computer system we know about. Wells and Matthews [1994] provide a review from a psychological perspective about attention and emotion; Fragopanagos and Taylor [2006] present a neurobiological model about the interplay of attention and emotions in the human brain. The interaction of attention, emotions, motivations, and goals is discussed by Balkenius [2000], but in his computer simulation, these aspects are not considered.

Top-down information that refers to knowledge of the outer world, which means of the background scene or of the objects that might occur, is considered in several systems. In these approaches, for example, all objects of a database that might occur in a scene are investigated in advance and their most discriminative regions are determined, that is, the regions that distinguish an object best from all others in the database [Fritz et al. 2004; Pessoa and Exel 1999]. Another approach is to regard context information, which means searching for a person in a street scene is restricted to the street region; the sky region is ignored. The contextual information is obtained from past search experiences in similar environments [Oliva et al. 2003; Torralba 2003b]. Another kind of context that can be integrated into attention models is the gist, that is, the semantic category of the scene, such as “office scene” or “forest” [Oliva 2005]. The gist is known to guide eye movements [Torralba 2003a] and is usually computed as a

vector of contextual features. In visual attention systems, the gist may be computed directly from the feature channels [Siagian and Itti 2009].

One important kind of top-down information is the prior knowledge about a target that is used to perform visual search. Systems regarding this kind of top-down information use knowledge of the target to influence the computation of the most salient region. This knowledge is usually learned in a preceding training phase, but might in simpler approaches also be provided manually by the user.

In existing systems, the target information influences the processing at different stages: The simplest solution computes the bottom-up saliency map and investigates the target similarity of the most salient regions [Rao et al. 2002; Lee et al. 2003]. Only the most salient targets in a scene can be found with this approach. More elaborated is the tuning of the conspicuity maps [Milanese et al. 1994; Hamker 2005], but biologically most plausible and also most useful from an engineering perspective is the approach to already bias the feature types [Tsotsos et al. 1995; Frintrop et al. 2005; Navalpakkam and Itti 2006a]. This is supported by findings of Navalpakkam and Itti [2006b]: Not only the information about the feature dimensions influence top-down search but also information about feature types.

Different methods exist for influencing the maps with the target information. Some approaches inhibit the target-irrelevant regions [Tsotsos et al. 1995; Choi et al. 2004], whereas others prefer to excite target-relevant regions [Hamker 2005]. Newer findings suggest that inhibition and excitation both play an important role [Navalpakkam et al. 2004]; this is realized in Navalpakkam et al. [2005] and Frintrop et al. [2005]. Navalpakkam and Itti [2006a] present an interesting approach in which knowledge about a target, as well as about distractors influences the search. Vincent et al. [2007] learn the optimal feature map weights with multiple linear regression.

If human behavior shall be imitated, the bottom-up and the top-down saliency have to be fused to obtain a single focus of attention. Note, however, that in a computational system, it is also possible to deal with both maps in parallel and use the bottom-up and the top-down information for different purposes. The decision whether to fuse the maps or not has to be done depending on the application. If the maps shall be fused, one difficulty is how to combine the weighting for uniqueness (bottom-up) and the weighting for target-relevance (top-down). One possibility is to multiply the bottom-up maps with the top-down feature weights after applying the uniqueness weight [Hamker 2005; Navalpakkam et al. 2005]. A problem with this approach is that it is difficult to find nonsalient objects, since the bottom-up computations assign a very low saliency to the target region. One approach to overcome this problem is to separate bottom-up and top-down computations and to finally fuse them again, as done by Frintrop et al. [2005]. Here, the contribution of bottom-up and top-down cues is adjusted by a parameter t, which has to be set according to the system state: In exploration mode, there is a high bottom-up contribution; in search mode, the parameter shall be set proportionally to the search priority. Rasolzadeh et al. [2009] have adopted this idea and present an extension in which t can vary over time depending on the energy of bottom-up and top-down saliency maps. Xu et al. [2009] propose an approach that switches automatically between bottom-up and top-down behavior depending on the two internal robot states “observing” and “operating.”

The evaluation of top-down attention systems is discussed in Section 3.6.

3.5 Important Attention Systems in Chronological Order

In this section, we will present some of the most important attention systems in a chronological order and mention their particularities.

The first computational architecture of visual attention was introduced by Koch and Ullman [1985], which was inspired by the Feature Integration Theory. When it was first published, the model was not yet implemented, but it provided the algorithmic reasoning serving as a foundation for later

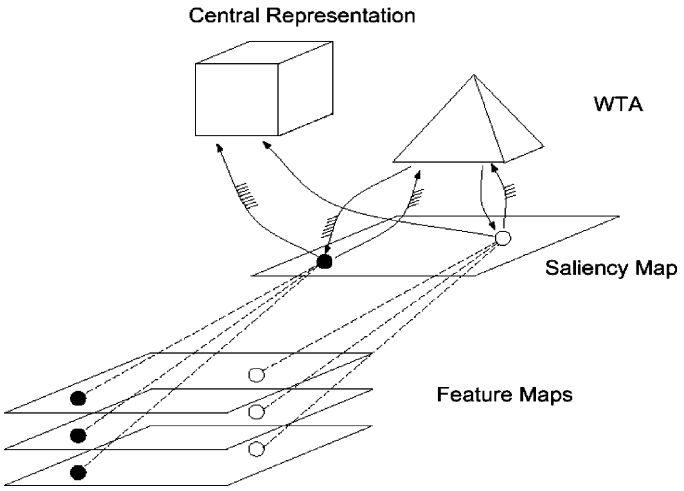


Fig. 9. The Koch-Ullman model. Different features are computed in parallel and their conspicuities are represented in several feature maps. A central saliency map combines the saliencies of the features and a winner take all network (WTA) determines the most salient location. This region is routed to the central representation where complex processing takes place (Figure reprinted with permission from Koch and Ullman [1985] © Springer Science and Business Media).

implementations and for many current computational attention systems. An important contribution of their work is the WTA network (see Figure 9).

One of the first implementations of an attention system was presented by Clark and Ferrier [1988]. Based on the Koch–Ullman model, it contains feature maps, which are weighted and summed up to a saliency map. The feature computations are performed by filter operations, realized by a special purpose image processing system, so the system belongs to the class of filter-based models.

Another early filter-based attention model was introduced by Milanese [1993]. In a derivative, Milanese et al. [1994] include top-down information from an object recognition system realized by distributed associative memories (DAMs). By first introducing concepts like conspicuity maps and feature computations based on center-surround mechanisms (called "conspicuity operator"), the system has set benchmarks for several techniques, which were used in computational attention models until today.

One of the oldest attention models, which is widely known and still developed further is Tsotsos' selective tuning (ST) model of visual attention [Tsotsos 1990; 1993; Tsotsos et al. 1995]. It is a connectionist model that consists of a pyramidal architecture with an inhibitory beam (see Figure 10). It is also possible to consider target-specific top-down cues by either inhibiting all regions with features different from the target features or regions of a specified location. The model has been implemented for several features, for example, luminance, orientation, color opponency [Tsotsos et al. 1995], motion [Tsotsos et al. 2005], and depth from stereo vision [Bruce and Tsotsos 2005a]. Originally, each version of the ST model processed only one feature dimension, but it was recently extended to perform feature binding [Rothenstein and Tsotsos 2006b; Tsotsos et al. 2008].

An unusual adaptation of Tsotsos's model is provided by Ramström and Christensen [2002]: The distributed control of the attention system is performed by game theory concepts. The nodes of the pyramid are subject to trading on a market, the features are the goods, rare goods are expensive (the features are salient), and the outcome of the trading represents the saliency.

One of the currently best-known attention systems is the Neuromorphic Vision Toolkit (NVT) (Figure 11), a derivative of the Koch–Ullman model, which is steadily kept up to date by the group

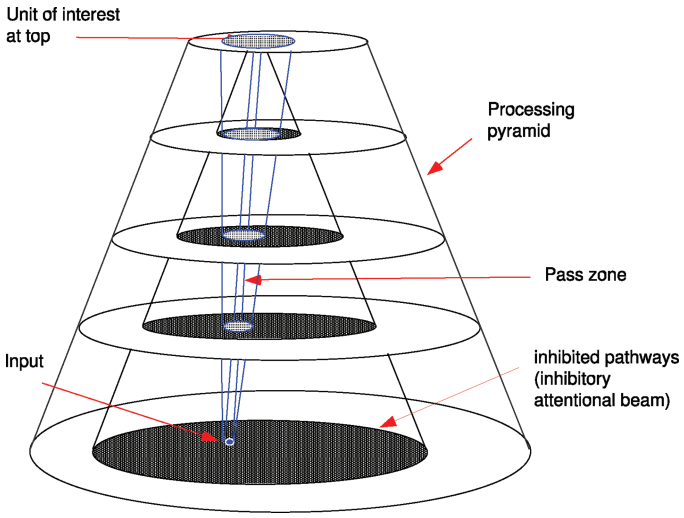


Fig. 10. The inhibitory attentional beam of Tsotsos et al. The selection process requires two traversals of the pyramid: First, the input traverses the pyramid in a feed-forward manner (pass zone). Second, the hierarchy of WTA processes is activated in a top-down manner to localize the strongest item in each layer while pruning parts of the pyramid that do not contribute to the most salient item (inhibit zone) (Figure kindly provided by John Tsotsos).

around Itti [Itti et al. 1998; Itti and Koch 2001a; Navalpakkam and Itti 2006a]. Their model as well as their implementation serve as a basis for many research groups; one reason for this is the good documentation and the online availability of the source code.[2] Itti et al. introduce image pyramids for the feature computations, which enables an efficient processing of real-world images. In its original version, the system concentrates on computing bottom-up attention. In newer work, Navalpakkam and Itti [2006a] introduce a derivative of the NVT, which is able to deal with top-down cues to enable visual search. Interesting to mention is that Itti and Baldi [2009] recently introduced a Bayesian model of surprise, which aims to predict eye movements. For tasks like watching video games, they found better correspondences to eye movements for the surprise model than for their saliency model.

Since the NVT belongs to the best-known and most distributed systems that exist, many groups tested it and suggested several improvements. For example, Draper and Lionelle [2005] came along with the system SAFE (selective attention as a front end), which shows several differences. For example, it does not combine the feature maps across scales but keeps them, resulting in a pyramid of saliency maps. They show that this approach is more stable with respect to geometric transformations such as translations, rotations, and reflections. Additionally, Frintrop [2005] suggested to separate the intensity feature computations into on-off and off-on computations instead of combining them in a single map and showed that certain pop-out effects are only detected by this separation. The same applies to the separation of red and green as well and blue and yellow.

The attention system of Hamker [2005; 2006] lays special emphasis on closely mimicking the neural processes in the human visual cortex. In addition to bottom-up saliency, which is similar to Itti's NVT, the system belongs to the few systems considering top-down influences. It is able to learn a target, which means it remembers the feature values of a presented stimulus. An interesting point is that Hamker's system is able to perform a very rough kind of object recognition called *match detection units*.

[2]http://ilab.usc.edu/

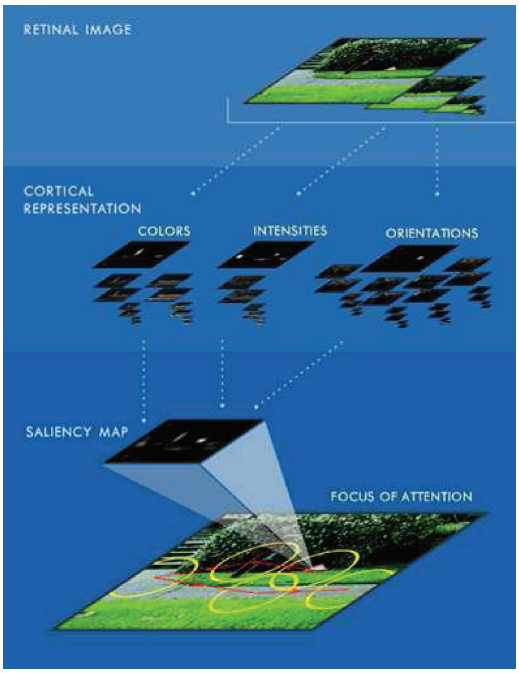


Fig. 11. Model of the Neuromorphic Vision Toolkit (NVT) by Itti et al. For each input image, image pyramids are computed to enable processing on different scales. Several feature channels investigate feature-dependent conspicuity independently. These are fused to a saliency map and a winner-take-all network determines the most salient location in this map (Reprinted with permission from http://ilab.usc.edu/).

An approach to hierarchical object-based selection of regions of interest is presented by Sun and Fisher [2003]. Regions of interest are computed on different scales, first on a coarse scale and then, if the region is sufficiently interesting, it is investigated on a finer scale. This yields foci of attention of different extents.

Backer presented an interesting model of attention with two selection stages [Backer et al. 2001; Backer 2004]. The first stage resembles standard architectures like Koch and Ullman [1985], but the result is not a single focus but a small number, usually 4, of salient locations. In the second selection stage, one of these locations is selected and yields a single focus of attention. The model investigates some of the more unregarded experimental data on multiple object tracking and object-based inhibition of return.

The system VOCUS of Frintrop [2005] has several aspects that make it well suitable for applications in computer vision and robotics. The top-down part enables an easy, user-friendly search for target objects. The system is largely robust to illumination and viewpoint changes, and it is real-time capable (50ms per frame for a 400 $\times$ 300 pixel image on a 2.8GHz PC) [Frintrop et al. 2007].

3.6 The Evaluation of Computational Attention Systems

There are mainly two possibilities to evaluate computational attention systems. First, the obtained saliency maps can be compared with the results from psychophysical experiments to determine how well the systems simulate human behavior. Second, one can evaluate how well systems perform a certain task, how they compare to standard algorithms for these tasks, and how different systems compare to each other.

Several groups have compared the performance of bottom-up attention systems with human eye movements. These evaluations are not trivial, since there is a high variability between scanpaths of different subjects and, in free-viewing tasks, there is usually no "best" scanpath. This variability may partly be explained by the fact that in human attention, always top-down cues like motivations, emotions, and preknowledge influence the processing. Easiest is the evaluation on simple, artificial scenes containing pop-outs, such as the one in Figure 3. Here, it is clear what the most salient spot is and most computational systems perform well in finding these pop-outs immediately (see Frintrop [2005]).

Several groups have also compared the correspondence of saliency models with eye movements for natural scenes. Parkhurst et al. [2002] reported a significant coherence of human eye movements with a computational saliency map, which was highest for the initial fixation. Especially high correspondence was found for fixations that followed stimulus onset. The correspondence was higher for artificial images like fractals than for natural images, probably because the top-down influence is lower for artificial scenes. Tatler et al. [2005] discovered that features such as contrast, orientation energy, and chromaticity all differ between fixated and nonfixated locations. The consistency of fixated locations between participants was highest for the first few fixations. In Tatler et al. [2006], they state that especially short saccades are dependent on the image features while long saccades are less so. It may also be noted that the first fixations of subjects who have the task of viewing scenes on a monitor tend to be clustered around the middle of the screen. This is called the *central bias*. While a final explanation is still to be found, Tatler [2007] provides several results and an interesting discussion on this topic. Probably the broadest evaluation of bottom-up saliency was presented by Elazary and Itti [2008]. They used the LabelMe database, which contained 24,836 photographs of natural scenes in which objects were manually marked and labeled by a large population of users. They found that the hot spots in the saliency map predict the locations of objects significantly above chance.

Henderson et al. [2007] investigated the influence of visual saliency on fixated locations during active search. They compared predictions from the bottom-up saliency model of Itti and Koch with fixation sequences of humans and concluded that the evidence for the visual saliency hypothesis in active visual search is relatively weak. This is not surprising, since obviously top-down cues are essential in active search. Attention systems able to perform active search, for example, Navalpakkam et al. [2005] or Frintrop [2005], are likely to achieve a larger correspondence in such settings. Other work comparing computational saliency with human visual attention is presented in Ouerhani et al. [2004], Bruce and Tsotsos [2005b], Itti [2005], Peters et al. [2005], and Peters and Itti [2008]. For example, Peters and Itti [2008] compared human eye movements with the prediction of a computational attention system in video games.

Several people have investigated how strongly the separate feature channels correspond to eye movements. Parkhurst et al. [2002] found that not one channel is generally superior to the others, but that the relative strength of each feature dimension depends on the image type: For fractal and home interior images, color was superior; for natural landscapes, buildings and city scenes, intensity was dominant. Color and intensity contributed in general more than orientation, but for buildings and city scenes, orientation was superior to color. Frey et al. [2008] found such a dependency of performance on different categories. While color had almost no influence on overt attention for some categories like faces, there is a high influence for images from other categories (e.g., Rainforest). This is especially interesting, since there is evidence that it is the rainforest where the trichromatic color vision evolved [Sumner and Mollon 2000]. Furthermore, Frey et al. [2008] found that the saliency model they investigated (Itti's NVT) exhibits good prediction performance of eye movements in more than half of the investigated categories. Kootstra et al. [2008] found that symmetry is a better predictor for human eye movements than contrast. Tatler et al. [2005] and Baddeley and Tatler [2006] compared the visual characteristics on images at fixated and nonfixated locations with signal detection and information theoretic techniques.

Tatler et al. [2005] state that "contrast and edge information was more strongly discriminatory than luminance or chromaticity." Baddeley and Tatler [2006] found that the mapping was dominated by high, frequency edges, and that low-frequency edges and contrast, on the other hand, had an inhibitory effect. They claim that previous correlations between fixations and contrast were simply artifacts of their correlations with edges. Color was not investigated in these experiments. In active search tasks, Vincent et al. [2007] discovered that color made the largest contribution for the search performance while edges made no important contribution. Altogether, it seems like further research is necessary to determine which features are most relevant in which settings and tasks.

Evaluating computational top-down attention systems in visual search tasks is easier than evaluating bottom-up systems, since a target is known and the detection rate for this target can be determined. As in human perception, the performance depends on the target and on the setting. Some results can be found in Hamker [2005], Navalpakkam et al. [2005], Frintrop [2005], Vincent et al. [2007]. For example, in Frintrop [2005], a target object in natural environments was in most cases found with the first fixation (e.g., a fire extinguisher in a corridor or a key fob on a desk). Vincent et al. [2007] have found a relatively low-fixation probability for real-world targets in their approach, especially for difficult search targets, such as a wine glass. These approaches are difficult to compare since they operate on different data, so it is hard to distinguish which differences come from the implementation of the system and which from the difference in data. In Frintrop [2005], we present a comparison of VOCUS with the systems in Hamker [2005] and Navalpakkam et al. [2005], each on the same image data sets.

Another possibility to evaluate the quality of attentional systems is their use in applications. If the system performance is increased in either time or quality, it is not necessarily important to achieve exact correspondences to human eye movements. Several application domains of visual attention systems will be presented in the next section.

4. APPLICATIONS IN COMPUTER VISION AND ROBOTICS

Restricting the large amount of visual data to a manageable rate has been an omnipresent topic during the last years in research areas concerned with image data. Although machines became much faster and hardware cheaper, processing all information is still not possible and will not be possible in the future. The reason is that the complexity of many problems is very high—as mentioned before, unbounded visual search is NP-complete—so finding a polynomial solution for such a problem is extremely unlikely.

Therefore, concepts like selective visual attention arouse much interest in computer vision and robotics. They provide an intuitive method to determine the most interesting regions of an image in a "natural," human-like way and are a promising approach to improve computational vision systems.

We organize the applications of computational attention systems roughly into three categories: In the first, low-level category, attentional regions are used as low-level features, so called interest points or regions of interest (ROIs) for tasks such as image matching (Section 4.1). The second, mid-level category considers attention as a front-end for high-level tasks as object recognition (Section 4.2). In the third, highest-level category, attention is used in a human-like way to guide the action of an autonomous system like a robot, that is, to guide object manipulation or human-robot interaction (Section 4.3).

4.1 Attention as Salient Interest Point Detector

Detecting regions of interest is an important method in many computer vision tasks. Many methods exist to detect interest points or regions in images, an overview is provided by Tuytelaars and Mikolajczyk [2007]. An alternative to these approaches are attention regions. While common detectors usually work on gray-scale images, computational attention systems integrate several features and determine the overall saliency from many cues. Another difference is that attention systems focus on a few highly discriminative features, while common detectors often tend to find many similar regions. Depending

on the application, the restriction to a few discriminative regions is favorable because it reduces computation complexity. We have shown that the repeatability of regions in different scenes is significantly higher for salient regions than for regions detected by standard detectors [Frintrop 2008].[3]

One application area of salient ROIs is *image segmentation*. Segmentation is the problem of grouping parts of an image together according to some measure of similarity. The automatic segmentation of images into regions usually deals with two major problems: First, setting the starting points for segmentation (seeds), and second, choosing the similarity criterion to segment regions. Ouerhani [2003] presents an approach that supports both aspects by visual attention: The saliency spots of the attention system serve as natural candidates for the seeds and the homogeneity criterion is adapted according to the features that discriminate a region from its surrounding. A comparison to other segmentation algorithms has, to our knowledge, not yet been done.

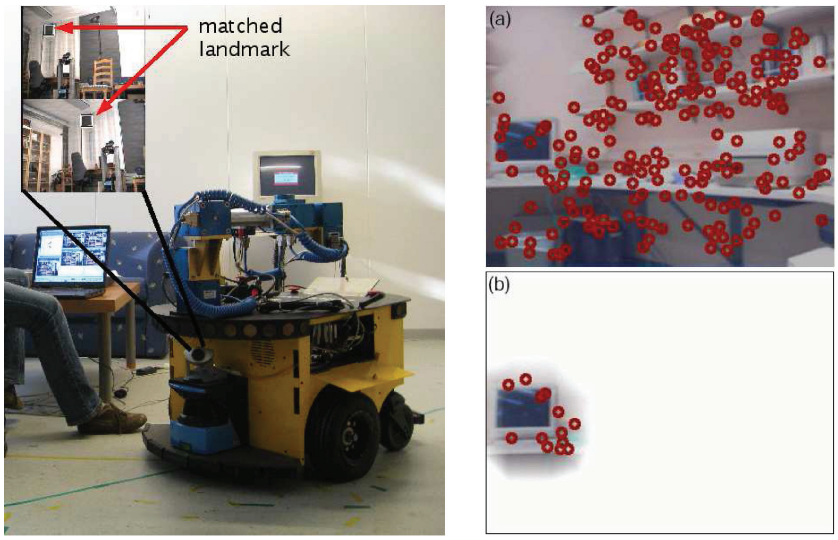
Another application area is *image and video compression*. The idea is to compress nonfocused regions stronger than focused ones, based on the findings that there is correspondence between the regions focused by humans and those detected by computational attention systems. Ouerhani [2003] performs focused image compression with a visual attention system. A color image compression method adaptively determines the number of bits to be allocated for coding image regions according to their saliency. Regions with high saliency have a higher reconstruction quality than less salient regions. Itti [2004] uses his attention system to perform video compression by blurring every frame, increasingly with distance from salient locations.

A large field with many application areas is *image matching*, that is, finding correspondences between two or more images that show the same scene or the same object. When searching for correspondences between two images, it is computationally too expensive to compare images on a pixel basis and variations in illumination and viewpoint make such a simple approach unsuitable. Instead, ROIs can be used to find such correspondences. This is necessary for tasks like stereo matching, building panoramas, place recognition, or robot localization.

To compare two ROIs, a descriptor is required. Attentional descriptors are vectors that determine the feature saliencies of the ROI and its surrounding (see Section 3.1) [Navalpakkam et al. 2005; Frintrop et al. 2005]. Since matching with an attentional descriptor alone is usually not powerful enough, several groups have combined their attention regions with other detectors or descriptors. A common approach is the SIFT descriptor (scale invariant feature transform), which captures the gradient magnitude in the surrounding of a region [Lowe 2004]. It is also very powerful under image transformations. Walther [2006] and Siagian and Itti [2009] detect SIFT keypoints (intensity extrema in scale space and combined with a SIFT descriptor) inside the attention regions, that is, the attentional regions determine a search area whereas the matching is based on the SIFT keypoints. Note, however, that this approach is sometimes problematic, since attention regions, favor homogeneous regions, whereas corner features are usually detected at textured areas. Thus, the combination often results in very few features, which makes matching difficult. In our work, we obtained better results by directly applying a SIFT descriptor to the attention regions [Frintrop and Jensfelt 2008].

One application scenario in which image matching is used is robot localization. Based on a known map of the surrounding, the robot has to determine its position in this map by interpreting its sensor data. Standard approaches for such problems use range sensors, such as laser scanners, and there are good and stable solutions for such problems. However, in outdoor environments and open areas, the standard methods for localization are likely to fail. Instead, a promising approach is localization by detecting visual landmarks with a known position. Attentional mechanisms can facilitate the search of landmarks during operation by selecting interesting regions in the sensor data. An early project that followed this

[3]See also http://www.informatik.uni-bonn.de/~frintrop/research/saliency.html



(a) Robot localization and mapping (b) Object recognition

Fig. 12. Two application scenarios for visual attention systems: (a) Robot localization and mapping: Robot Dumbo corrects its position estimate by detecting a landmark that it has seen before. Landmark detection is done with the attention system VOCUS. The top-left corner shows the currently seen frame (top) and the frame from the database (bottom) with the matched landmark [Frintrop and Jensfelt 2008]. (b) Object recognition: top: SIFT keypoints are extracted for the whole image. Bottom: Attentional regions of interest restrict the keypoints to regions, which are likely to contain objects. This enables unsupervised learning in cluttered scenes (Figure reprinted with permission from [Walther 2006]).

approach was the ARK project [Nickerson et al. 1998]. It relied on hand-coded maps, including the locations of known static obstacles, as well as the locations of natural visual landmarks. Ouerhani et al. [2005] track salient spots over time and use them as landmarks for robot localization. The results must be considered preliminary, since testing was done on the training sequence on a straight corridor without loops. *Scene classification* and global localization based on salient landmarks is presented in Siagian and Itti [2009]. In addition to the landmarks, the authors use the "gist" of the scene, a feature vector, which captures the appearance of the scene, to obtain a coarse localization hypothesis.

In the previously described examples, a map of the environment is initially known. Usually, it is obtained in a training phase. A more difficult task is *simultaneous localization and mapping (SLAM)*, in which a robot initially does not know anything about its environment and has to build a map and localize itself inside the map at the same time. Until now, this topic was rarely investigated in combination with visual attention. Frintrop and Jensfelt [2008] investigated the combination of visual attention and SLAM. The salient regions are detected with the attention systems VOCUS, matched with a SIFT descriptor, and tracked over several frames to obtain a 3D position of the landmarks. Finally, they are matched to database entries of the landmarks to detect if the robot closed a loop (i.e., returned to a previously visited area [see Figure 12(a)]).

In addition to the presented application areas, image matching with attentional ROIs is sometimes also used for object recognition. This aspect will be described in the next section.

4.2 Attention as Front-End for Object Recognition

Probably the most suggestive application of an attention system is object recognition, since the two-stage approach of a preprocessing attention system and a classifying recognizer mimics human perception

[Neisser 1967]. Miau et al. [2001] present a biologically motivated approach that combines an attentional front-end with the biologically motivated object recognition system HMAX [Riesenhuber and Poggio 1999], which simulates processes in human cortex and has rather limited capabilities. It is restricted to recognize simple artificial objects like circles or rectangles. Miau et al. [2001] also replaced the HMAX system by a support vector machine to detect pedestrians in natural images. This approach is much more powerful with respect to the recognition rate but computationally expensive.

Salah et al. [2002] combine an attention system with neural networks and an observable Markov model for handwritten digit recognition and face recognition, and Ouerhani [2003] presents an attention-based traffic sign recognition system. In Frintrop et al. [2004], we have combined an attention system with an AdaBoost-based object classifier [Viola and Jones 2004], which was trained for objects in laser scanner data. Walther [2006] combines an attention system with an object recognizer based on SIFT features [Lowe 2004] and shows that the recognition results are improved by the attentional front-end (see Figure 12(b)).

All of these systems rely only on bottom-up information and, therefore, on the assumption that the objects of interest are sufficiently salient by themselves. Nonsalient objects are not detected. For some object classes such as traffic signs, which are intentionally designed salient, this works quite well; for other applications, top-down information is needed to enable the system to focus on the desired objects. A combination of a top-down modulated computational attention system with a classifier is presented by Mitri et al. [2005]. Here, the attention system VOCUS generates object hypotheses, which are verified or falsified by a classifier. For the application of ball detection in the robot soccer scenario ROBOCUP,[4] the amount of false detections is reduced significantly.

In the previously mentioned approaches, the attentional part is separated from the object recognition; both systems work independently. In human perception, these processes are strongly intertwined. A few groups have recently started to work on approaches in which both processes share resources. Hamker [2005] introduces *match detection units* that compare the encoded pattern with the target template. If these patterns are similar, an eye movement is initiated toward this region and the target is said to be detected. Currently, results have to be considered conceptual, since recognition does not consider spatial configuration of features and recognizes only patterns that are presented with the same orientation as during learning. An interesting approach is presented by Walther and Koch [2007]. The authors suggest a unifying framework for object recognition and attention. It is based on the HMAX model for object recognition and modulates the activity by spatial and feature modulation functions, which suppress or enhance locations or features due to spatial attention.

Another interesting approach is provided by Rybak et al. [1998]: Although the attentional part of their system is rather limited (it uses only one feature [orientation] and no target-specific tuning of the feature computations), they present a sophisticated approach to investigate an image guided by prior knowledge. In a memorizing mode, a sequence of fixation points is determined and stored in two kinds of memories: The sensory memory ("what"-structure) stores the features of the fixations, and the motor memory ("where"-structure) stores the relative shifts between the fixations. This information is used in search mode to guide the visual search and to compare the stored fixation patterns with the current image.

A different view on attention for object recognition is presented by Fritz et al. [2004]: An information-theoretic saliency measure is used to determine discriminative regions of interest in objects. The saliency measure is computed by the conditional entropy of estimated posteriors of the local appearance patterns. That means regions of an object are considered as salient if they discriminate the object from other objects in an object database. A similar approach pursues Pessoa and Exel [1999].

[4]http://www.robocup.org

4.3 Attention Systems for Guiding Robot Action

A robot that has to act in a complex world faces the same problems as a human: It has to decide what to do next. Because of limited resources, usually only one task can be performed at a time: The robot can only manipulate one object, it can only follow one object with the camera, and it can only interact with one person at the same time (even if these capabilities could be slightly extended by additional hardware to a few parallel tasks, such extensions are very limited). Thus, even if computational power would allow us to find all correspondences, to recognize all objects in an image and process everything of interest, it would still be necessary to filter out the relevant information to determine the next action. This decision is based first on the current sensor input and second on the internal state, for example, the current tasks and goals.

A topic in which the decision about the next action is intrinsically based on visual data is *active vision*, that is, the problem of where to look next. It deals with controlling "the geometric parameters of the sensory apparatus ... in order to improve the quality of the perceptual results" [Aloimonos et al. 1988]. Thus, it is the technical equivalent for overt attention: It directs the camera to regions of potential interest as the human visual system directs the gaze. Active vision is of special interest in robotics: It makes "vision processing more robust and more closely tied to the activities that a robotic system may be engaged in" [Clark and Ferrier 1989].

One of the first approaches to realize an active vision system with the help of visual attention was presented by Clark and Ferrier [1988]. They describe how to steer a binocular robotic head with visual attention and perform simple experiments to fixate and track the most salient region in artificial scenes composed of geometric shapes. Mertsching et al. [1999] and Bollmann [1999] use the neural active vision system NAVIS once with a fixed stereo camera head and once on a mobile robot with a monocular camera head. Vijayakumar et al. [2001] present an attention system, which is used to guide the gaze of a humanoid robot. The authors consider only one feature, visual flow, which enables the system to attend to moving objects. To simulate the different resolutions of the human eye, two cameras per eye are used: one wide-angle camera for peripheral vision and one narrow-angle camera for foveal vision. Dankers et al. [2007] introduced an architecture for reactive visual analysis of dynamic scenes as part of an active stereo vision system. Saliency is computed for each camera separately. Active gaze control for simultaneous robot localization and mapping was recently presented in Frintrop and Jensfelt [2008]. The robot actively controls the camera by switching between the behaviors tracking, redetection, and exploration. Thus, it obtains a better distribution of landmarks and facilitates the redetection of landmarks.

Many of the above examples include the *visual tracking* problem, that is, the problem of consistently following a region or object over several frames. The problem becomes difficult if illumination changes, if the object is partially and/or temporary occluded and if not only the object or the camera, but also both of them are mobile. Walther et al. [2004] track objects in underwater videos by detecting them with a bottom-up attention system and tracking them with Kalman filters. Currently, we investigate general object tracking based on visual attention [Frintrop and Kessel 2009]. The appearance of an object is quickly learned from a single frame and the most salient part of the person is redetected with top-down directed attention in subsequent frames. An extension of this work deals with people tracking from a mobile platform, an important task for service robots [Frintrop et al. 2010].

Another area in which the visual input determines the next action is *object manipulation*. A robot that has to grasp and manipulate objects has to detect and probably also recognize the object first. Attentional mechanisms can support these tasks. For example, Bollmann et al. [1999] present a robot that uses the active vision system NAVIS to play at dominoes. In Rae [2000], a robot arm has to grasp an object a human has pointed at. The group around Tsotsos is working on a smart wheelchair to support

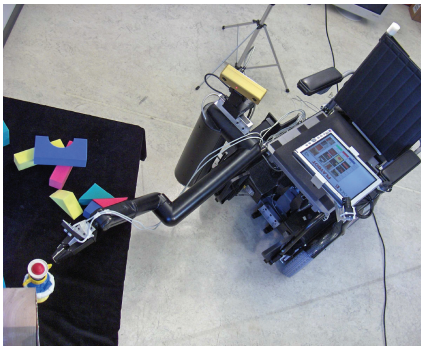


Fig. 13. PlayBot: A visually guided robotic wheelchair for disabled children. The selective tuning model of visual attention supports the detection of objects of interest (Figure reprinted with permission from http://www.cse.yorku.ca/~playbot).

disabled children. The wheelchair has a display with an easily accessible user interface which shows pictures of places and toys. Once a task like "go to table, point to toy" is selected, the system drives to the selected location and searches for the specified toy, using mechanisms based on a visual attention system (see Figure 13) [Tsotsos et al. 1998; Rotenstein et al. 2007].

In the field of *robot navigation*, the problem of *visual servoing* has become a well-established robot control technique, which integrates vision in feedback control loops. The technique is mainly employed for controlling the robot's position. Clark and Ferrier [1992] describe how to realize a visual servo control system, which implements attentive control of a binocular vision system. Results on simple artificial scenes in which the most salient region is fixated and tracked are shown in Clark and Ferrier [1988]. In Scheier and Egner [1997], a mobile robot uses an attention system to approach large objects. Since larger objects have a higher saliency, only the regions with the highest saliency have to be approached. In Baluja and Pomerleau [1997], an attention system is used to support autonomous road following by highlighting relevant regions in a saliency map. Borji [2009] investigates the control of motor commands for an artificial agent in a navigation scenario by reinforcement learning. The current state of the system is derived from object and scene recognition at the focus of attention.

Finally, *human-robot interaction* is an intuitive application area for computational attention systems. If robots shall purposefully interact with humans, it is convenient if both attend to the same object or region of interest. A computational attention system similar to the human one can help a robot to focus on the same region as a human. Breazeal [1999] introduces a robot that shall actively look at people or toys. Although top-down information would be necessary to focus on a particular object relevant for a certain task, bottom-up information can be useful, too, if it is combined with other cues. For example, Heidemann et al. [2004] combine an attention system with a system that follows the direction of a pointing finger and can adjust to the selected region accordingly. This approach was used by Rae [2000] to guide a robot arm towards an object and grasp it. Belardinelli [2008] presents methods to let a robot learn visual scene exploration by imitating human gaze shifts. Nagai [2009] developed an action learning model based on spatial and temporal continuity of bottom-up features. Finally, an interesting sociological study in which the interaction of a human with a robot simulation is investigated is presented by Muhl et al. [2007]. Human subjects had to show an object to a robot face on a screen, which attended to the object with help of a visual attention system. If the robot was artificially diverted and directed its gaze away from the object, humans tried to reobtain the robots attention by waving hands, making noise, or approaching the robot. This shows that people established a communicative space with the robot and accepted it as a social partner.

5. DISCUSSION AND CONCLUSION

This article gives a broad overview of computational visual attention systems and their cognitive foundations and aims to bridge the gap between different research areas. Visual attention is a highly interdisciplinary field and the disciplines investigate the area from different perspectives. Psychologists usually investigate human behavior on special tasks to understand the internal processes of the brain, often resulting in psychophysical theories or models. Neurobiologists take a view directly into the brain with new techniques such as functional magnetic resonance imaging (fMRI). These methods visualize which brain areas are active under certain conditions. Computer scientists use the findings from psychology and biology to build improved technical systems.

During the last years, the different disciplines have profited considerably from each other. Psychologists refer to neurobiological findings to improve their attention models and neurobiologists consider psychological experiments to interpret their data. Additionally, more and more psychologists implement their models computationally or refer to computational models to verify if the behavior of the systems equals human perception. These findings help to improve the understanding of the mechanisms and can also lead to improved computational systems.

Of course, in all of the three areas presented in this article, namely human attention, computational systems, and applications, there are still many open questions. Let us try to address some of them.

One important question is, what are the basic features of attention? Although intensively studied, this question is still not fully answered (see, e.g., Wolfe and Horowitz [2004]). Other research questions relate to how these features interact. The theory that peak salience computed from local feature contrast maxima in several feature dimensions determine human fixations has been questioned in some articles. For example, the correlations between local image statistics and the locations of human fixations have been investigated, leading to new hypotheses, for instance, that high spatial frequency edges guide attention rather than contrast in other feature dimensions [Baddeley and Tatler 2006]. These new ideas require more investigation.

Other questions concern the nature of top-down cues and processes. Visual search in artificial search arrays has been well investigated and also studies on natural images have been done (e.g., Peters et al. [2005]). For both, especially for the research on natural scenes, certain open questions remain. A still largely unexplored area is the investigation of visual perception in dynamic scenes (an exception is Peters and Itti [2008]) and, even more challenging, during interactions of humans in the real-world (e.g., Land [2006]). Additionally, top-down influences are not limited to target search. Other cues, such as prior knowledge, motivations, and emotions, influence the visual system and are worth being investigated further. Also interesting are questions such as "how much learning is involved in visual processing?", "how does context influence the search?", and "how much memory is involved in these mechanisms?" Some current findings on these topics can be found in Kunar et al. [2008]. When going beyond visual attention, questions arise like "how does visual attention interact with other senses?" [Fritz et al. 2007], "which concepts of selective attention are shared in the brain among different senses?" [Ghazanfar and Schroeder 2006], and "how do visual attention and object recognition interact?"

For computational attention systems, similar questions remain, starting from "which are the optimal features?" and "how are these features integrated?" to "how do top-down cues influence the computation?" and "how do bottom-up and top-down cues interact?" However, we want to claim here that computational systems do not necessarily have to mimic biology perfectly to achieve similar performance. A camera differs from the eye and a computer is not the brain. Even parallel hardware like multiprocessors or parallel computations on GPUs differ considerably from the architecture of neurons. Especially interesting is to find out which concepts of human perception make sense in computational systems and which have to be adapted accordingly.

Finally, concerning the applications of computational attention systems, a current challenge is to capacitate the systems to be used in the real world. That means the systems have to be robust to noise, image transformations and illumination changes, and they have to be fast enough to process images at frame rate. Robustness to noise has been shown by Itti et al. [1998], invariance to 2D similarity transformations to a large extend is achieved by Draper and Lionelle [2005], and robustness of a top-down attention system to viewpoint changes and illumination variations has been shown by Frintrop [2005]. Recently, there have been approaches to extend to the concept of 2D saliency maps to 3D [Fleming et al. 2006; Schauerte et al. 2009]. The speed of the systems has prevented real-time applications for a long time. Parallelizations on several CPUs [Itti 2002], on dedicated hardware [Ouerhani 2003], or on a GPU [May et al. 2007; Xu et al. 2009] enable a significant speed-up. Also, software solutions based on integral images have enabled real-time performance, making the systems' flexibly applicable without special hardware [Frintrop et al. 2007]. Also interesting is the investigation of how the concepts of attention apply to other sensors than cameras, for example, laser scanners (a visual attention system based on laser scanner data is presented by Frintrop et al. [2005]). More research is necessary to find out how these concepts might be adapted to best fit the properties of different sensors and how the information from different sensors may be fused.

Computational attention has gained significantly in popularity over the last decade. First of all, adequate computational resources are now available to study attentional mechanisms with a high degree of fidelity. In addition, a large number of cognitive projects have been launched, particularly in Europe. Good examples include MACS, CogVis, POP, and SEARISE.[5] In most of these approaches, visual attention is included in the perception module and helps to deal with the complexity of the real world. Over the next few years, a number of embodied cognitive agents will be studied as part of new generation systems both in Europe and in the United States. The European efforts are part of the emphasis on cognitive systems, whereas the United States efforts are part of the NSF Cyber Physical Systems program [Lee 2008]. As vision systems are integrated into complete systems, the need for optimization of the visual process in terms of overt and covert attention becomes more explicit. In addition, the interplay between attention and tasking can be studied more explicitly. The more complex the systems and their tasks become, the more urgent the need for a preselecting attention system, which determines in advance the regions of highest potential interest in the sensor data.

REFERENCES

Abdi, H. 2007. Signal detection theory (SDT). In *Encyclopedia of Measurement and Statistics*, N. Salkind, Ed. Sage, CA.

Aloimonos, Y., Weiss, I., and Bandopadhay, A. 1988. Active vision. *Int. J. Comput. Vision 1,* 4, 333–356.

Aristotle. On Sense and the Sensible. The Internet Classics Archive, 350 B.C.E., translated by J. I. Beare.

Awh, E. and Pashler, H. 2000. Evidence for split attentional foci. *J. Exp.Psych. Hum. Percept. Perform. 26*, 2, 834–846.

Aziz, M. Z. and Mertsching, B. 2007. Pop-out and IOR in static scenes with region-based visual attention. In *Proceedings of the ICVS Workshop on Computational Attention and Applications*. ACM, New York.

Backer, G. 2004. Modellierung visueller Aufmerksamkeit im Computer-Sehen: Ein zweistufiges Selektionsmodell f¨ur ein Aktives Sehsystem. Ph.D. thesis, Universität Hamburg, Germany.

Backer, G., Mertsching, B., and Bollmann, M. 2001. Data- and model-driven gaze control for an active-vision system. *IEEE Trans. Pattern Anal. Mach. Intell. 23*, 12, 1415–1429.

Bacon, W. and Egeth, H. 1994. Overriding stimulus-driven attentional capture. *Percept. Psychophysics 55*, 5, 485–496.

Baddeley, R. J. and Tatler, B. W. 2006. High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis. *Vision Res. 46*, 2824–2833.

Balkenius, C. 2000. Attention, habituation and conditioning: towards a computational model. *Cognitive Sci. Q. 1*, 2, 171–214.

Baluja, S. and Pomerleau, D. 1997. Expectation-based selective attention for visual monitoring and control of a robot vehicle. *Rob. Auton. Syst. 22*, 3-4, 329–344.

[5]http://cordis.europa.eu/ist/cognition/projects.htm#list

Belardinelli, A. 2008. Salience features selection: Deriving a model from human evidence. Ph.D. thesis, Sapienza Universita di Roma, Rome, Italy.

Ben-Shahar, O., Scholl, B., and Zucker, S. 2007. Attention, segregation, and textons: Bridging the gap between object-based attention and texton-based segregation. *Vision Res. 47*, 6, 173–178.

Bichot, N. P. 2001. Attention, eye movements, and neurons: Linking physiology and behavior. In *Vision and Attention*, M. Jenkin and L. R. Harris, Eds. Springer Verlag, Berlin.

Bichot, N. P., Rossi, A. F., and Desimone, R. 2005. Parallel and serial neural mechanisms for visual search in macaque area V4. *Science 308*, 5721, 529–534.

Bisley, J. and Goldberg, M. 2003. Neuronal activity in the lateral intraparietal area and spatial attention. *Science 299*, 5603, 81–86.

Björkman, M. and Eklundh, J.-O. 2007. Vision in the real world: Finding, attending and recognizing objects. *Int. J. Imaging Syst. Technol. 16*, 2, 189–208.

Bollmann, M. 1999. Entwicklung einer Aufmerksamkeitssteuerung für ein aktives Sehsystem. Ph.D. thesis, Universität Hamburg, Germany.

Bollmann, M., Hoischen, R., Jesikiewicz, M., Justkowski, C., and Mertsching, B. 1999. Playing domino: A case study for an active vision system. In *Computer Vision Systems*, H. Christensen, Ed. Springer, Berlin, 392–411.

Borji, A. 2009. Interactive learning of task-driven visual attention control. Ph.D. thesis, Institute for Research in Fundamental Sciences (IPM), School of Cognitive Sciences (SCS), Tehran, Iran.

Breazeal, C. 1999. A context-dependent attention system for a social robot. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI 99)*. Springer, Berlin, 1146–1151.

Bruce, N. D. B. and Tsotsos, J. K. 2005a. An attentional framework for stereo vision. In *Proceedings of the Canadian Conference on Computer and Robot Vision*. IEEE, Los Alamitos.

Bruce, N. D. B. and Tsotsos, J. K. 2005b. Saliency-based on information maximization. In *Proceedings of Neural Information Processing Systems (NIPS)*. MIT Press, Cambridge, MA.

Bundesen, C. 1990. A theory of visual attention. *Psych. Rev. 97*, 523–547.

Bundesen, C. 1998. A computational theory of visual attention. *Philos. Trans. R. Soc., Series B 353*, 1271–1281.

Bundesen, C. and Habekost, T. 2005. Attention. In *Handbook of Cognition*, K. Lamberts and R. Goldstone, Eds. Sage Publications, London.

Bur, A., Wurtz, P., Müri, R., and Hügli, H. 2007. Motion integration in visual attention models for predicting simple dynamic scenes. In *Proceedings of the SPIE Conference on Human Vision and Electronic Imaging.* Springer, Berlin.

Cameron, E., Tai, J., Eckstein, M., and Carrasco, M. 2004. Signal detection theory applied to three visual search tasks. *Spatial Vision 17*, 4–5.

Carrasco, M., Evert, D. L., Chang, I., and Katz, S. M. 1995. The eccentricity effect: Target eccentricity affects performance on conjunction searches. *Percept. Psychophysics 57*, 8, 1241–1261.

Cassin, B. and Solomon, S. 1990. *Dictionary of Eye Terminology*. Triad Publishing Company, Gainsville, FL.

Cave, K. R. 1999. The Feature Gate model of visual selection. *Psych. Res. 62*, 182–194.

Cave, K. R. and Wolfe, J. M. 1990. Modeling the role of parallel processing in visual search. *Cognitive Psych. 22*, 2, 225–271.

Cherry, E. C. 1953. Some experiments on the recognition of speech, with one and two ears. *J. Acoust. Soc. Am. 25*, 975–979.

Choi, S.-B., Ban, S.-W., and Lee, M. 2004. Biologically motivated visual attention system using bottom-up saliency map and top-down inhibition. *Neural Inform. Process. Lett. Rev. 2*, 1.

Chun, M. M. and Jiang, Y. 1998. Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psych. 36*, 28–71.

Clark, J. J. and Ferrier, N. J. 1988. Modal control of an attentive vision system. In *Proceedings of the 2nd International Conference on Computer Vision*. IEEE, Los Alamitos, CA.

Clark, J. J. and Ferrier, N. J. 1989. Control of visual attention in mobile robots. In *Proceedings of the IEEE Conference on Robotics and Automation*. IEEE, Los Alamitos, CA, 826–831.

Clark, J. J. and Ferrier, N. J. 1992. Attentive visual serving. In *An Introduction to Active Vision*, A. Blake and A. Yuille, Eds. MIT Press, Cambridge, MA.

Connor, C. E., Egeth, H. E., and Yantis, S. 2004. Visual attention: Bottom-up versus top-down. *Curr. Biol. 14*.

Corbetta, M. 1990. Frontoparietal cortical networks for directing attention and the eye to visual locations: Identical, independent, or overlapping neural systems? In *Proceedings of the National Academy of Sciences of the United States of America 95*, 831–838.

CORBETTA, M. AND SHULMAN, G. L. 2002. Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. 3*, 3, 201–215.

DANKERS, A., BARNES, N., AND ZELINSKY, A. 2007. A reactive vision system: Active-dynamic saliency. In *Proceedings of the 5th International Conference on Computer Vision Systems (ICVS'07)*. IEEE, Los Alamitos, CA.

DESIMONE, R. AND DUNCAN, J. 1995. Neural mechanisms of selective visual attention. *Ann. Rev. Neurosci. 18*, 193–222.

DEUBEL, H. AND SCHNEIDER, W. X. 1996. Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Res. 36*, 12, 1827–1837.

DRAPER, B. A. AND LIONELLE, A. 2005. Evaluation of selective attention under similarity transformations. *J. Comput. Vision Image Understanding 100*, 1-2, 152–171.

DRIVER, J. AND BAYLIS, G. C. 1998. Attention and visual object segmentation. In *The Attentive Brain*, R. Parasuraman, Ed. MIT Press, Cambridge, MA, 299–326.

DUNCAN, J. 1984. Selective attention and the organization of visual information. *J. Exp. Psych. 113*, 501–517.

ECKSTEIN, M., THOMAS, J., PALMER, J., AND SHIMOZAKI, S. 2000. A signal detection model predicts the effects of set size on visual search accuracy for feature, conjunction, triple conjunction, and disjunction displays. *Percept. Psychophys. 62*, 3, 425–451.

EGETH, H. E. AND YANTIS, S. 1997. Visual attention: Control, representation, and time course. *Ann. Rev. Psych. 48*, 269–297.

EINHÄUSER, W., SPAIN, M., AND PERONA, P. 2008. Objects predict fixations better than early saliency. *J. Vision 8*, 14, 1–26.

ELAZARY, L. AND ITTI, L. 2008. Interesting objects are visually salient. *J. Vision 8*, 3:3, 1–15.

ERIKSEN, C. W. AND ST. JAMES, J. D. 1986. Visual attention within and around the field of focal attention: A zoom lens model. *Percept. Psychophys. 40*, 225–240.

FINDLAY, J. M. AND GILCHRIST, I. D. 2001. Active vision perspective. In *Vision & Attention*, M. Jenkin and L. R. Harris, Eds. Springer Verlag, Berlin, 83–103.

FINDLAY, J. M. ANDWALKER, R. 1999. A model of saccade generation based on parallel processing and competitive inhibition. *Behav. Brain Sci. 22*, 661–721.

FINK, G., DOLAN, R., HALLIGAN, P., MARSHALL, J., AND FRITH, C. 1997. Space-based and object-based visual attention: Shared and specific neural domains. *Brain 120*, 11, 2013–2028.

FLEMING, K. A., PETERS II, R. A., AND BODENHEIMER, R. E. 2006. Image mapping and visual attention on a sensory ego-sphere. In *Proceedings of the Conference on Intelligent Robots and Systems (IROS)*. IEEE, Los Alamitos, CA, 241–246.

FRAGOPANAGOS, N. AND TAYLOR, J. 2006. Modelling the interaction of attention and emotion. *Neurocomputing 69*, 16–18, 1977–1983.

FRAUNDORFER, F. AND BISCHOF, H. 2003. Utilizing saliency operators for image matching. In *Proceedings of the International Workshop on Attention and Performance in Computer Vision*. Springer, Berlin, 17–24.

FREY, H.-P., HONEY, C., AND KÖNIG, P. 2008. What's color got to do with it? The influence of color on visual attention in different categories. *J. Vision 8*, 14, 1–17.

FRINTROP, S. 2005. VOCUS: A visual attention system for object detection and goal-directed search. Ph.D. thesis, Rheinische Friedrich-Wilhelms-Universität Bonn, Germany. Lecture Notes in Artificial Intelligence, vol. 3899, Springer Verlag.

FRINTROP, S. 2008. The high repeatability of salient regions. In *Proceedings of Workshop on Efficient Strategies for Cognitive Agents in Complex Environments*. Springer, Berlin.

FRINTROP, S., BACKER, G., AND ROME, E. 2005. Goal-directed search with a top-down modulated computational attention system. In *Proceedings of the Annual Meeting of the German Association for Pattern Recognition*. Springer, Berlin.

FRINTROP, S. AND CREMERS, A. B. 2007. Top-down attention supports visual loop closing. In *Proceedings of the European Conference on Mobile Robotics*. Springer, Berlin.

FRINTROP, S. AND JENSFELT, P. 2008. Attentional landmarks and active gaze control for visual SLAM. *IEEE Trans. Rob. 24*, 5.

FRINTROP, S. AND KESSEL, M. 2009. Most salient region tracking. In *Proceedings of the IEEE International Conference on Robotics and Automation*. IEEE, Los Alamitos, CA.

FRINTROP, S., KLODT, M., AND ROME, E. 2007. A real-time visual attention system using integral images. In *Proceedings of the 5th International Conference on Computer Vision Systems*.

FRINTROP, S., KÖNIGS, A., HOELLER, F., AND SCHULZ, D. 2010. A component-based approach to visual person tracking from a mobile platform. *Int. J. So. Rob.*

FRINTROP, S., NÜCHTER, A., SURMANN, H., AND HERTZBERG, J. 2004. Saliency-based object recognition in 3D data. In *Proceedings of the International Conference on Intelligent Robots and Systems*. IEEE, Los Alamitos, CA, 2167–2172.

FRINTROP, S., ROME, E., NÜCHTER, A., AND SURMANN, H. 2005. A bimodal laser-based attention system. *J. Comput. Vision Image Understand. 100*, 1–2, 124–151.

Fritz, G., Seifert, C., and Paletta, L. 2004. Attentive object detection using an information theoretic saliency measure. In *Proceedings of the 2nd International Workshop on Attention and Performance in Computational Vision*. Springer, Berlin, 136–143.

Fritz, J. B., Elhilali, M., David, S. V., and Shamma, S. A. 2007. Auditory attention focusing the searchlight on sound. *Curr. Opin. Neurobiol. 17*, 437–455.

Garey, M. and Johnson, D. S. 1979. *Computers and Intractability, A Guide to the Theory of NP-Completeness*. Freeman, San Francisco.

Gegenfurtner, K. R. 2003. Cortical mechanisms of color vision. *Nat. Rev. Neurosci. 4*, 563–572.

Ghazanfar, A. and Schroeder, C. 2006. Is neocortex essentially multisensory? *Trends Cogn. Sci. 10*, 278–285.

Giesbrecht, B., Wodorff, M., Song, A., and Mangun, G. 2003. Neural mechanisms of topdown control during spatial and feature attention. *Neuroimage 19*, 496–512.

Gottlieb, J. P., Kusunoki, M., and Goldberg, M. E. 1998. The representation of visual salience in monkey parietal cortex. *Nature 391*, 481–484.

Green, D. M. and Swets, J. A. 1966. *Signal Detection Theory and Psychophysics*. Wiley, New York.

Hamker, F. H. 2005. The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision. *J. Comput. Vision Image Understanding 100*, 1–2, 64–106.

Hamker, F. H. 2006. Modeling feature-based attention as an active top-down inference process. *BioSystems 86*, 91–99.

Harel, J., Koch, C., and Perona, P. 2007. Graph-based visual saliency. In *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. MIT Press, Cambridge, MA, 545–552.

Heidemann, G., Rae, R., Bekel, H., Bax, I., and Ritter, H. 2004. Integrating context-free and context-dependent attentional mechanisms for gestural object reference. *Mach. Vision Appl. 16*, 1, 64–73.

Heinke, D. and Humphreys, G. W. 2003. Attention, spatial representation and visual neglect: Simulating emergent attention and spatial memory in the selective attention for identification model (SAIM). *Psych. Rev. 110*, 1, 29–87.

Heinke, D. and Humphreys, G. W. 2004. Computational models of visual selective attention. A review. In *Connectionist Models in Psychology*, G. Houghton, Ed. Psychology Press, Florence, KY, 273–312.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., and Mack, M. 2007. Visual saliency does not account for eye movements during visual search in real-world scenes. In *Eye Movements: A Window on Mind and Brain*, R. van Gompel, M. Fischer, W. Murray, and R. Hill, Eds. Elsevier, Oxford, 537–562.

Horowitz, T. S. and Wolfe, J. M. 2003. Memory for rejected distractors in visual search? *Visual Cognition 10*, 3, 257–298.

Humphreys, G. W. and Müller, H. J. 1993. Search via recursive rejection (SERR): A connectionist model of visual search. *Cognitive Psych. 25*, 43–110.

Itti, L. 2002. Real-time high-performance attention focusing in outdoors color video streams. In *Proceedings of the SPIE Conference Human Vision and Electronic Imaging*. IEEE, Los Alamitos, CA.

Itti, L. 2004. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Trans. Image Process. 13*, 10.

Itti, L. 2005. Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition 12*, 6, 1093–1123.

Itti, L. and Baldi, P. 2009. Bayesian surprise attracts human attention. *Vision Res. 49*, 10, 1295–1306.

Itti, L., Dhavale, N., and Pighin, F. 2003. Realistic avatar eye and head animation using a neurobiological model of visual attention. In *Proceedings of the SPIE 48th Annual International Symposium on Optical Science and Technology*. IEEE, Los Alamitos, CA.

Itti, L. and Koch, C. 2001a. Computational modeling of visual attention. *Nat. Rev. Neurosci. 2*, 3, 194–203.

Itti, L. and Koch, C. 2001b. Feature combination strategies for saliency-based visual attention systems. *J. Electr. Imaging 10*, 1, 161–169.

Itti, L., Koch, C., and Niebur, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell. 20*, 11, 1254–1259.

Johansson, R., Westling, G., Backstrom, A., and Flanagan, J. 2001. Eye-hand coordination in object manipulation. *J. Neurosci. 21*, 17, 6917–6932.

Johnson, A. and Proctor, R. 2003. *Attention: Theory and Practice*. Sage Publications, Newbury Park, CA.

Jonides, J. 1981. Voluntary versus automatic control over the mind's eye movements. In *Attention and Performance IX*, A. D. Long, Ed. Lawrence Erlbaum Associates, NJ, 187–203.

Kadir, T. and Brady, M. 2001. Saliency, scale and image description. *Int. J. Comput. Vision 45*, 2, 83–105.

KAHNEMAN, D. AND TREISMAN, A. 1992. The reviewing of object files: Object-specific integration of information. *Cognitive Psych. 24*, 175–219.

KANDEL, E. R., SCHWARTZ, J. H., AND JESSELL, T. M. 1996. *Essentials of Neural Science and Behavior*. McGraw-Hill/Appleton & Lange, New York.

KASTNER, S. AND UNGERLEIDER, L. G. 2001. The neural basis of biased competition in human visual cortex. *Neuropsychologia 39*, 1263–1276.

KOCH, C. AND ULLMAN, S. 1985. Shifts in selective visual attention: Towards the underlying neural circuitry. *Hum. Neurobiol. 4*, 4, 219–227.

KOOTSTRA, G., NEDERVEEN, A., AND DE BOER, B. 2008. Paying attention to symmetry. In *Proceedings of the British Machine Vision Conference*. Springer, Berlin.

KUNAR, M., FLUSBERG, S., AND WOLFE, J. 2008. The role of memory and restricted context in repeated visual search. *Percept. Psychophys. 70*, 314–328.

LAND, M. F. 2006. Eye-movements and the control of actions in everyday life. *Prog. Retinal Eye Res. 25*, 296–324.

LEE, E. A. 2008. Cyber physical systems: Design challenges. Tech. rep. UCB/EECS-2008-8, EECS Department, University of California, Berkeley.

LEE, K., BUXTON, H., AND FENG, J. 2003. Selective attention for cue-guided search using a spiking neural network. In *Proceedings of the International Workshop on Attention and Performance in Computer Vision*. IEEE, Los Alamitos, CA, 55–62.

LEVIN, D. 1996. Classifying faces by race: the structure of face categories. *J. Exp. Psych. 22*, 1364–1382.

LI, Z. 2005. The primary visual cortex creates a bottom-up saliency map. In *Neurobiology of Attention*, L. Itti, G. Rees, and J. Tsotsos, Eds. Elsevier Academic Press.

LIU, T., SLOTNICK, S. D., SERENCES, J. T., AND YANTIS, S. 2003. Cortical mechanisms of feature-based intentional control. *Cerebral Cortex 13*, 12.

LIVINGSTONE, M. S. AND HUBEL, D. H. 1987. Psychophysical evidence for separate channels for the perception of form, color, movement, and depth. *J. Neurosci. 7*, 11, 3416–3468.

LOGAN, G. D. 1996. The CODE theory of visual attention: an integration of space-based and object-based attention. *Psych. Rev. 103*, 603–649.

LOWE, D. G. 2004. Distinctive image features from scale-invariant key points. *Int. J. Comput. Vision 60*, 2, 91–110.

MAKI, A., NORDLUND, P., AND EKLUNDH, J.-O. 2000. Attentional scene segmentation: Integrating depth and motion. *Comput. Vision Image Understanding 78*, 3, 351–373.

MARR, D. 1982. *VISION – A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman and Company, New York.

MAUNSELL, J. H. R. 1995. The brain's visual world: Representation of visual targets in cerebral cortex. *Science 270*, 764–769.

MAY, S., KLODT, M., AND ROME, E. 2007. GPU-accelerated Affordance Cueing based on Visual Attention. In *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Los Alamitos, CA, 3385–3390.

MAZER, J. A. AND GALLANT, J. L. 2003. Goal-related activity in V4 during free viewing visual search. Evidence for a ventral stream visual salience map. *Neuron 40*, 6, 1241–50.

MCMAINS, S. A. AND SOMERS, D. C. 2004. Multiple spotlights of attentional selection in human visual cortex. *Neuron 42*, 677–686.

MERTSCHING, B., BOLLMANN, M., HOISCHEN, R., AND SCHMALZ, S. 1999. The neural active vision system. In *Handbook of Computer Vision and Applications*, B. Jähne, H. Haussecke, and P. Geissler, Eds. vol. 3. Academic Press, 543–568.

MIAU, F., PAPAGEORGIOU, C., AND ITTI, L. 2001. Neuromorphic algorithms for computer vision and attention. In *Proceedings of the 46th Annual SPIE International Symposium on Optical Science and Technology*. IEEE, Los Alamitos, CA, 12–23.

MILANESE, R. 1993. Detecting salient regions in an image: From biological evidence to computer implementation. Ph.D. thesis, University of Geneva, Switzerland.

MILANESE, R., WECHSLER, H., GIL, S., BOST, J., AND PUN, T. 1994. Integration of bottom-up and top-down cues for visual attention using non-linear relaxation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Los Alamitos, CA, 781–785.

MITRI, S., FRINTROP, S., PERVÖLZ, K., SURMANN, H., AND NÜCHTER, A. 2005. Robust object detection at regions of interest with an application in ball recognition. In *Proceedings of the IEEE International Conference on Robotics and Automation*. IEEE, Los Alamitos, CA, 126–131.

MOZER, M. C. 1987. Early parallel processing in reading: A connectionist approach. In *Attention and Performance XII: The Psychology of Reading*, M. Coltheart, Ed. Lawrence Erlbaum Associated Ltd., Philadelphia, 83–104.

MUHL, C., NAGAI, Y., AND SAGERER, G. 2007. On constructing a communicative space in HRI. In *Proceedings of the 30th German Conference on Artificial Intelligence*, J. Hertzberg, M. Beetz, and R. Englert, Eds. Springer, Berlin.

NAGAI, Y. 2009. From bottom-up visual attention to robot action learning. In *Proceedings of the 8th International IEEE Conference on Development and Learning*. IEEE, Los Alamitos, CA.

NAKAYAMA, K. AND MACKEBEN, M. 1989. Sustained and transient components of focal visual attention. *Vision Res. 29*, 1631–1647.

NAKAYAMA, K. AND SILVERMAN, G. H. 1986. Serial and parallel processing of visual feature conjunctions. *Nature 320*, 264–265.

NAVALPAKKAM, V. AND ITTI, L. 2006a. An integrated model of top-down and bottom-up attention for optimizing detection speed. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, Los Alamitos, CA.

NAVALPAKKAM, V. AND ITTI, L. 2006b. Top-down attention selection is fine-grained. *J. Vision 6*, 11, 1180–1193.

NAVALPAKKAM, V., REBESCO, J., AND ITTI, L. 2004. Modeling the influence of knowledge of the target and distractors on visual search. *J. Vision 4*, 8, 690.

NAVALPAKKAM, V., REBESCO, J., AND ITTI, L. 2005. Modeling the influence of task on attention. *Vision Res. 45*, 2, 205–231.

NEISSER, U. 1967. *Cognitive Psychology*. Appleton-Century-Crofts, New York.

NICKERSON, S. B., JASIOBEDZKI, P., WILKES, D., JENKIN, M., MILIOS, E., TSOTSOS, J. K., JEPSON, A., AND BAINS, O. N. 1998. The ARK project: Autonomous mobile robots for known industrial environments. *Rob. Auton. Syst. 25*, 1–2, 83–104.

NOTHDURFT, H.-C. 2005. Salience of feature contrast. In *Neurobiology of Attention*, L. Itti, G. Rees, and J. K. Tsotsos, Eds. Elsevier, Burlington, MA, 233–239.

OGAWA, T. AND KOMATSU, H. 2004. Target selection in area V4 during a multidimensional visual search task. *J. Neurosci. 24*, 28, 6371–6382.

OLIVA, A. 2005. Gist of the scene. In *Neurobiology of Attention*, L. Itti, G. Rees, and J. Tsotsos, Eds. Elsevier Academic Press, 251–257.

OLIVA, A., TORRALBA, A., CASTELHANO, M. S., AND HENDERSON, J. M. 2003. Top-down control of visual attention in object detection. In *Proceedings of the International Conference on Image Processing*. IEEE, Los Alamitos, CA, 253–256.

OLSHAUSEN, B., ANDERSON, C., AND VAN ESSEN, D. 1993. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J. Neurosci. 13*, 11, 4700–4719.

OLSHAUSEN, B. A. AND FIELD, D. J. 2005. How close are we to understanding V1? *Neural Comput. 17*, 8, 1665–1699.

OLSHAUSEN, B. A. AND FIELD, D. J. 2006. What is the other 85% of V1 doing? In *23 Problems in Systems Neuroscience*, L. V. Hemmen and T. Sejnowslti, Eds. Oxford University Press, Oxford, UK.

OUERHANI, N. 2003. Visual attention: From bio-inspired modeling to real-time implementation. Ph.D. thesis, Institut de Microtechnique Université de Neuchâtel, Switzerland.

OUERHANI, N., BUR, A., AND HÜGLI, H. 2005. Visual attention-based robot self-localization. In *Proceedings of the European Conference on Mobile Robotics (ECMR 2005)*. IEEE, Los Alamitos, CA, 8–13.

OUERHANI, N. AND HÜGLI, H. 2000. Computing visual attention from scene depth. In *Proceedings of the International Conference on Pattern Recognition (ICPR 2000)*. Vol. 1. IEEE, Los Alamitos, CA, 375–378.

OUERHANI, N., JOST, T., BUR, A., AND HÜGLI, H. 2006. Cue normalization schemes in saliency-based visual attention models. In *Proceedings International Cognitive Vision Workshop*. Springer, Berlin.

OUERHANI, N., VON WARTBURG, R., HÜGLI, H., AND MÜRI, R. 2004. Empirical validation of the saliency-based model of visual attention. *Electr. Lett. Comput. Vision Image Anal. 3*, 1, 13–24.

PALMER, J., AMES, C., AND LINDSEY, D. 1993. Measuring the effect of attention on simple visual search. *J. of Experimental Psychology. Hum. Percept. Perform. 19*, 1, 108–130.

PALMER, S. E. 1999. *Vision Science, Photons to Phenomenology*. MIT Press, Cambridge, MA.

PARKHURST, D., LAW, K., AND NIEBUR, E. 2002. Modeling the role of salience in the allocation of overt visual attention. *Vision Res. 42*, 1, 107–123.

PASHLER, H. 1997. *The Psychology of Attention*. MIT Press, Cambridge, MA.

PESSOA, L. AND EXEL, S. 1999. Attentional strategies for object recognition. In *Proceedings of the International Work Conference on Artificial and Natural Neural Networks*. Springer, Berlin, 850–859.

PETERS, R., IYER, A., ITTI, L., AND KOCH, C. 2005. Components of bottom-up gaze allocation in natural images. *Vision Res. 45*, 2397–2416.

PETERS, R. J. AND ITTI, L. 2008. Applying computational tools to predict gaze direction in interactive visual environments. *ACM Trans. Appl. Percept. 5*, 2.

PHAF, R. H., VAN DER HEIJDEN, A. H. C., AND HUDSON, P. T. W. 1990. SLAM: A connectionist model for attention in visual selection tasks. *Cognitive Psych. 22*, 273–341.

POSNER, M. AND COHEN, Y. 1984. Components of visual orienting. In *Attention and Performance X*, H. Bouma and D. Bouwhuis, Eds. Erlbaum, London, 531–556.

POSNER, M. I. 1980. Orienting of attention. *Q. J. Exp. Psych. 32*, 3–25.

Posner, M. I. and Petersen, S. E. 1990. The attentional system of the human brain. *Ann. Rev. Neurosci. 13*, 25–42.

Postma, E. 1994. Scan: A neural model of covert attention. Ph.D. thesis, Rijksuniversiteit Limburg, Wageningen.

Pylyshyn, Z. and Storm, R. 1988. Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision 3*, 179–197.

Pylyshyn, Z. W. 2003. *Seeing and Visualizing: It's Not What You Think*. MIT Press, Cambridge, MA.

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

Ramström, O. and Christensen, H. I. 2002. Visual attention using game theory. In *Proceedings of the Workshop on Biologically Motivated Computer Vision*. Springer, Berlin.

Ramström, O. and Christensen, H. I. 2004. Object-based visual attention: Searching for objects defined by size. In *Proceedings of International Workshop on Attention and Performance in Computational Vision*. Springer, Berlin, 9–16.

Rao, R., Zelinsky, G., Hayhoe, M., and Ballard, D. 2002. Eye-movements in iconic visual search. *Vision Res. 42*, 1447–1463.

Rasolzadeh, B., Björkman, M., Huebner, K., and Kragic, D. 2009. An active vision system for detecting, fixating and manipulating objects in real world. *Int. J. Rob. Res*.

Rauschenberger, R. 2003. Attentional capture by auto- and allo-cues. *Psychonomic Bull. Rev. 10*, 4, 814–842.

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

Rensink, R. A., O'Regan, J. K., and Clark, J. J. 1997. To see or not to see: The need for attention to perceive changes in scenes. *Psych. Sci. 8*, 368–373.

Riesenhuber, M. and Poggio, T. 1999. Hierarchical models of object recognition in cortex. *Nature Neurosci. 2*, 11, 1019–1025.

Rosenholtz, R. 2001. Search asymmetries? What search asymmetries? *Percept. Psychophys*. 63,3, 476–489.

Rotenstein, A., Andreopoulos, A., Fazl, E., Jacob, D., Robinson, M., Shubina, K., Zhu, Y., and Tsotsos, J. 2007. Towards the dream of intelligent, visually-guided wheelchairs. In *Proceedings of the 2nd International Conference on Technology and Aging*.

Rothenstein, A. and Tsotsos, J. 2006a. Attention links sensing to recognition. *Image Vision Comput. J. 26*, 1, 114–126.

Rothenstein, A. and Tsotsos, J. 2006b. Selective tuning: Feature binding through selective attention. In *Proceedings of International Conference on Artificial Neural Networks*. IEEE, Los Alamitos, CA.

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

Sabra, A. I. 1989. *The Optics of Ibn Al-Haytham*. The Warburg Institute, University of London.

Salah, A., Alpaydin, E., and Akrun, L. 2002. A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition. *IEEE Trans. Pattern Anal. Mach. Intell. 24*, 3, 420–425.

Sandini, G. and Metta, G. 2002. Retina-like sensors: Motivations, technology and applications. In *Sensors and Sensing in Biology and Engineering*. Springer Verlag, Berlin.

Schauerte, B., Richarz, J., Plötz, T., Thurau, C., and Fink, G. A. 2009. Multi-modal and multi-camera attention in smart environments. In *Proceedings of Multi-Modal Interfaces and Workshop on Machine Learning for Multi-Modal Interaction*. ACM, New York.

Scheier, C. and Egner, S. 1997. Visual attention in a mobile robot. In *Proceedings of the IEEE International Symposium on Industrial Electronics*. IEEE, Los Alamitos, CA, 48–53.

Scholl, B. J. 2001. Objects and attention: The state of the art. *Cognition 80*, 1–46.

Shulman, G., Remington, R., and McLean, J. 1979. Moving attention through visual space. *J. Exp. Psych. 5*, 3, 522–526.

Siagian, C. and Itti, L. 2009. Biologically inspired mobile robot vision localization. *IEEE Trans. Rob. 25*, 4, 861–873.

Simons, D. J. and Levin, D. T. 1997. Change blindness. *Trends Cognitive Sci. 1*, 261–267.

Styles, E. A. 1997. *The Psychology of Attention*. Psychology Press Ltd, Florence, KY.

Sumner, P. and Mollon, J. 2000. Catarrhine photopigments are optimized for detecting targets against a foliage background. *J. Exp. Biol. 203*, 1963–1986.

Sun, Y. and Fisher, R. 2003. Object-based visual attention for computer vision. *Artif. Intell. 146*, 1, 77–123.

Tatler, B. W. 2007. The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. *J. Vision 14*, 7, 1–17.

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

Tatler, B. W., Baddeley, R. J., and Vincent, B. T. 2006. The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. *Vision Res. 46*, 1857–1862.

Theeuwes, J. 2004. Top-down search strategies cannot override attentional capture. *Psychonomic Bull. Rev. 11*, 65–70.

Torralba, A. 2003a. Contextual priming for object detection. *Int. J. Comput.Vision 53*, 2, 169–191.

TORRALBA, A. 2003b. Modeling global scene factors in attention. *J. Opt. Soc. Am. 20*, 7, 1407–1418.

TREISMAN, A. M. 1993. The perception of features and objects. In *Attention: Selection, Awareness, and Control*, A. Baddeley and L. Weiskrantz, Eds. Clarendon Press, Oxford, 5–35.

TREISMAN, A. M. AND GELADE, G. 1980. A feature integration theory of attention. *Cognitive Psych. 12*, 97–136.

TREISMAN, A. M. AND GORMICAN, S. 1988. Feature analysis in early vision: Evidence from search asymmetries. *Psych. Rev. 95*, 1, 15–48.

TSOTSOS, J., RODRIGUEZ-SANCHEZ, A., ROTHENSTEIN, A., AND SIMINE, E. 2008. Different binding strategies for the different stages of visual recognition. *Brain Res. 1225*, 119–132.

TSOTSOS, J. K. 1987. A "complexity level" analysis of vision. In *Proceedings of the International Conference on Computer Vision: Human and Machine Vision Workshop*. IEEE, Los Alamitos, CA.

TSOTSOS, J. K. 1990. Analyzing vision at the complexity level. *Behav. Brain Sci. 13*, 3, 423–445.

TSOTSOS, J. K. 1993. An inhibitory beam for attentional selection. In *Spatial Vision in Humans and Robots*, L. R. Harris and M. Jenkin, Eds. Cambridge University Press, Cambridge, UK, 313–331.

TSOTSOS, J. K., CULHANE, S. M., WAI, W. Y. K., LAI, Y., DAVIS, N., AND NUFLO, F. 1995. Modeling visual attention via selective tuning. *Artif. Intell. 78*, 1-2, 507–545.

TSOTSOS, J. K., LIU, Y., MARTINEZ-TRUJILLO, J. C., POMPLUN, M., SIMINE, E., AND ZHOU, K. 2005. Attenting to visual motion. *J. Comput. Vision Image Understanding 100*, 1-2, 3–40.

TSOTSOS, J. K., VERGHESE, G., STEVENSON, S., BLACK, M., METAXAS, D., CULHANE, S., DICKINSON, S., JENKIN, M., JEPSON, A., ET AL. 1998. PLAYBOT: A visually-guided robot to assist physically disabled children in play. *Image Vision Comput. 16*, 275–292.

TUYTELAARS, T. AND MIKOLAJCZYK, K. 2007. Local invariant feature detectors: A survey. *Found. Trends Comput. Graphics Vision 3*, 3, 177–280.

VAN OEFFELEN, M. P. AND VOS, P. G. 1982. Configurational effects on the enumeration of dots: Counting by groups. *Memory Cognition 10*, 396–404.

VECERA, S. AND FARAH, M. 1994. Does visual attention select objects or locations? *J. Exp. Psych. 123*, 2, 146–160.

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

VICKERY, T. J., KING, L.-W., AND JIANG, Y. 2005. Setting up the target template in visual search. *J. Vision 5*, 1, 81–92.

VIJAYAKUMAR, S., CONRADT, J., SHIBATA, T., AND SCHAAL, S. 2001. Overt visual attention for a humanoid robot. In *Proceedings of the International Conference on Intelligence in Robotics and Autonomous Systems (IROS 2001)*. ACM, New York, 2332–2337.

VINCENT, B. T., TROSCIANKO, T., AND GILCHRIST, I. D. 2007. Investigating a space-variant weighted salience account of visual selection. *Vision Res. 47*, 1809–1820.

VIOLA, P. AND JONES, M. J. 2004. Robust real-time face detection. *Int. J. Comput. Vision 57*, 2, 137–154.

VON HELMHOLTZ, H. 1896. *Handbuch der physiologischen Optik*. Von Leopold Voss Verlag, Hamburg, Germany. (an English Quote is included in Nakayama & Mackeben, 1989).

WALTHER, D. 2006. Interactions of visual attention and object recognition: computational modeling, algorithms, and psychophysics. Ph.D. thesis, California Institute of Technology, Pasadena, CA.

WALTHER, D., EDGINGTON, D. R., AND KOCH, C. 2004. Detection and tracking of objects in underwater video. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*. IEEE, Los Alamitos, CA.

WALTHER, D. AND KOCH, C. 2007. Attention in hierarchical models of object recognition. *Comput. Neurosci. 165*, 57–78.

WELLS, A. AND MATTHEWS, G. 1994. *Attention and Emotion: A Clinical Perspective*. Psychology Press, Florence, KY.

WOLFE, J. M. 1994. Guided search 2.0: A revised model of visual search. *Psychonomic Bull. Rev. 1*, 2, 202–238.

WOLFE, J. M. 1998a. Visual search. In *Attention*, H. Pashler, Ed. Psychology Press, Florence, KY, 13–74.

WOLFE, J. M. 1998b. What can 1,000,000 trials tell us about visual search? *Psych. Sci. 9*, 1, 33–39.

WOLFE, J. M. 2001a. Asymmetries in visual search: An introduction. *Percept. Psychophys. 63*, 3, 381–389.

WOLFE, J. M. 2001b. Guided search 4.0: A guided search model that does not require memory for rejected distractors. *J. Vision 1*, 3, 349a.

WOLFE, J. M. 2007. 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.

WOLFE, J. M., CAVE, K., AND FRANZEL, S. 1989. Guided search: An alternative to the feature integration model for visual search. *J. Exp. Psych. 15*, 419–433.

WOLFE, J. M. AND GANCARZ, G. 1996. *Guided Search 3.0: Basic and Clinical Applications of Vision Science*. Kluwer Academic, The Netherlands, 189–192.

WOLFE, J. M., HOROWITZ, T., KENNER, N., HYLE, M., AND VASAN, N. 2004. How fast can you change your mind? The speed of top-down guidance in visual search. *Vision Res. 44*, 1411–1426.

Wolfe, J. M. and Horowitz, T. S. 2004. What attributes guide the deployment of visual attention and how do they do it? *Nat. Rev. Neurosci. 5*, 1–7.

Xu, T., Chenkov, N., Kühnlenz, K., and Buss, M. 2009. Autonomous switching of top-down and bottom-up attention selection for vision guided mobile robots. In *Proceedings of the International Conference on Intelligent Robots and Systems*. ACM, New York.

Xu, T., Pototschnig, T., Kühnlenz, K., and Buss, M. 2009. A high-speed multi-GPU implementation of bottom-up attention using CUDA. In *Proceedings of the International Conference on Robotics and Automation*. IEEE, Los Alamitos, CA.

Yantis, S. 2000. Goal-directed and stimulus-driven determinants of attentional control. In *Attention and Performance*, S. Monsell and J. Driver, Eds. Vol. 18. MIT Press, Cambridge, MA.

Yantis, S., Ach, J. S., Serences, J., Carlson, R., Steinmetz, M., Pekar, J., and Courtney, S. 2002. Transient neural activity in human parietal cortex during spatial attention shifts. *Nat. Neurosci. 5*, 995–1002.

Yantis, S. and Serences, J. T. 2003. Cortical mechanisms of space-based and object-based attentional control. *Curr. Opin. Neurobiol. 13*, 187–193.

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

Zeki, S. 1993. *A Vision of the Brain*. Blackwell Scientific, Cambridge, MA.

Zelinsky, G. J. and Sheinberg, D. L. 1997. Eye-movements during parallel-serial visual search. *J. Exp. Psych. Hum. Percept. Perform. 23*, 1, 244–262.

Received February 2007; revised July 2008; accepted November 2008

COMPUTATIONAL MODELLING OF VISUAL ATTENTION

Laurent Itti[*] *and Christof Koch*[‡]

We review recent work on computational models of focal visual attention, with emphasis on the bottom-up, image-based control of attentional deployment. We highlight five important trends that have emerged from the computational literature. First, the perceptual saliency of stimuli critically depends on the surrounding context. Second, a unique 'saliency map' that topographically encodes for stimulus conspicuity over the visual scene has proved to be an efficient and plausible bottom-up control strategy. Third, inhibition-of-return, the process by which the currently attended location is prevented from being attended again, is a crucial element of attentional deployment. Fourth, attention and eye movements tightly interplay, posing computational challenges with respect to the coordinate system used to control attention. And last, scene understanding and object recognition strongly constrain the selection of attended locations. Insights from these five key areas provide a framework for a computational and neurobiological understanding of visual attention. **[Author: OK?]**

CENTRE-SURROUND MECHANISMS
These involve neurons that respond to image differences between a small central region and a broader concentric antagonistic surround region [Author: OK?].

[*]*Hedco Neuroscience Building, University of Southern California, 3641 Watt Way, Los Angeles, California 90089-2520, USA.* [‡]*Division of Biology, Caltech, Pasadena, California 91125, USA. Correspondence to C.K. e-mail: koch@klab.caltech.edu [Author: Addresses OK?]*

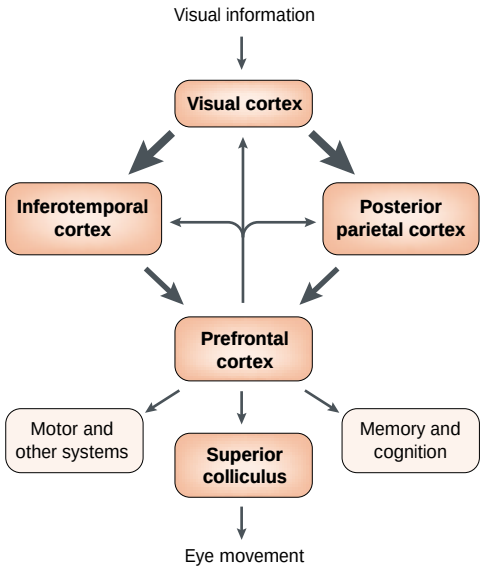
The most important function of selective visual attention is to direct our gaze rapidly towards objects of interest in our visual environment[1–9]. This ability to orientate rapidly towards salient objects in a cluttered visual scene has evolutionary significance because it allows the organism to detect quickly possible prey, mates or predators in the visual world. A two-component framework for attentional deployment has recently emerged, although the idea dates back to William James[1], the father of American psychology. This framework suggests that subjects selectively direct attention to objects in a scene using both bottom-up, image-based saliency cues and top-down, task-dependent cues.

Some stimuli are intrinsically conspicuous or salient in a given context. For example, a red dinner jacket among black tuxedos at a sombre state affair, or a flickering light in an otherwise static scene, automatically and involuntarily attracts attention. Saliency is independent of the nature of the particular task, operates very rapidly and is primarily driven in a bottom-up manner, although it can be influenced by contextual, figure-ground effects. If a stimulus is sufficiently salient, it will pop out of a visual scene. This suggests that saliency is computed in a pre-attentive manner across the entire visual field, most probably in terms of hierarchical CENTRE-SURROUND MECHANISMS. The speed of this saliency-based form of attention is on the order of 25 to 50 ms per item.

The second form of attention is a more deliberate and powerful one that has variable selection criteria, depending on the task at hand (for example, 'look for the red, horizontal target'). The expression of this top-down attention is most probably controlled from higher areas, including the frontal lobes, which connect back into visual cortex and early visual areas. Such volitional deployment of attention has a price, because it takes time — 200 ms or more, which rivals the time needed to move the eyes. So, whereas certain features in the visual world automatically attract attention and are experienced as 'visually salient', directing attention to other locations or objects requires voluntary 'effort'. Both mechanisms can operate in parallel.

Attention implements an information-processing bottleneck that allows only a small part of the incoming sensory information to reach short-term memory and

Box 1 | Neuronal mechanisms for the control of attention



The brain regions that participate in the deployment of visual attention include most of the early visual processing area. A simplified overview of the main brain areas involved is shown in the figure. Visual information enters the primary visual cortex via the lateral geniculate nucleus (not shown), although weaker pathways, for example, to the superior colliculus (SC), also exist. From there, visual information progresses along two parallel hierarchical streams. Cortical areas along the 'dorsal stream' (including the posterior parietal cortex; PPC) are primarily concerned with spatial localization and directing attention and gaze towards objects of interest in the scene. The control of attentional deployment is consequently believed to mostly take place in the dorsal stream. Cortical areas along the 'ventral stream' (including the inferotemporal cortex; IT) are mainly concerned with the recognition and identification of visual stimuli. Although probably not directly concerned with the control of attention, these ventral-stream areas have indeed been shown to receive attentional feedback modulation, and are involved in the representation of attended locations and objects (that is, in what passes through the attentional bottleneck). In addition, several higher-function areas are thought to contribute to attentional guidance, in that lesions in those areas causes a condition of 'neglect' in which patients seem unaware of parts of their visual environment (see REF. 111 for an overview of the regions involved).

From a computational viewpoint, the dorsal and ventral streams must interact, as scene understanding involves both recognition and spatial deployment of attention. One region where such interaction has been extensively studied is the prefrontal cortex (PFC), which is bidirectionally connected to both the PPC and the IT (see REF. 15). So, in addition to being responsible for planning of action (such as the execution of eye movements through the SC), the PFC has an important role in modulating, in a feedback manner, the dorsal and ventral processing streams.

visual awareness[10,11]. So, instead of attempting to fully process the massive sensory input (estimated to be on the order of 10^7–10^8 bits per second at the optic nerve) in parallel, a serial strategy has evolved that achieves near real-time performance despite limited computational capacity. Attention allows us to break down the problem of understanding a visual scene into rapid series of computationally less demanding, localized visual analysis problems. In addition to these orientating and scene analysis functions, attention is also characterized by a feedback modulation of neural activity for the visual attributes and at the location of desired or selected targets. This feedback is believed to be essential for binding the different visual attributes of an object, such as colour and form, into a unitary percept[2,12,13]. By this account, attention not only serves to select a location of interest but also enhances the cortical representation of objects at that location. As such, focal visual attention has been compared to a 'stagelight', successively illuminating different players as they take centre stage[14]. Finally, attention is involved in triggering behaviour, and is consequently intimately related to recognition, planning and motor control[15].

Developing computational models that describe how attention is deployed within a given visual scene has been an important challenge for computational neuroscience. The potential application of these architectures in artificial vision for tasks such as surveillance, automatic target detection, navigational aids and robotics control provides additional motivation. Here, we focus on biologically plausible computational modelling of a saliency-based form of focal bottom-up attention. Much less is known about the neural instantiation of the top-down, volitional component of attention[16,17]. As this aspect of attention has not been modelled in such detail, it is not our primary focus here.

The control of focal visual attention involves an intricate network of brain areas (BOX 1). In a first approximation, selecting where to attend next is primarily controlled by the DORSAL STREAM of visual processing[18], although object recognition in the VENTRAL STREAM can bias the next attentional shift through top-down control (see below). The basis of most computational models are the experimental results obtained using the visual search paradigm of Treisman and colleagues, in particular the distinction between pop-out and conjunctive searches developed in the early 1980s[2].

The first explicit, neurally plausible computational architecture for controlling visual attention was proposed by Koch and Ullman[19] in 1985 (FIG. 1) (for an earlier related model of vision and eye movements, see Didday and Arbib[20]). Koch and Ullman's model was centred around a 'saliency map', that is, an explicit two-dimensional topographical map that encodes stimulus conspicuity, or saliency, at every location in the visual scene. The saliency map receives inputs from early visual processing, and provides an efficient control strategy by which the focus of attention simply scans the saliency map in order of decreasing saliency. Following this basic architecture, we concentrate on five essential compo-

DORSAL STREAM
Brain areas involved in the localization of objects and mostly found in the posterior/superior part of the brain.

VENTRAL STREAM
Brain areas involved in the identification of objects and mostly found in the posterior/inferior part of the brain.

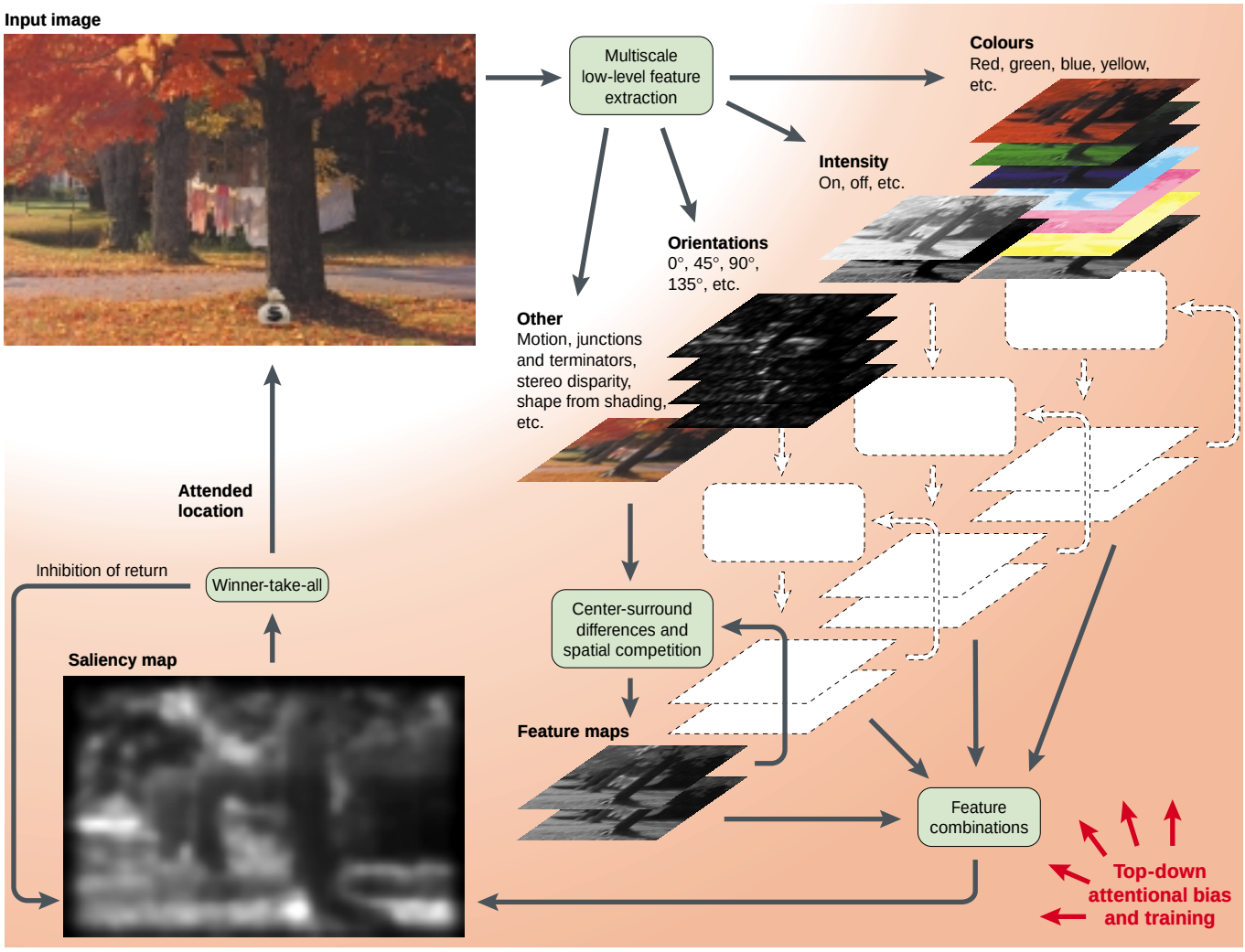


Figure 1 | **Flow diagram of a typical model of the control of bottom-up attention**. This diagram is based on Koch and Ullman's[19] hypothesis that a centralized two-dimensional saliency map can provide an efficient control strategy for the deployment of attention on the basis of bottom-up cues. The input image is decomposed through several pre-attentive feature detection mechanisms (sensitive to colour, intensity and so on), which operate in parallel over the entire visual scene. Neurons in the feature maps then encode for spatial contrast in each of those feature channels. In addition, neurons in each feature map spatially compete for salience, through long-range connections that extend far beyond the spatial range of the classical receptive field of each neuron. After competition, the feature maps are combined into a unique saliency map, which topographically encodes for saliency irrespective of the feature channel in which stimuli appeared salient. The saliency map is sequentially scanned by attention through the interplay between a winner-take-all network (which detects the point of highest saliency at any given time) and inhibition-of-return (which suppresses the last attended location from the saliency map, so that attention can focus onto the next most salient location). Top-down attentional bias and training can modulate, top-down, most stages of this bottom-up model.

nents of any model of bottom-up attention. These are the pre-attentive computation of early visual features across the entire visual scene, their integration to yield a single attentional control command, the generation of attentional scanpaths, the interaction between covert and overt attentional deployment (that is, eye movements) and the interplay between attention and scene understanding.

Pre-attentive computation of visual features

The first processing stage in any model of bottom-up attention is the computation of early visual features. In biological vision, visual features are computed in the retina, superior colliculus, lateral geniculate nucleus and early visual cortical areas[21]. Neurons at the earliest stages are tuned to simple visual attributes such as INTENSITY CONTRAST, COLOUR OPPONENCY, orientation, direction and velocity of motion, or stereo disparity at several spatial scales. NEURONAL TUNING becomes increasingly more specialized with the progression from low-level to high-level visual areas, such that higher-level visual areas include neurons that respond only to corners or junctions[22], shape-from-shading cues[23,24] or views of specific real-world objects[25–28].

Early visual features are computed pre-attentively in a massively parallel manner across the entire visual field (note, however, that we do not imply here that such computation is purely feedforward, as object recognition and attention can influence it[29]). Indeed, neurons fire vigorously in these early areas even if the

INTENSITY CONTRAST
Spatial difference (for example, detected by centre-surround mechanisms) in light intensity (luminance) in a visual scene.

COLOUR OPPONENCY
Spatial difference in colours, computed in the brain using red/green and blue/yellow centre-surround mechanisms.

NEURONAL TUNING
Property of visual neurons to only respond to certain classes of stimuli (for example, vertically orientated bars).

animal is attending away from the receptive field at the site of recording[30], or is anaesthetized[31]. In addition, several psychophysical studies, as well as introspection, indicate that we are not blind to the world outside the focus of attention. Thus we can make simple judgments on objects to which we are not attending[32], although those judgments are limited and less accurate than those made in the presence of attention[2,12,13,33–36]. So although attention does not seem to be mandatory for early vision, it has recently become clear that attention can vigorously modulate, in a top-down manner, early visual processing, both in a spatially-defined and in a non-spatial but feature-specific manner[37–39]. This modulatory effect of attention has been described as enhanced gain[30], biased[40,41] or intensified[33] competition, or enhanced spatial resolution[34], or as modulated background activity[42], effective stimulus strength[43] or noise[44]. That attention can modulate early visual processing in a manner equivalent to an increase of stimulus strength[43] is computationally an important finding, which directly supports the metaphor of attention as a stagelight. Of particular interest from a computational perspective is a recent study by Lee *et al.*[33] that measured PSYCHOPHYSICAL THRESHOLDS for three simple pattern-discrimination tasks (contrast, orientation and spatial-frequency discriminations) and two spatial-masking tasks (32 thresholds in total). A dual-task paradigm was used to measure thresholds either when attention was fully available to the task of interest, or when it was less available because it was engaged elsewhere by a concurrent attention-demanding task. The mixed pattern of attentional modulation observed in the thresholds (up to threefold improvement in orientation discrimination with attention, but only 20% improvement in contrast discrimination) can be quantitatively accounted for by a computational model. This model predicts that attention activates a winner-take-all competition among neurons tuned to different orientations and spatial frequencies within one cortical HYPERCOLUMN[33,45], a proposition that has recently received further experimental support[46]. Because feedback modulation influences the computation of bottom-up features, models of bottom-up attention need to take this into account. An example of a mixed bottom-up and top-down model in which attention enhances spatial resolution[47] is discussed later.

Computational models may or may not include explicit details about early visual feature extraction. Models that do not are restricted to images for which the responses of feature detectors can reasonably be guessed. Models that do have the widest applicability to any visual stimulus, including natural scenes. Computer implementations of early visual processes are often motivated by an imitation of biological properties. For example, the response of a neuron tuned to INTENSITY CENTRE-SURROUND CONTRAST can be computed by convolving the luminance channel of the input image by a DIFFERENCE-OF-GAUSSIANS (Mexican Hat) filter. Similarly, the responses of orientation-selective neurons are usually obtained through convolution by GABOUR WAVELETS, which resemble biological IMPULSE RESPONSE FUNCTIONS[48,49]. Another interesting approach consists of implementing detectors that respond best to those features that are present at the locations visited by observers while free-viewing images[50,51]. For instance, Zetzsche *et al.*[50,52] showed using an eye-tracking device how the eyes preferentially fixate regions with multiple superimposed orientations such as corners, and derived nonlinear operators that specifically detect those regions.

Irrespective of the method used for early feature detection, several fundamental computational principles have emerged from both experimental and modelling studies. First, different features contribute with different strengths to perceptual saliency[53], and this relative feature weighting can be influenced depending on the demands of the task through top-down modulation[38,54] and through training[55–58]. Second, at a given visual location, there is little evidence for strong interactions across different visual modalities, such as colour and orientation[53]. This is not too surprising from a computational viewpoint, as one would otherwise expect these interactions to also be subject to training and top-down modulation, and this would result in the ability to learn to detect conjunctive targets efficiently, which we lack[2,59]. Within a given broad feature dimension, however, strong local interactions between filters sensitive to different properties of that feature (for example, between different orientations within the broad orientation feature) have been precisely characterized, both in physiology[60] and psychophysics[45]. Less evidence exists for within-feature competition across different spatial scales[45].

Last and most importantly, what seems to matter in guiding bottom-up attention is feature contrast rather than local absolute feature strength[61]. Indeed, not only are most early visual neurons tuned to some type of local spatial contrast (such as centre-surround or orientated edges), but neuronal responses are also strongly modulated by context, in a manner that extends far beyond the range of the classical receptive field (cRF)[62]. In a first approximation, the computational consequences of non-classical surround modulation are twofold. First, a broad inhibitory effect is observed when a neuron is excited with its preferred stimulus but that stimulus extends beyond the neuron's cRF, compared with when the stimulus is restricted to the cRF and the surrounding visual space is either empty or contains non-preferred stimuli[63–65]. Second, long-range excitatory connections in V1 seem to enhance responses of orientation-selective neurons when stimuli extend to form a contour[66,67]. These interactions are thought to be crucial in perceptual grouping[68,69]. The net result is that activity in early cortical areas is surprisingly sparse when monkeys are free-viewing natural scenes[70], compared with the vigorous responses that can be elicited by small laboratory stimuli presented in isolation.

So, the computation of early visual features entails more than localized operations limited to the cRF of visual neurons, as local responses crucially depend on longer-range contextual influences. To explicitly demonstrate this idea with a computer model, Itti *et al.*[71] compared purely local spatial frequency 'richness'

PSYCHOPHYSICAL THRESHOLDS
Smallest difference between two visual stimuli (for example, vertical versus tilted bar) than can reliably (that is, with a given probability of error) be reported by an observer.

HYPERCOLUMN
A patch of cortex including neurons responding to all orientations and many spatial scales, all for a single location in the visual field.

INTENSITY CENTRE-SURROUND CONTRAST
Author: please define

DIFFERENCE-OF-GAUSSIAN
A filter obtained by taking the difference between a narrow Gaussian distribution (the excitatory centre) and a broader Gaussian distribution with same mean (the inhibitory surround).

GABOUR WAVELET
Product of a sinusoidal grating and a two-dimensional Gaussian function.

IMPULSE RESPONSE FUNCTION
The response of a filter to a single point (Dirac) stimulus.

(as measured by computing local Fourier components with a magnitude above a certain threshold) with a saliency measure that included broad non-classical surround inhibition. They designed images with uniformly rich spatial frequency content (using colour speckle noise), but containing a perceptually salient target. Although the target was undifferentiated from its surround in terms of spatial frequency content, it was correctly detected by the mechanism including contextual competition.

Pre-attentive mechanisms that extract early visual features across the entire visual scene should not be overlooked in future modelling efforts. Indeed, it has recently become clear that early vision is far from being a passive and highly prototyped image-processing front-end that can be accurately modelled by linear filtering operations. Perceptually, whether a given stimulus is salient or not cannot be decided without knowledge of the context in which the stimulus is presented. So computationally, one must also account for nonlinear interactions across distant spatial locations, which mediate contextual modulation of neuronal responses.

Saliency

We have seen how the early stages of visual processing decompose the incoming visual input through an ensemble of feature-selective filtering processes endowed with contextual modulatory effects. The question that arises next is how to control a single attentional focus based on multiple neuronal networks that encode the incoming sensory signals using multiple representations. To solve this problem, most models of bottom-up attention follow that of Koch and Ullman[19] and hypothesize that the various feature maps feed into a unique 'saliency' or 'master' map[2,19]. The saliency map is a scalar, two-dimensional map whose activity topographically represents visual saliency, irrespective of the feature dimension that makes the location salient. That is, an active location in the saliency map encodes the fact that this location is salient, no matter whether it corresponds to a red object in a field of green objects, or to a stimulus moving towards the right while others move towards the left. On the basis of this scalar topographical representation, biasing attention to focus onto the most salient location is reduced to drawing attention towards the locus of highest activity in the saliency map.

Computationally, an explicit representation of saliency in a dedicated map reinforces the idea that some amount of spatial selection should be performed during pre-attentive feature detection. Otherwise, the divergence from retinal input to many feature maps could not be followed by a convergence into a saliency map without ending up with a representation in the saliency map that is as complex, cluttered and difficult to interpret as the original image. On the basis of this divergence, selection and convergence process, a location is defined as salient if it wins the spatial competition in one or more feature dimensions at one or more spatial scales. The saliency map then encodes for an aggregate measure of saliency not tied to any particular feature dimension, providing an efficient control strategy for focusing attention to salient locations without consideration of the detailed feature responses that made those locations salient.

Not surprisingly, many successful models for the bottom-up control of attention are built around a saliency map. What differentiates the models, then, is the strategy used to prune the incoming sensory input and extract saliency. In an influential model mostly aimed at explaining visual search experiments, Wolfe[54] hypothesized that the selection of relevant features for a given search task could be performed top-down, through spatially-defined and feature-dependent weighting of the various feature maps. Saliency is then computed in this model as the likelihood that a target will be present at a given location, based both on bottom-up feature contrast and top-down feature weight. This view has recently received experimental support from the many studies of top-down attentional modulation mentioned earlier.

Tsotsos and colleagues[72] implemented attentional selection using a combination of a feedforward bottom-up feature extraction hierarchy and a feedback selective tuning of these feature extraction mechanisms. In this model, the target for attention focusing is selected at the top level of the processing hierarchy (the equivalent of a saliency map), on the basis of feedforward activation and possible additional top-down biasing for certain locations or features. That location is then propagated back through the feature extraction hierarchy, through the activation of a cascade of winner-take-all networks embedded within the bottom-up processing pyramid. Spatial competition for saliency is thus refined at each level of processing, as the feedforward paths not contributing to the winning location are pruned (resulting in the feedback propagation of an 'inhibitory beam' around the selected target).

Milanese and colleagues[73] used a relaxation process to minimize an energy measure consisting of four terms: first, minimizing inter-feature incoherence favours those regions that excite several feature maps; second, minimizing intra-feature incoherence favours grouping of initially spread activity into small numbers of clusters; third, minimizing total activity in each map enforces intra-map spatial competition for saliency; and last, maximizing the dynamic range of each map ensures that the process does not converge towards uniform maps at some average value. Although the biological plausibility of this process remains to be tested, it has yielded a rare example of a model that can be applied to natural colour images.

Itti *et al.*[71,74] consider a purely bottom-up model, in which spatial competition for saliency is directly modelled after non-classical surround modulation effects. The model uses an iterative spatial competition scheme with early termination. At each iteration, a feature map receives additional inputs from the convolution of itself by a large difference-of-Gaussians filter. The result is half-wave rectified, a nonlinear process that ensures that the locations losing the competition are entirely eliminated. The net effect of this competitive process is similar to a winner-take-all process with limited inhibitory

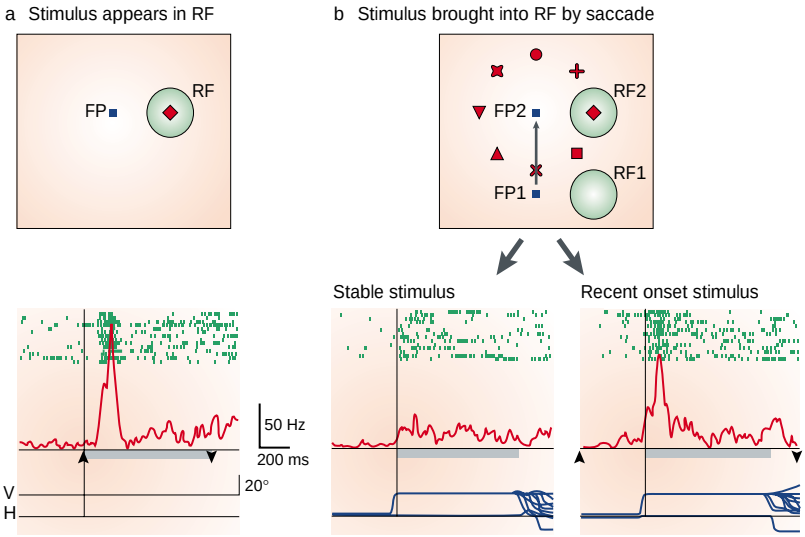


Figure 2 | **Recording saliency. [Author: OK?]** Once a purely computational hypothesis, the idea that saliency might be explicitly encoded by specific neurons in the cortex has recently received experimental support from many electrohpysiological studies[77–82]. How can one design an experiment that specifically tests whether a neuron responds to the saliency of a stimulus, rather than to the mere presence of that stimulus in the visual environment? In a particularly interesting experiment, Gottlieb and colleagues[80], recording from the lateral intraparietal sulcus of the awake monkey, found neurons that responded to visual stimuli only when those stimuli were made salient (by rapidly flashing them on a computer screen), but not otherwise. Their experiment cleverly used the retinotopic nature of the receptive fields of these neurons to bring a stimulus into their receptive field (RF) through a saccadic eye movement. **a** | In the control condition, a stimulus is presented in the RF of the neuron being recorded from, and elicits a response. That response could be simply visual, or indicating the saliency of this stimulus suddenly appearing in the visual field. **b** | To differentiate between these possibilities, two additional experiments were designed to be identical for the neuron of interest: a stimulus entered the RF through a saccade. However, a vigorous response was observed only when the stimulus had been made salient shortly before the beginning of the trial (by flashing it on and off while it still was outside the RF of the neuron; 'recent onset' condition). **[Author: please define FP]** (Adapted with permission from REF. 80 © (1998) Macmillan Magazines Ltd.)

spread, and allows only a sparse population of locations to remain active. After competition, all feature maps are simply summed to yield the scalar saliency map at the core of the model. Because it includes a complete front-end, this model has been widely applied to the analysis of natural colour scenes. Experimental results include the reproduction by the model of human behaviour in classical visual search tasks (popout versus conjunctive search, and search asymmetries[2,74]), a demonstration of very robust saliency computation with respect to image noise[71], the automatic detection of traffic signs and other salient objects in natural environments[58] and the detection of pedestrians in natural scenes (see below for preliminary results). Finally, the performance of the model at detecting military vehicles in the high-resolution Search2 NATO database of colour rural scenes[75] exceeded human performance in terms of the estimated number of locations that need to be visited by the attentional searchlight before the target is located[74].

In view of the numerous models based on a saliency map, it is important to note that postulating centralized control based on such a map is not the only computational alternative for the bottom-up guidance of attention. In particular, Desimone and Duncan[10] argued that saliency is not explicitly represented by specific neurons and by a saliency map, but instead is implicitly coded in a distributed modulatory manner across the various feature maps. Attentional selection is then performed on the basis of top-down enhancement of the feature maps relevant to a target of interest and extinction of those that are distracting, but without an explicit computation of salience. At least one model successfully applied this strategy to synthetic stimuli[76]; note, however, that such top-down biasing (also used in Wolfe's Guided Search model to select the weights of various feature contributions to the saliency map) requires that a specific search task be performed for the model to yield useful predictions.

Although originally a theoretical construct supported by sparse experimental evidence, the idea of a unique, centralized saliency map seems today to be challenged by the apparent existence of multiple areas that encode stimulus saliency in the visual system of the monkey. These regions include areas in the lateral intraparietal sulcus of the posterior parietal cortex (FIG. 2), the frontal eye fields, the inferior and lateral subdivisions of the pulvinar and the superior colliculus[77–82].

One possible explanation for this multiplicity could be that some of the neurons in these areas are indeed concerned with the explicit computation of saliency, but are located at different stages along the sensorimotor processing stream. For example, other functions have also been assigned to the posterior parietal cortex, such as that of mapping retinotopic to head-centred coordinate systems and of memorizing targets for eye or arm movements[83,84]. So more detailed experimental studies are needed to reveal subtle differences in the functions and representations found in these brain areas. Most probably, the main difference between these brain regions is the balance between their role in perception and action[15,82]. Meanwhile, it is worth noting that, in addition to the physiological findings just mentioned, recent psychophysical results also support the idea of an explicit encoding of saliency in the brain[85].

Attentional selection and inhibition-of-return

The saliency map guides where the attentional stagelight or spotlight[86] is to be deployed, that is, to the most salient location in the scene. One plausible neural architecture to detect the most salient location is that of a winner-take-all network, which implements a neurally distributed maximum detector[19,87]. Using this mechanism, however, raises another computational problem: how can we prevent attention from permanently focusing onto the most active (winner) location in the saliency map? One efficient computational strategy, which has received experimental support, consists of transiently inhibiting neurons in the saliency map at the currently attended location. After the currently attended location is thus suppressed, the winner-take-all network naturally converges towards the next most salient location, and repeating this process generates attentional scanpaths[19,71].

Such inhibitory tagging of recently attended locations has been widely observed in human psychophysics as a phenomenon called 'inhibition-of-return' (IOR)[88,89]. A typical psychophysical experiment

Box 2 | Attention and eye movements

Most of the models and experiments reviewed here are concerned with covert attention, that is, shifts of the focus of attention in the absence of eye movements. In normal situations, however, we move our eyes 3–5 times per second (that is, 150,000 to 250,000 times every day), to align locations of interest with our foveas. Overt and covert attention, however, are closely related, as revealed by psychophysical[112–115] physiological[79,81,83,116] and imaging[112,117] studies. The neuronal structures involved include the deeper parts of the superior colliculus; parts of the pulvinar; the frontal eye fields in the macaque and its homologue in humans, the precentral gyrus; and areas in the intraparietal sulcus in the macaque and around the intraparietal and postcentral sulci and adjacent gyri in humans. An example of overlapping functionality in humans is the study by Hoffman and Subramaniam[114]. They designed an experiment in which subjects performed a saccade just preceded by a target detection task; the greater accuracy found when the target to be detected appeared at the endpoint location of the saccade suggests that covert attention had been deployed to that endpoint in preparation to the saccade [Author: please clarify this sentence] .

For models, the addition of eye movements poses several additional computational challenges. Of particular interest is the need for compensatory mechanisms to shift the saliency map (typically in retinotopic coordinates) as eye movements occur. Dominey and Arbib[118] proposed a biologically plausible computational architecture that could perform such dynamic remapping in posterior parietal cortex (PPC). They noted that eye velocity signals have not been found in PPC; however, cells modulated by eye position have been reported[83]. They thus devised an iterative scheme to shift the contents of the saliency map according to the difference between current eye position and a temporally damped eye position signal. Their algorithm builds a convolution kernel from the difference between current and damped eye positions, which, when applied to the saliency map, translates it in the direction opposite to that difference. A related approach was proposed by Pouget and Sejnowski[84], in which the observed modulation of neuronal responses in PPC by retinal location and eye position ('gain field'[80]) is modelled by a set of basis functions, then used to transform from retinotopic to head-centred coordinates.

The interaction between overt and covert attention is particularly important for models concerned with visual search[119–121]. Further modelling of such interactions promises a better understanding of many mechanisms, including saccadic suppression, dynamic remapping of the saliency map and inhibition of return, covert pre-selection of targets for overt saccades, and the online understanding of complex visual scenes.

to evaluate IOR consists of performing speeded local pattern discriminations at various locations in the visual field; when a discrimination is performed at a location to which the observer has been previously cued, reaction times are slightly, but significantly, higher than at locations not previously visited[90]. These results indicate that visual processing at recently attended locations might be slower, possibly owing to some inhibitory tagging at attended locations. Several authors have specifically isolated an attentional component of IOR in addition to a motor (response delay) component[91–93].

Computationally, IOR implements a short-term memory of the previously visited locations and allows the attentional selection mechanism to focus instead on new locations. The simplest implementation of IOR consists of triggering transient inhibitory conductances in the saliency map at the currently attended location[74]. However, this only represents a coarse approximation of biological IOR, which has been shown to be object-bound, so that it should track and follow moving objects, and compensate for a moving observer as well[94–97]. The frame of reference in which IOR is expressed is an important issue because the eyes and the body can move (BOX 2). This frame-of-reference problem should be accounted for in computational models. Note that the idea of IOR is not necessarily contradicted by the recent findings of Horowitz and Wolfe[98] that visual search does not seem to involve memory: when elements of a search array were randomly reorganized at short time intervals while subjects were searching for a specific target, search efficiency was not degraded compared with when the search array remained stationary. Although these results preclude perfect memorization of all previously attended locations (otherwise, search on a stable array should be more efficient than on a constantly changing array), they do not preclude that the positions of the last few visited items were remembered, in accordance with the limited lifespan reported for IOR[90].

Although simple in principle, IOR is computationally a very important component of attention, in that it allows us — or a model — to rapidly shift the attentional focus over different locations with decreasing saliency, rather than being bound to attend only to the location of maximal saliency at any given time. The role of IOR in active vision and overt attention also poses challenges that will need to be addressed in more detail by future models (BOX 2).

Attention and recognition

So far, we have reviewed computational modelling and supporting experimental evidence for a basic architecture concerned with the bottom-up control of attention: early visual features are computed in a set of topographical feature maps; spatial competition for saliency prunes the feature responses to only preserve a handful of active locations; all feature maps are then combined into a unique scalar saliency map; and, finally, the saliency map is scanned by the focus of attention through the interplay between winner-take-all and IOR. Although such a simple computational architecture might accurately describe how attention is deployed within the first few hundreds of milliseconds after the presentation of a new scene, it is obvious that a more complete model of attentional control must include top-down, volitional biasing influences as well. The computational challenge, then, lies in the integration of bottom-up and top-down cues, such as to provide coherent control signals for the focus of attention, and in the interplay between attentional orientating and scene or object recognition.

One of the earliest models that combines object recognition and attention is MORSEL[99], in which attentional selection was shown to be necessary for object recognition. This model is applied to the recognition of words processed through a recognition hierarchy. Without attentional selection, the representations of several words in a scene would conflict and confuse that recognition hierarchy, yielding multiple superimposed representations at the top level. The addition of a top-down attentional selection process allowed the model to disambiguate recognition by focusing on one word at a time. Another early model that is worth mentioning here is described in REF. 100.

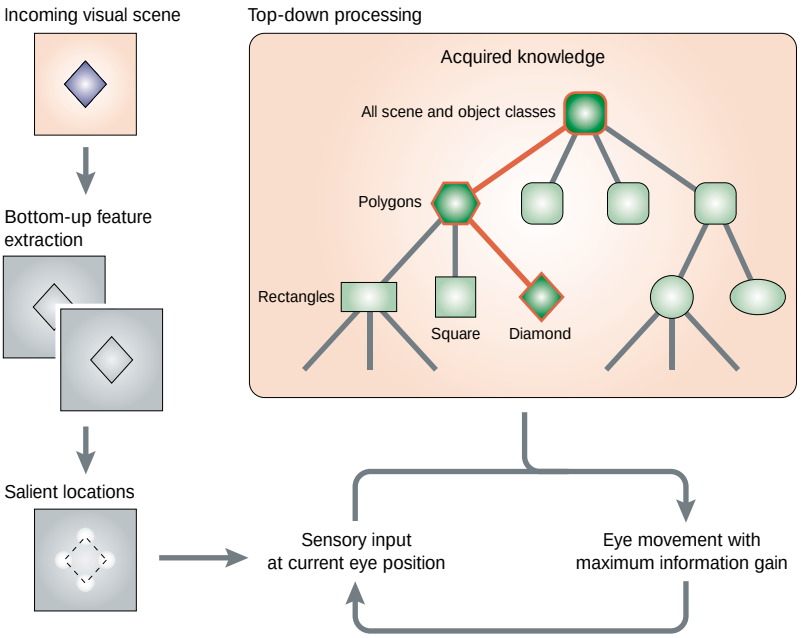


Figure 3 | **Combined model of attentional selection and object recognition.** Attention scans the scene such as to gather as much information as possible that can help discriminate between several recognition hypotheses. The model has two main components. First, a bottom-up feature extraction pathway extracts informative image regions from an incoming visual scene (for example, the corners of the diamond in the present illustration). Second, a trained knowledge base hierarchically represents object classes and encodes for both expected visual features at a set of critical points on these objects, and motor commands to move the eyes from one critical point to another. Recognition is then achieved by choosing for the next eye movement the one that maximizes information gain, that is, that best prunes the tree of known objects. In the hypothetical example shown to illustrate this idea, the first eye movement might thus go to the top corner of the object; finding sharp edges there would then suggest that this is a polygonal drawing. The knowledge base would then direct gaze to the most salient point directly below the currently fixated one, as that eye movement would best discriminate between the several known polygonal shapes; looking at the orientations of the features there, it becomes clear that the object is a diamond rather than a square or one of several possible rectangles. (Adapted with permission from REF. 101 **xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx.**)

A very interesting model that uses spatial shifts of attention during recognition was recently provided by Schill *et al.*[101]. Their model performs scene (or object) recognition, using attention (or eye movements) to focus on those parts of the scene that are most informative when disambiguating identity. To this end, a hierarchical knowledge tree is built through training. Here, leaves represent identified objects, intermediary nodes represent more general object classes and links between nodes contain sensorimotor information used for discrimination between possible objects (that is, bottom-up feature response to be expected for particular points in the object and eye movements targeted at those points). During the iterative recognition of an object, the system programs its next fixation towards the location that will maximize the gain of information about the object. This permits the model to discriminate between the current candidate object classes (FIG. 3).

Rybak *et al.*[102] proposed a related model, in which scanpaths (containing motor control directives stored in a 'where' memory and locally expected bottom-up features stored in a 'what' memory) are learned for each scene or object to be recognized. When presented with a new image, the model starts by selecting candidate scanpaths by [Author: OK?] matching bottom-up features in the image to those stored in the 'what' memory. For each candidate scanpath, the model deploys attention according to the directives in the 'where' memory and compares the local contents of the 'what' memory at each fixation with the local image features. This model can recognize complex greyscale scenes and faces in a translation-, rotation- and scale-independent manner.

Deco and Zihl have recently proposed another model that combines attentional selection and object recognition[47]. Their model starts by selecting candidate object locations bottom-up through a coarse-scale analysis of the image. An attentional mechanism scans the candidate locations in a serial fashion and performs object recognition at progressively finer scales until a sufficient recognition score is obtained for an object stored in memory. This model has been sucessfully applied to psychophysical experiments that show attentional enhancement of spatial resolution (see also REFS 34,103 for related experiments and modelling).

A more extreme view is expressed by the 'scanpath theory' of Stark[104], in which the control of eye movements is almost exclusively under top-down control. The theory proposes that what we see is only remotely related to the patterns of activation in our retinas. This is suggested by our permanent illusion of vivid perception over the entire field of view, although only the central two degrees of our foveal vision provide crisp sampling of the visual world. Rather, the scanpath theory argues that a cognitive model of what we expect to see is the basis for our percept; the sequence of eye movements that we make to analyse a scene, then, is mostly controlled by our cognitive model of that scene. This theory has had several successful applications to robotics control, in which an internal model of a robot's working environment was used to restrict the analysis of incoming video sequences to a small number of circumscribed regions important for a given task[105].

One important challenge for combined models of attention and recognition is finding suitable neuronal correlates for the various components. Despite the biological inspiration in these architectures, the models reviewed here do not relate in much detail to biological correlates of object recognition. Although several biologically plausible models have been proposed for object recognition in the ventral 'what' stream (in particular, REFS 106,107), their integration with neurobiological models concerned with attentional control in the dorsal 'where' stream remains an open issue. This integration will, in particular, have to account for the increasing experimental support for an object-based spatial focus of attention[108–110].

Summary

Here, we have discussed recent advances in the study of biologically plausible computational models of attention, with a particular emphasis on bottom-up control of attentional deployment. Throughout this review, we have stressed five important computational trends that

have emerged from this literature. First, saliency is derived from low-level visual features but, more than absolute feature strength or other detailed characteristics of the features, what seems to be important for the computation of saliency is feature contrast with respect to the contextual surround. Second, saliency increasingly seems to be a quantity that is coded explicitly in cortex separate from the visual features. This reinforces the once hypothetical concept of an explicit saliency map. Furthermore, several models have demonstrated the computational usefulness and plausibility of such an explicit map by successfully reproducing the behaviour of humans and monkeys in search tasks. Meanwhile, neural analogues of the saliency map are being found at multiple locations in the visual system of the macaque, hence posing a new challenge of integration of these many maps to yield unitary behaviour. Third, attention will not shift unless the currently attended (most salient) location is somehow disabled (otherwise, any model looking for saliency will keep coming back to the most salient location). IOR is consequently an essential computational component of attention and, indeed, it has been recently described as a complex, object-based and dynamically adaptive process that needs to be better modelled. Fourth, covert attention and eye movements are increasingly believed to share a common neuronal substrate. This poses serious computational problems with respect to the frame of reference in which saliency and IOR are computed. Recent evidence for world-centred and object-centred frames of reference need to be integrated into models. Last, the control of attentional deployment is intimately related to scene understanding and object recognition. Although several computer vision models with restricted biological plausibility have been proposed that integrate both attentional orientating and object identification, many exciting research challenges still lie in attempting to provide a more complete account of the dorsal and ventral processing streams in primate brains.

Controlling where attention should be deployed is not an autonomous feedforward process. Possible future directions for modelling work include modelling of interactions between task demands and top-down cues, bottom-up cues, mechanistic constraints (for example, when eye and body movements are executed) and neuroanatomical constraints such as feedback modulation.

Links

FURTHER INFORMATION Supplementary material for Figure 2

1. James, W. *The Principles of Psychology* (Harvard Univ. Press, Cambridge, Massachusetts, 1980/1981).
2. Treisman, A. M. & Gelade, G. A feature-integration theory of attention. *Cogn. Psychol.* **12**, 97–136 (1980).
An influencial theory of attention and visual search.
3. Bergen, J. R. & Julesz, B. Parallel versus serial processing in rapid pattern discrimination. *Nature* **303**, 696–698 (1983).
4. Treisman, A. Features and objects: The fourteenth Bartlett memorial lecture. *Q. J. Exp. Psychol. A* **40**, 201–237 (1988).
5. Nakayama, K. & Mackeben, M. Sustained and transient components of focal visual attention. *Vision Res.* **29**, 1631–1647 (1989).
6. Braun, J. & Sagi, D. Vision outside the focus of attention. *Percept. Psychophys.* **48**, 45–58 (1990).
7. Hikosaka, O., Miyauchi, S. & Shimojo, S. Orienting a spatial attention-its reflexive, compensatory, and voluntary mechanisms. *Brain Res. Cogn. Brain Res.* **5**, 1–9 (1996).
8. Braun, J. & Julesz, B. Withdrawing attention at little or no cost: detection and discrimination tasks. *Percept. Psychophys.* **60**, 1–23 (1998).
9. Braun, J., Itti, L., Lee, D. K., Zenger, B. & Koch, C. in *Visual Attention and Neural Circuits* (eds Braun, J., Koch,C. & Davis, J.) (MIT, Cambridge, Massachusetts, in the press).
10. Desimone, R. & Duncan, J. Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* **18**, 193–222 (1995).
11. Crick, F. & Koch, C. Constraints on cortical and thalamic projections: the no-strong-loops hypothesis. *Nature* **391**, 245–250 (1998).
12. Hummel, J. E. & Biederman, I. Dynamic binding in a neural network for shape recognition. *Psychol. Rev.* **99**, 480–517 (1992).
13. Reynolds, J. H. & Desimone, R. The role of neural mechanisms of attention in solving the binding problem. *Neuron* **24**, 19–29 (1999).
14. Weichselgartner, E. & Sperling, G. Dynamics of automatic and controlled visual attention. *Science* **238**, 778–780 (1987).
15. Miller, E. K. The prefrontal cortex and cognitive control. *Nature Rev. Neurosci.* **1**, 59–65 (2000).
16. Hopfinger, J. B., Buonocore, M. H. & Mangun, G. R. The neural mechanisms of top-down attentional control. *Nature Neurosci.* **3**, 284–291 (2000).
17. Corbetta, M., Kincade, J. M., Ollinger, J. M., McAvoy, M. P. & Shulman, G. L. Voluntary orienting is dissociated from target detection in human posterior parietal cortex. *Nature Neurosci.* **3**, 292–297 (2000); erratum **3**, 521 (2000).
18. Ungerlaider, L. G. & Mishkin, M. in *Analysis of Visual Behavior* (eds Ingle, D. J., Goodale, M. A. & Mansfield, R. J. W.) (MIT, **[Author: City?]**, 1982).
19. Koch, C. & Ullman, S. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum. Neurobiol.* **4**, 219–227 (1985).
One of the first explicit computational models of bottom-up attention, at the origin of the idea of a "saliency map".
20. Didday, R. L. & Arbib, M. A. Eye movements and visual perception: A "two visual system" model. *Int. J. Man–Machine Studies* **7**, 547–569 (1975).
21. Suder, K. & Worgotter, F. The control of low-level information flow in the visual system. *Rev. Neurosci.* **11**, 127–146 (2000).
22. Pasupathy, A. & Connor, C. E. Responses to contour features in macaque area v4. *J. Neurophysiol.* **82**, 2490–2502 (1999).
23. Braun, J. Shape-from-shading is independent of visual attention and may be a 'texton'. *Spat. Vis.* **7**, 311–322 (1993).
24. Sun, J. & Perona, P. Early computation of shape and reflectance in the visual system. *Nature* **379**, 165–168 (1996).
25. Logothetis, N. K., Pauls, J. & Poggio, T. Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.* **5**, 552–563 (1995).
26. Bar, M. & Biederman, I. Localizing the cortical region mediating visual awareness of object identity. *Proc. Natl Acad. Sci. USA* **96**, 1790–1793 (1999).
27. Vogels, R., Biederman, I., Bar, M. & Lorincz, A. Inferior temporal neurons show greater sensitivity to nonaccidental than metric differences. *J. Cogn. Neurosci.* (in the press).
[Author: update?]
28. Kreiman, G., Koch, C. & Fried, I. Category-specific visual responses of single neurons in the human medial temporal lobe. *Nature Neurosci.* **3**, 946–953 (2000).
29. He, Z. J. & Nakayama, K. Perceiving textures: beyond filtering. *Vision Res.* **34**, 151–162 (1994).
30. Treue, S. & Maunsell, J. H. Attentional modulation of visual motion processing in cortical areas mt and mst. *Nature* **382**, 539–541 (1996).
31. Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol. (Lond.)* **160**, 106–154 (1962).
32. DeSchepper, B. & Treisman, A. Visual memory for novel shapes: implicit coding without attention. *J. Exp. Psychol. Learn. Mem. Cogn.* **22**, 27–47 (1996).
33. Lee, D. K., Itti, L., Koch, C. & Braun, J. Attention activates winner-take-all competition among visual filters. *Nature Neurosci.* **2**, 375–381 (1999).
A detailed neural model is used to quantitatively predict attentional modulation of psychophysical pattern discrimination performance in terms of intensified competition between visual neurons.
34. Yeshurun, Y. & Carrasco, M. Attention improves or impairs visual performance by enhancing spatial resolution. *Nature* **396**, 72–75 (1998).
35. Mack, A., Tang, B., Tuma, R., Kahn, S. & Rock, I. Perceptual organization and attention. *Cogn. Psychol.* **24**, 475–501 (1992).
36. Moore, C. M. & Egeth, H. Perception without attention: evidence of grouping under conditions of inattention. *J. Exp. Psychol. Hum. Percept. Perform.* **23**, 339–352 (1997).
37. Motter, B. C. Neural correlates of attentive selection for color or luminance in extrastriate area v4. *J. Neurosci.* **14**, 2178–2189 (1994).
38. Treue, S. & Trujillo, J. C. M. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature* **399**, 575–579 (1999).
Investigates two types of feedback attentional modulation: spatial-based, and non-spatial but feature-based.
39. Barcelo, F., Suwazono, S. & Knight, R. T. Prefrontal modulation of visual processing in humans. *Nature Neurosci.* **3**, 399–403 (2000).
40. Moran, J. & Desimone, R. Selective attention gates visual processing in the extrastriate cortex. *Science* **229**, 782–784 (1985).
41. Niebur, E., Koch, C. & Rosin, C. An oscillation-based model for the neuronal basis of attention. *Vision Res.* **33**, 2789–2802 (1993).
42. Chawla, D., Rees, G. & Friston, K. J. The physiological basis of attentional modulation in extrastriate visual areas. *Nature Neurosci.* **2**, 671–676 (1999).
43. Reynolds, J. H., Pasternak, T. & Desimone, R. Attention increases sensitivity of v4 neurons. *Neuron* **26**, 703–714 (2000).
44. Dosher, B. A. & Lu, Z. L. Mechanisms of perceptual attention in precuing of location. *Vision Res.* **40**, 1269–1292 (2000).
45. Itti, L., Koch, C. & Braun, J. Revisiting spatial vision: Towards a unifying model. *J. Opt. Soc. Am. A* **17**, 1899–1917 (2000).
46. Carrasco, M., Penpeci-Talgar, C. & Eckstein, M. Spatial covert attention increases contrast sensitivity across the csf: support for signal enhancement. *Vision Res.* **40**, 1203–1215 (2000).

47. Deco, G. & Zihl, J. A neurodynamical model of visual attention: Feedback enhancement of spatial resolution in a hierarchical system. *J. Comp. Neurosci.* (in the press). **[Author: update?]**
48. Daugman, J. G. Spatial visual channels in the fourier plane. *Vision Res.* **24**, 891–910 (1984).
49. Palmer, L. A., Jones, J. P. & Stepnoski, R. A. in *The Neural Basis of Visual Function* (ed. Leventhal, A. G.) 246–265 (CRC, Boca Raton,Florida, 1991).
50. Zetzsche, C. *et al.* Investigation of a sensorimotor system for saccadic scene analysis: an integrated approach. *Proc. 5th Int. Conf. Simulation Adaptive Behav.* **5**, 120–126 (1998).
51. Reinagel, P. & Zador, A. M. Natural scene statistics at the centre of gaze. *Network Comp. Neural Sys.* **10**, 341–350 (1999).
52. Barth, E., Zetzsche, C. & Rentschler, I. Intrinsic two-dimensional features as textons. *J. Opt. Soc. Am. A Opt. Image. Sci. Vis.* **15**, 1723–1732 (1998).
53. Nothdurft, H. Salience from feature contrast: additivity across dimensions. *Vision Res.* **40**, 1183–1201 (2000). **Studies psychophysically how orientation, motion, luminance and colour contrast cues combine to yield the saliency of visual stimuli.**
54. Wolfe, J. M. Visual search in continuous, naturalistic stimuli. *Vision Res.* **34**, 1187–1195 (1994).
55. Braun, J. Vision and attention: the role of training. *Nature* **393**, 424–425 (1998).
56. Ahissar, M. & Hochstein, S. The spread of attention and learning in feature search: effects of target distribution and task difficulty. *Vision Res.* **40**, 1349–1364 (2000).
57. Sigman, M. & Gilbert, C. D. Learning to find a shape. *Nature Neurosci.* **3**, 264–269 (2000).
58. Itti, L. & Koch, C. Feature combination strategies for saliency-based visual attention systems. *J. Electronic Imaging* (in the press). **[Author: update?]**
59. Wolfe, J. in *Attention* (ed. Pashler, H.) **[Author: page no's]** (University College London, London, 1996).
60. Carandini, M. & Heeger, D. J. Summation and division by neurons in primate visual cortex. *Science* **264**, 1333–1336 (1994).
61. Nothdurft, H. C. Texture discrimination by cells in the cat lateral geniculate nucleus. *Exp. Brain Res.* **82**, 48–66 (1990).
62. Allman, J., Miezin, F. & McGuinness, E. Stimulus specific responses from beyond the classical receptive field: neurophysiological mechanisms for local-global comparisons in visual neurons. *Annu. Rev. Neurosci.* **8**, 407–430 (1985). **One of the first reports that activity of a visual neuron can be modulated by the presence of distant stimuli, far outside the neuron's receptive field.**
63. Cannon, M. W. & Fullenkamp, S. C. Spatial interactions in apparent contrast: inhibitory effects among grating patterns of different spatial frequencies, spatial positions and orientations. *Vision Res.* **31**, 1985–1998 (1991).
64. Sillito, A. M., Grieve, K. L., Jones, H. E., Cudeiro, J. & Davis, J. Visual cortical mechanisms detecting focal orientation discontinuities. *Nature* **378**, 492–496 (1995).
65. Levitt, J. B. & Lund, J. S. Contrast dependence of contextual effects in primate visual cortex. *Nature* **387**, 73–76 (1997).
66. Gilbert, C. D. & Wiesel, T. N. Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. *J. Neurosci.* **9**, 2432–2442 (1989).
67. Gilbert, C., Ito, M., Kapadia, M. & Westheimer, G. Interactions between attention, context and learning in primary visual cortex. *Vision Res.* **40**, 1217–1226 (2000).
68. Ben-Av, M. B., Sagi, D. & Braun, J. Visual attention and perceptual grouping. *Percept. Psychophys.* **52**, 277–294 (1992).
69. Grossberg, S. & Raizada, R. D. Contrast-sensitive perceptual grouping and object-based attention in the laminar circuits of primary visual cortex. *Vision Res.* **40**, 1413–1432 (2000).
70. Vinje, W. E. & Gallant, J. L. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* **287**, 1273–1276 (2000).
71. Itti, L., Koch, C. & Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Patt. Anal. Mach. Intell.* **20**, 1254–1259 (1998).
72. Tsotsos, J. K. *et al.* Modeling visual-attention via selective tuning. *Artif. Intell.* **78**, 507–545 (1995).
73. Milanese, R., Gil, S. & Pun, T. Attentive mechanisms for dynamic and static scene analysis. *Opt. Eng.* **34**, 2428–2434 (1995).
74. Itti, L. & Koch, C. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res.* **40**, 1489–1506 (2000).
75. Toet, A., Bijl, P., Kooi, F. L. & Valeton, J. M. *A High-Resolution Image Dataset for Testing Search and Detection Models (TNO-TM-98–A020)* (TNO Human Factors Research, Institute, Soesterberg, The Netherlands, 1998).
76. Hamker, F. H. in *Proc. 5th Neural Comp. Psychol. Workshop (NCPW'98)* (eds von Heinke, D., Humphreys, G. W. & Olson, A.) **[Author: page no's]** (Springer Verlag, London, 1999).
77. Laberge, D. & Buchsbaum, M. S. Positron emission tomographic measurements of pulvinar activity during an attention task. *J. Neurosci.* **10**, 613–619 (1990).
78. Robinson, D. L. & Petersen, S. E. The pulvinar and visual salience. *Trends Neurosci.* **15**, 127–132 (1992).
79. Kustov, A. A. & Robinson, D. L. Shared neural control of attentional shifts and eye movements. *Nature* **384**, 74–77 (1996).
80. Gottlieb, J. P., Kusunoki, M. & Goldberg, M. E. The representation of visual salience in monkey parietal cortex. *Nature* **391**, 481–484 (1998). **Electrophysiological experiments in the awake monkey indicating that some neurons might explicitly encode for saliency in the posterior parietal cortex.**
81. Colby, C. L. & Goldberg, M. E. Space and attention in parietal cortex. *Annu. Rev. Neurosci.* **22**, 319–349 (1999).
82. Thompson, K. G. & Schall, J. D. Antecedents and correlates of visual detection and awareness in macaque prefrontal cortex. *Vision Res.* **40**, 1523–1538 (2000).
83. Andersen, R. A., Bracewell, R. M., Barash, S., Gnadt, J. W. & Fogassi, L. Eye position effects on visual, memory, and saccade-related activity in areas lip and 7a of macaque. *J. Neurosci.* **10**, 1176–1196 (1990).
84. Pouget, A. & Sejnowski, T. J. Spatial transformations in the parietal cortex using basis functions. *J. Cogn. Neurosci.* **9**, 222–237 (1997).
85. Blaser, E., Sperling, G. & Lu, Z. L. Measuring the amplification of attention. *Proc. Natl Acad. Sci. USA* **96**, 11681–11686 (1999).
86. Brefczynski, J. A. & DeYoe, E. A. A physiological correlate of the 'spotlight' of visual attention. *Nature Neurosci.* **2**, 370–374 (1999).
87. Amari, S. & Arbib, M. A. in *Systems Neuroscience* (ed. Metzler, J.) 119–165 (Academic, **[Author; City?]**, 1977).
88. Posner, M. I. & Cohen, Y. in *Attention and Performance* Vol. X (eds Bouma, H. & Bouwhuis, D.) 531–556 (Erlbaum, **[Author; City?]**, 1984).
89. Kwak, H. W. & Egeth, H. Consequences of allocating attention to locations and to other attributes. *Percept. Psychophys.* **51**, 455–464 (1992).
90. Klein, R. M. Inhibition of return. *Trends Cogn. Sci.* **4**, 138–147 (2000). **A very complete review of IOR.**
91. Shimojo, S., Tanaka, Y. & Watanabe, K. Stimulus-driven facilitation and inhibition of visual information processing in environmental and retinotopic representations of space. *Brain Res. Cogn. Brain Res.* **5**, 11–21 (1996).
92. Kingstone, A. & Pratt, J. Inhibition of return is composed of attentional and oculomotor processes. *Percept. Psychophys.* **61**, 1046–1054 (1999).
93. Taylor, T. L. & Klein, R. M. Visual and motor effects in inhibition of return. *J. Exp. Psychol. Hum. Percept. Perform.* **26**, 1639–1656 (2000).
94. Tipper, S. P., Driver, J. & Weaver, B. Object-centred inhibition of return of visual attention. *Q. J. Exp. Psychol. A* **43**, 289–298 (1991).
95. Gibson, B. S. & Egeth, H. Inhibition of return to object-based and environment-based locations. *Percept. Psychophys.* 55, 323–339 (1994).
96. Ro, T. & Rafal, R. D. Components of reflexive visual orienting to moving objects. *Percept. Psychophys.* **61**, 826–836 (1999).
97. Becker, L. & Egeth, H. Mixed reference frames for dynamic inhibition of return. *J. Exp. Psychol. Hum. Percept. Perform.* **26**, 1167–1177 (2000).
98. Horowitz, T. S. & Wolfe, J. M. Visual search has no memory. *Nature* **394**, 575–577 (1998).
99. Mozer, M & Sitton, S. in *Attention* **[Author: eds?] [Author: page numbers]** (University College London, London, 1996)
100. Guigon, E., Grandguillaume, P., Otto, I., Boutkhil, L. & Burnod, Y. Neural network models of cortical functions based on the computational properties of the cerebral cortex. *J. Physiol. (Paris)* **88**, 291–308 (1994).
101. Schill, K., Umkehrer, E., Beinlich, S., Krieger, G. & Zetzsche, C. Scene analysis with saccadic eye movements: top-down and bottom-up modeling. *J. Electronic Imaging* (in the press). **[Author: update?]**
102. Rybak, I. A., Gusakova, V. I., Golovan, A. V., Podladchikova, L. N. & Shevtsova, N. A. A model of attention-guided visual perception and recognition. *Vision Res.* **38**, 2387–2400 (1998).
103. Deco, G. & Schurmann, B. A hierarchical neural system with attentional top-down enhancement of the spatial resolution for object recognition. *Vision Res.* **40**, 2845–2859 (2000).
104. Stark, L. W. & Choi, Y. S. in *Visual Attention and Cognition* (eds Zangemeister, W. H., Stiehl, H. S. & Freska, C.) 3–69 (Elsevier Science B. V., **[Author; City?]**, 1996).
105. Stark, L. W. *et al.* Representation of human vision in the brain: how does human perception recognize images? *J. Electronic Imaging* (in the press). **[Author: update?]**
106. Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. *Nature Neurosci.* **2**, 1019–1025 (1999).
107. Riesenhuber, M. & Poggio, T. Models of object recognition. *Nature Neurosci.* **S3**, 1199–1204 (2000).
108. O'Craven, K. M., Downing, P. E. & Kanwisher, N. fmri evidence for objects as the units of attentional selection. *Nature* **401**, 584–587 (1999).
109. Roelfsema, P. R., Lamme, V. A. & Spekreijse, H. Object-based attention in the primary visual cortex of the macaque monkey. *Nature* **395**, 376–381 (1998).
110. Abrams, R. A. & Law, M. B. Object-based visual attention with endogenous orienting. *Percept. Psychophys.* **62**, 818–833 (2000).
111. Corbetta, M. Frontoparietal cortical networks for directing attention and the eye to visual locations: identical, independent, or overlapping neural systems? *Proc. Natl Acad. Sci. USA* **95**, 831–838 (1998).
112. Webster, M. J. & Ungerleider, L. G. in *The Attentive Brain* (ed. Parasuraman, R.) 19–34 (MIT, Cambridge, Massachusetts, 1998).
113. Shepherd, M., Findlay, J. M. & Hockey, R. J. The relationship between eye movements and spatial attention. *Q. J. Exp. Psychol.* **38**, 475–491 (1986).
114. Sheliga, B. M., Riggio, L. & Rizzolatti, G. Orienting of attention and eye movements. *Exp. Brain Res.* **98**, 507–22 (1994).
115. Hoffman, J. E. & Subramaniam, B. The role of visual attention in saccadic eye movements. *Percept. Psychophys.* **57**, 787–795 (1995).
116. Kowler, E., Anderson, E., Dosher, B. & Blaser, E. The role of attention in the programming of saccades. *Vision Res.* **35**, 1897–916 (1995).
117. Schall, J. D., Hanes, D. P. & Taylor, T. L. Neural control of behavior: countermanding eye movements. *Psychol. Res.* **63**, 299–307 (2000).
118. Nobre, A. C., Gitelman, D. R., Dias, E. C. & Mesulam, M. M. Covert visual spatial orienting and saccades: overlapping neural systems. *Neuroimage* **11**, 210–216 (2000).
119. Dominey, P. F. & Arbib, M. A. A cortico-subcortical model for generation of spatially accurate sequential saccades. *Cereb. Cortex* **2**, 153–175 (1992). **A complete neural network model for the control of saccadic eye movements.**
120. Motter, B. C. & Belky, E. J. The guidance of eye movements during active visual search. *Vision Res.* **38**, 1805–1815 (1998).
121. Gilchrist, I. D., Heywood, C. A. & Findlay, J. M. Saccade selection in visual search: evidence for spatial frequency specific between-item interactions. *Vision Res.* **39**, 1373–1383 (1999).
122. Wolfe, J. M. & Gancarz, G. in *Basic and Clinical Applications of Vision Science* (ed. Lakshminarayanan, V.) 189–192 (Kluwer Academic, Dordrecht, The Netherlands, 1996).

Acknowledgements

The research carried out in the laboratories of the authors on visual attention is supported by the National Science Foundation, the National Institute of Mental Health and the Office of Naval Research.

Bios

Laurent Itti received his M.S. in Image Processing from the Ecole Nationale Superieure des Telecommunications (Paris) in 1994, and Ph.D. in Computation and Neural Systems from Caltech (Los Angeles) in 2000. In September 2000, he became assistant professor of Computer Science at the University of Southern California. His primary research interest is in biologically plausible computational brain modelling, and in the comparison of model simulations to empirical measurements from living systems. Of particular interest in his laboratory is the development of computational models of biological vision with applications to machine vision.

Professor Christof Koch has a Ph.D. in Physics from the University of Tubingen, Germany. His research focuses on understanding the biophysical mechanisms underlying information processing in individual nerve cells as well as the neuronal operations underlying spatial vision, motion, shape perception and visual attention in the primate visual system, using electrophysiological, brain imaging, psychophysical and computational tools. Together with Dr Francis Crick, he works on the neuronal basis of visual consciousness.

Summary

- We review recent work on computational models of focal visual attention, with emphasis on the bottom-up, saliency- or image-based control of attentional deployment. We highlight five important trends that have emerged from the computational literature:
- First, the perceptual saliency of stimuli critically depends on surrounding context; that is, a same object may or may not appear salient depending on the nature and arrangement of other objects in the scene. Computationally, this means that contextual influences, such as non-classical surround interactions, must be included in models.
- Second, a unique 'saliency map' topographically encoding for stimulus conspicuity over the visual scene has proved to be an efficient and plausible bottom-up control strategy. Many successful models are based on such architecture, and electrophysiological as well as psychophysical studies have recently supported the idea that saliency is explicitly encoded in the brain.
- Third, inhibition-of-return (IOR), the process by which the currently attended location is prevented from being attended again, is a critical element of attentional deployment. Without IOR, indeed, attention would endlessly be attracted towards the most salient stimulus. IOR thus implements a memory of recently visited locations, and allows attention to thoroughly scan our visual environment.
- Fourth, attention and eye movements tightly interplay, posing computational challenges with respect to the coordinate system used to control attention. Understanding the interaction between overt and covert attention is particularly important for models concerned with visual search.
- Last, scene understanding and object recognition strongly constrain the selection of attended locations. Although several models have approached, in an information-theoretical sense, the problem of optimally deploying attention to analyse a scene, biologically plausible implementations of such a computational strategy remain to be developed.

Links

Supplementary information for Figure 2
http://ilab.usc.edu/itti/nrn/

Learning to Detect A Salient Object

Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, *Fellow, IEEE,* Xiaoou Tang, *Fellow, IEEE,* and Heung-Yeung Shum, *Fellow, IEEE*

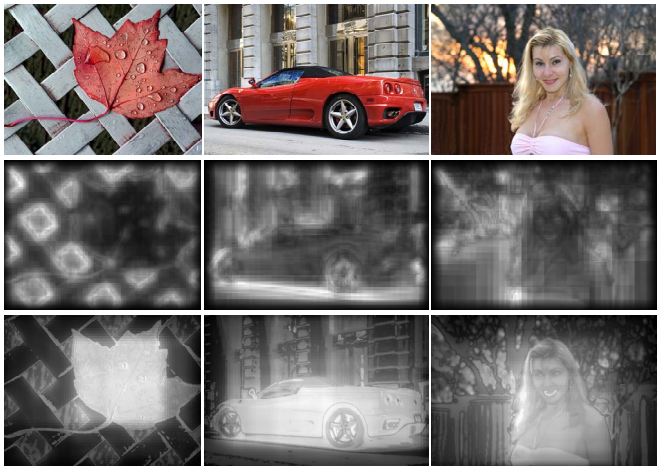
Abstract—In this paper, we study the salient object detection problem for images. We formulate this problem as a binary labeling task, where we separate the salient object from the background. We propose a set of novel features including multi-scale contrast, center-surround histogram, and color spatial distribution to describe a salient object locally, regionally, and globally. A conditional random field is learned to effectively combine these features for salient object detection. Further, we extend the proposed approach to detect a salient object from sequential images, by introducing the dynamic salient features. We collected a large image database containing tens of thousands of carefully labeled images by multiple users and a video segment database, and conducted a set of experiments over them to demonstrate the effectiveness of the proposed approach.

Index Terms—Salient object detection, conditional random field, visual attention, saliency map.

1 Introduction

The human brain and visual system pay more attention to some parts of an image. Visual attention has been studied by researchers in physiology, psychology, neural systems, and computer vision for a long time. There are many applications for visual attention, for example, automatic image cropping [1], adaptive image display on small devices [2], image/video compression, advertising design [3], and image collection browsing [4]. Recent studies [5]–[7] demonstrated that visual attention helps object recognition, tracking, and detection as well. In this paper, we study one aspect of visual attention — salient object detection. Fig. 1 shows some examples of salient objects.

For instance, people are usually interested in the objects in images from Fig. 1, and the leaf, car, and woman attract the most visual attention in each respective image. We call them salient objects, or foreground objects that we are familiar with, or objects with the most interest. In many applications, such as image display on small devices [2] and image collection browsing [4], people want to show the regions with the most interest, or the salient objects. In this paper, we try to locate these salient objects automatically with the supposition that a salient object exists in an image.



(a) (b) (c)

Fig. 1. Salient object detection. From top to bottom: input image with a salient object, saliency map computed by Itti's attention algorithm (http://www.saliencytoolbox.net), and saliency map computed by our salient object detection approach.

1.1 Related work

Most existing visual attention approaches are based on the bottom-up computational framework [8]–[16], where visual attention is supposed to be driven by low-level stimulus in the scene, such as intensity, contrast, and motion. These approaches consist of the following three steps. The first step is *feature extraction*, in which multiple low-level visual features, such as intensity, color, orientation, texture and motion are extracted from the image at multiple scales. The second step is *saliency computation*. The saliency is computed by a center-surround operation [13], self-information [8], or graph-based random walk [9] using multiple features. After normalization and linear/non-linear combination, a master map [17] or a saliency map [14] is computed to represent the saliency of each image pixel. Last, a few key locations on the saliency map are identified by winner-take-all, or inhibition-of-return, or other non-linear operations. Recently, a saliency model based on low, middle and high-level image features is trained using the collected eye

Thanks for the helpful comments from reviewers.
Tie Liu is with Xi'an Jiaotong University and IBM Research China. Zejian Yuan and Nanning Zheng are with Xi'an Jiaotong University, P.R.China. Jian Sun, Jingdong Wang and Heung-Yeung Shum are with Microsoft Research Asia. Xiaou Tang is with the Chinese University of HongKong.

Digital Object Indentifier 10.1109/TPAMI.2010.70 0162-8828/10/$26.00 © 2010 IEEE

tracking data [18]. While these approaches have worked well in finding a few fixation locations in synthetic and natural images, they have not been able to accurately detect where the salient object should be.

For instance, the middle row in Fig. 1 shows three saliency maps computed using Itti's algorithm [13]. Notice that the visual saliency concentrates on several small local regions with high contrast structures, e.g., the background grid in (a), the shadow in (b), and the foreground boundary in (c). Although the leaf in (a) commands much attention, the saliency for the leaf is low. Therefore, these saliency maps computed from low-level features don't have the notation of objects, and they are not good indications for where a salient object is located while perusing these images.

Figure-ground segregation is somehow related to salient object detection. However, the usually figure-ground segregation algorithm works with the supposition of the category of objects [19] [20] [21] or with interactions [22] [23]. If the object is assigned with a given category, the specific features, for example for cows, can be defined specially, and these features cannot be adopted for other categories. For interactive figure-ground segmentation, the appearance model is usually set up, where for our salient object detection we have not such an appearance model.

Visual attention is also studied for sequential images, where the spatiotemporal cues from image sequences are indicated to be helpful for visual attention detection. For instance, motion from objects or backgrounds helps to indicate the salient fixations [24] [25] [26]. Large motion [27] and motion contract [24] are supposed to induce prominent attention respectively. Usually, the visual saliency from a single image is combined with the motion saliency for better visual attention detection, and different combination strategies are introduced in [27]. Video surprising [11] is also related, where it describes the KullbackLeibler divergence between the prior and posterior distribution of a feature map. These visual attention approaches suffer from the similar shortcoming to the visual attention approaches for single image. Automatic object discovery [28] [29] [30] deals with a similar salient object detection task for sequential images. The objects are extracted and tracked using motion-based layer segmentation in [28], and a generative model of objects by defining switch variables for combinatorial model selection is adopted in [29]. The unsupervised video object discovery [30] combines the topic model and the temporal model for videos.

1.2 Our approach

In this paper, we investigate one aspect of visual attention, namely, salient object detection. We incorporate the high level concept of the salient object into the process of saliency map computation. As can be observed in Fig. 2, people naturally pay more attention to salient objects in images such as a person, a face, a car, an animal, or a road sign. Therefore, we formulate *salient object detection* as a binary labeling problem that separates a salient object from the background. Like face detection, we learn to detect a familiar object; unlike face detection, we detect a familiar yet unknown object in an image.

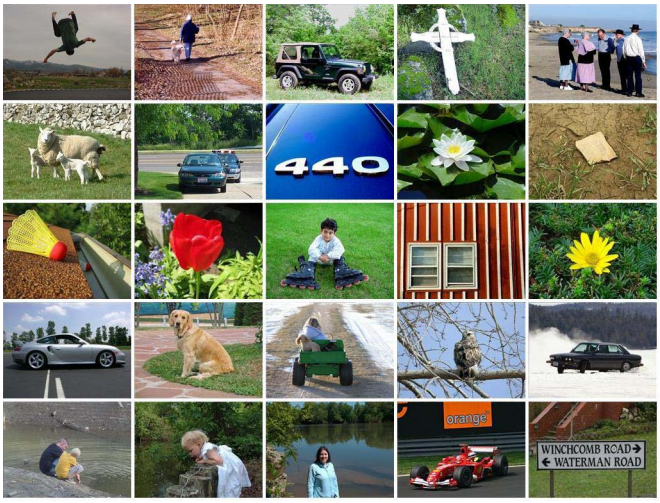


Fig. 2. Sample images in our image database for salient object detection.

We present a supervised approach to learn to detect a salient object in an image or sequential images. First, we model the salient object detection problem by a condition random field (CRF), where a group of salient features are combined through CRF learning. Moreover, the segmentation is also incorporated into the CRF to detect a salient object with unknown size and shape. The last row in Fig. 1 shows the saliency maps computed by our approach. Second, to overcome the challenge that we do not know what a specific object or object category is, we propose a set of novel local, regional, and global salient features to define a generic salient object. We also define the salient features on the motion field similarly to capture the spatiotemporal cues. Then, we construct a large image database with 20,000+ well labeled images for training and evaluation. To the best of our knowledge, it is the first time a large image database has been made available for quantitative evaluation.

The remainder of the paper is organized as follows. Section 2 introduces the formulation of the salient object detection problem, and the salient object features are presented in section 3. Section 4 introduces the image database and the evaluation experiments. Section 5 discuss the connections between our approach and related approaches, and the conclusion follows in section 6.

2 Formulation

Given an image I, we represent the salient object as a binary mask $A = \{a_x\}$. For each pixel x, $a_x \in \{1, 0\}$ is a binary label to indicate whether the pixel x belongs to the salient object. Similarly, the salient objects in sequential images, $\{I_1, \cdots, I_t, \cdots, I_N\}$, are represented

by a sequence of binary masks $\{A_1, \cdots, A_t, \cdots, A_N\}$, with A_t corresponding to image I_t.

In this paper, we formulate the salient object detection problem as a binary labeling task by inspecting whether each pixel belongs to the salient object. We first present the conditional random field formulation to the single image case, and then extend it to the sequential image case by exploring the extra temporal information.

2.1 Formulation of Salient Object Detection in a Single Image

In the CRF framework [31], the probability of a labeling configuration $A = \{a_x\}$, given the observation image I, is modeled as a conditional distribution $P(A|I) = \frac{1}{Z}\exp(-E(A|I))$, where Z is the partition function. We define the energy $E(A|I)$ as a linear combination of a set of static salient features, including a number of K unary features $F_k(a_x, I)$ and a pairwise feature $S(a_x, a_{x'}, I)$:

$$E(A|I) = \sum_x \sum_{k=1}^{K} \lambda_k F_k(a_x, I) + \sum_{x,x'} S(a_x, a_{x'}, I), \quad (1)$$

where λ_k is the weight of the kth feature, and x, x' are two adjacent pixels. Compared with Markov random field, one of the advantages of CRF is that the features $F_k(a_x, I)$ and $S(a_x, a_{x'}, I)$ can be arbitrary low-level or high-level features extracted from the whole image. CRF also provides an elegant framework to learn an optimal combination of multiple features.

Salient object feature. $F_k(a_x, I)$ indicates whether a pixel x belongs to the salient object. In the next section, we propose a set of local, regional, and global salient object features. The salient object feature $F_k(a_x, I)$ is formulated from a normalized feature map $f_k(x, I) \in [0, 1]$ for every pixel, and is written as follows:

$$F_k(a_x, I) = \begin{cases} f_k(x, I) & a_x = 0 \\ 1 - f_k(x, I) & a_x = 1. \end{cases} \quad (2)$$

Pairwise feature. $S(a_x, a_{x'}, I)$ exploits the spatial relationship between two adjacent pixels. Following the contrast-sensitive potential function in interactive image segmentation [22], we define $S(a_x, a_{x'}, I)$ as:

$$S(a_x, a_{x'}, I) = |a_x - a_{x'}| \cdot \exp(-\beta d_{x,x'}), \quad (3)$$

where $d_{x,x'} = \|I_x - I_{x'}\|_2$ is the L2 norm of the color difference, β is a robust parameter that weights the color contrast and can be set as $\beta = (2\langle||I_x - I_{x'}||^2\rangle)^{-1}$ [32], with $\langle\cdot\rangle$ being the expectation operator. This feature function can be viewed as a penalty term when adjacent pixels are assigned with different labels. The more similar the colors of the two pixels are, the less likely they are assigned different labels.

2.2 Formulation of Salient Object Detection in Sequential Images

We exploit the extra temporal cues to formulate salient object detection in sequential images. Besides the static salient features from a single image, the temporal features, called dynamic features, are further defined. Different from previous work [24] [25] [26], we propose new dynamic features and learn a CRF model to combine the dynamic features and static features. Instead of building a complex 3D graph formulation, e.g., a large graph in interactive video cutout [33] [34], we integrate the cues from multiple images into a 2D graph for effective and efficient optimization.

Given the sequential images $\{I_t\}$, $t \in \{1, \cdots, N\}$, the probability of the sequential binary maps, $\{A_t\}$, $t \in \{1, \cdots, N\}$, can be modeled as a conditional distribution:

$$P(A_{1,\cdots,N}|I_{1,\cdots,N}) = \frac{1}{Z}\exp(-E(A_{1,\cdots,N}|I_{1,\cdots,N})), \quad (4)$$

where Z is the partition function. A reasonable supposition is that the salient object detection A_t can be inferred from the associated frame I_t and the previous frame I_{t-1}. Then the energy function $E(A_{1,\cdots,N}|I_{1,\cdots,N})$ can be decomposed as:

$$E(A_{1,\cdots,N}|I_{1,\cdots,N}) = \sum_{t=1}^{N} E(A_t|I_{1,\cdots,N}) = \sum_{t=1}^{N} E(A_t|I_{t-1}, I_t). \quad (5)$$

Here $E(A_t|I_{t-1}, I_t)$ is composed of a static term and a dynamic term. The static term is the same as the single image case. In the dynamic term, we compute a motion field M_t from a pair of successive images I_{t-1} and I_t, and build salient features from the motion field, and in addition introduce an appearance coherent feature between the salient objects in the successive frames. Specifically, the energy $E(A_t|I_{t-1}, I_t)$ is formulated as a linear combination of static salient features $F_k(a_x, I_t)$, a pairwise feature $S(a_x, a_{x'}, I)$ and a set of dynamic salient features, including motion salient features $F_k(a_x, M_t)$ and appearance coherent features $F_k(a_x, I_{t-1}, I_t)$:

$$E(A_t|I_{t-1}, I_t) = \sum_x \Big(\sum_{k=1}^{K} \lambda_k F_k(a_x, I_t) + \sum_{k=K+1}^{K+L} \lambda_k F_k(a_x, M_t) + \lambda_0 F(a_x, I_{t-1}, I_t)\Big) + \sum_{x,x'} S(a_x, a_{x'}, I_t), \quad (6)$$

where $\{\lambda_k\}$ are the weights of the features, M_t is the motion field corresponding to image I_t, and x, x' are two adjacent pixels in image I_t. $F_k(a_x, I_t)$ are the static salient features, and $S(a_x, a_{x'}, I_t)$ describes the spatial relationship between two adjacent pixels. These two categories of features are defined as in (1). Different from (1), more features from the temporal information are included, the motion salient features $F_k(a_x, M_t)$ from the motion field M_t, and the appearance coherent feature $F(a_x, I_{t-1}, I_t)$ between the salient objects from two adjacent frames.

Motion salient feature. $F_k(a_x, M_t)$ is defined, similar to (2), as the indicator of a normalized feature map $f_k(x, M_t) \in [0, 1]$, where M_t is the motion field of the image I_t and obtained based on the SIFT flow technique [35].

Appearance coherent feature. $F(a_x, I_{t-1}, I_t)$ models the appearance coherence of the salient objects from two adjacent frames, which is defined as an indicator of a normalized feature map $f(x, I_{t-1}, I_t) \in [0, 1]$, similar to (2). This feature function $f(x, I_{t-1}, I_t)$ penalizes the pixels that are identified to be in the salient object but with a large color difference between the surrounding regions from two adjacent frames. With this appearance coherent feature, the salient objects from two adjacent frames can be labeled more consistently.

2.3 Learning and inference for the CRF model

The objective functions of the salient object detection for single image and sequential image cases, in (1) and (6), are essentially very similar to the perspective of the CRF formulation, i.e., a linear combination of a set of features. To get the linear combination of features, the goal of CRF learning is to estimate the linear weights $\overrightarrow{\lambda} = \{\lambda_k\}$ under the Maximized Likelihood (ML) criteria. In the following, we present the parameter learning scheme for the single image case. The parameter learning scheme for the sequential image case can be similarly obtained. Given N training image pairs $\{I^n, A^n\}_{n=1}^N$, the optimal parameters maximize the sum of the log-likelihood:

$$\overrightarrow{\lambda}^* = \arg\max_{\overrightarrow{\lambda}} \sum_n \log P(A^n | I^n; \overrightarrow{\lambda}). \tag{7}$$

The derivative of the log-likelihood with respect to the parameter λ_k is the difference between two expectations:

$$\frac{d \log P(A^n | I^n; \overrightarrow{\lambda})}{d\lambda_k} = \\ < F_k(A^n, I^n) >_{P(A^n | I^n; \overrightarrow{\lambda})} - < F_k(A^n, I^n) >_{P(A^n | G^n)} . \tag{8}$$

Then, the gradient descent direction is:

$$\Delta\lambda_k \propto \sum_n (\sum_{x, a_x^n} (F_k(a_x^n, I^n) p(a_x^n | I^n; \overrightarrow{\lambda}) - F_k(a_x^n, I^n) p(a_x^n | g_x^n))), \tag{9}$$

where $p(a_x^n | I^n; \overrightarrow{\lambda}) = \int_{A^n \setminus a_x^n} P(A_x^n | I^n; \overrightarrow{\lambda})$ is the marginal distribution. $p(a_x^n | g_x^n)$ is from the labeled ground-truth g_x^n, and it is defined as:

$$p(a_x^n | g_x^n) = \begin{cases} 1 - g_x^n & a_x = 0 \\ g_x^n & a_x = 1 \end{cases} . \tag{10}$$

Exact computation of marginal distribution $p(a_x^n | I^n; \overrightarrow{\lambda})$ is intractable. However, the pseudo-marginal (belief) computed by belief propagation can be used as a good approximation [36] [19]. The tree-reweighted belief propagation [37] can be run under the current parameters in each step of gradient descent to compute an approximation of the marginal distribution $p(a_x^n | I^n; \overrightarrow{\lambda})$.

When the combination parameters of salient features are learned, we can infer the most probable labeling A to minimize the energy from (1) and (6). We still apply the tree-reweighted belief propagation to infer the label using the learned parameters, and we will discuss the details of implementations in section 4.

Fig. 3. Multi-scale contrast. From left to right: input image, contrast maps at multiple scales, and the feature map from linearly combining the contrasts at multiple scales.

3 Salient Object Feature

In this section, we instantiate the formulation of salient object detection by presenting the salient object features: static salient features for the single image case and dynamic salient features specifically for the sequential images.

3.1 Static Salient Feature

We introduce local, regional, and global features that define a salient object. Since the scale selection is one of the fundamental issues in feature extraction, we resize all images so that the max(width,height) of the image is 400 pixels. In the following, all parameters are set with respect to this basic image size.

3.1.1 Multi-scale contrast

Contrast is the most commonly used local feature for attention detection [13], [38], [39] because the contrast operator simulates the human visual receptive fields. Without knowing the size of the salient object, contrast is usually computed at multiple scales. In this paper, we simply define the multi-scale contrast feature $f_c(x, I)$ as a linear combination of contrasts in the Gaussian image pyramid:

$$f_c(x, I) = \sum_{l=1}^{L} \sum_{x' \in N(x)} ||I^l(x) - I^l(x')||^2 \tag{11}$$

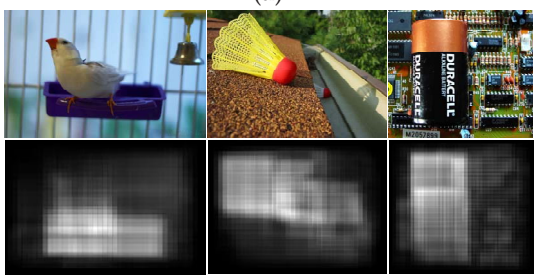
where I^l is the lth-level image in the pyramid and the number of pyramid levels L is 6. $N(x)$ is a 9×9 window. The feature map $f_c(\cdot, I)$ is normalized to a fixed range $[0, 1]$. An example is shown in Fig. 3. Multi-scale contrast highlights the high contrast boundaries by giving low scores to the homogenous regions inside the salient object.

3.1.2 Center-surround histogram

As shown in Fig. 2, the salient object usually has a larger extent than local contrast and can be distinguished from its surrounding context. Therefore, we propose a regional salient feature.

Suppose the salient object is enclosed by a rectangle R. We construct a surrounding contour R_S with the same area of R, as shown in Fig. 4 (a). To measure

(a)



(b)

Fig. 4. Center-surround histogram. (a) center-surround histogram distances with different locations and sizes. (b) top row are input images and bottom row are center-surround histogram feature maps.

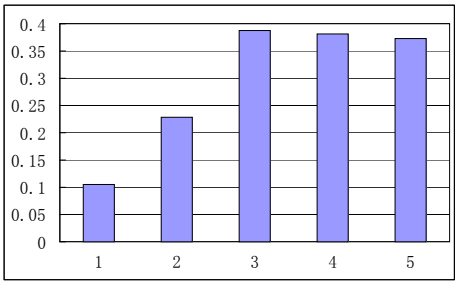


Fig. 5. The average center-surround histogram distance on the image set $\mathcal{A}$. 1. a randomly selected rectangle. 2. a rectangle centered at the image center with 55% ratio of area to image. 3-5. rectangles labeled by three users.

(a) (b) (c)

Fig. 6. Color spatial-distribution feature. (a) input images. (b) color spatial variance feature maps. (c) center-weighted, color spatial variance feature maps.

how distinct the salient object in the rectangle is with respect to its surroundings, we can measure the distance between R and R_S using various visual cues such as intensity, color, and texture/texton. In this paper, we use the χ^2 distance between histograms of RGB color: $\chi^2(R, R_S) = \frac{1}{2}\sum \frac{(R^i - R^i_S)^2}{R^i + R^i_S}$. We use histograms because they are a robust global description of appearance. They are insensitive to small changes in size, shape, and viewpoint. Another reason is that the histogram of a rectangle with any location and size can be very quickly computed by means of an integral histogram introduced recently [40]. Fig. 4 (a) shows that the salient object (the girl) is most distinct using the χ^2 histogram distance. We have also tried the intensity histograms and histograms of oriented gradient [41].We found that the former is redundant with the color histogram and the latter is not a good measurement because the texture distribution in a semantic object is usually not coherent.

To handle varying aspect ratios of the object, we use five templates with different aspect ratios $\{0.5, 0.75, 1.0, 1.5, 2.0\}$. We find the most distinct rectangle $R^*(x)$ centered at each pixel x by varying the size and aspect ratio:

$$R^*(x) = \arg\max_{R(x)} \chi^2(R(x), R_S(x)). \quad (12)$$

The size range of the rectangle $R(x)$ is set to $[0.1, 0.7] \times \min(w, h)$, where w, h are image width and height. Then, the center-surround histogram feature $f_h(x, I)$ is defined as a sum of spatially weighted distances:

$$f_h(x, I) \propto \sum_{\{x' | x \in R^*(x')\}} w_{xx'} \chi^2(R^*(x'), R^*_S(x')), \quad (13)$$

where $R^*(x')$ is the rectangle centered at x' and containing the pixel x. The weight $w_{xx'} = \exp(-0.5\sigma^{-2}_{x'}||x - x'||^2)$ is a Gaussian falloff weight with variance $\sigma^2_{x'}$, which is set to one third of the size of $R^*(x')$. Finally, the feature map $f_h(\cdot, I)$ is also normalized to the range $[0, 1]$.

Fig. 4 (b) shows several center-surround feature maps. The salient objects are well located by the center-surround histogram feature. Especially, the last image in Fig. 4 (b) is a difficult case for color or contrast based approaches but the center-surround histogram feature can capture the "object-level" salient region.

To further verify the effectiveness of this feature, we compare the center-surround histogram distance of a randomly selected rectangle, a rectangle centered at the image center, and three user-labeled rectangles in the image. Fig. 5 shows the average distances on the image set $\mathcal{A}$, and this image set is introduced in section 4. It is no surprise that the salient object has a large center-surround histogram distance.

3.1.3 Color spatial-distribution

The center-surround histogram is a regional feature. Is there a global feature related to the salient object? We observe from Fig. 2 that the wider a color is distributed in the image, the less possible a salient object contains

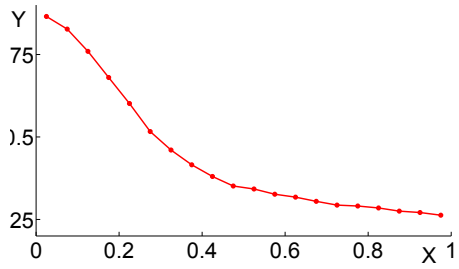


Fig. 7. Color spatial variance (x-coordinate) v.s. average saliency probability (y-coordinate) on the image set $\mathcal{A}$. The saliency probability is computed from the "ground truth" labeling.

this color. The global spatial distribution of a specific color can be used to describe the saliency of an object.

To describe the spatial-distribution of a specific color, the simplest approach is to compute the spatial variance of the color. First, all colors in the image are represented by Gaussian Mixture Models (GMMs) $\{w_c, \mu_c, \Sigma_c\}_{c=1}^{C}$, where $\{w_c, \mu_c, \Sigma_c\}$ is the weight, the mean color and the covariance matrix of the cth component. Each pixel is assigned to a color component with the probability:

$$p(c|I_x) = \frac{w_c \mathcal{N}(I_x|\mu_c, \Sigma_c)}{\sum_c w_c \mathcal{N}(I_x|\mu_c, \Sigma_c)}. \quad (14)$$

Then, the horizontal variance $V_h(c)$ of the spatial position for each color component c is:

$$V_h(c) = \frac{1}{|X|_c} \sum_x p(c|I_x) \cdot |x_h - M_h(c)|^2, \quad (15)$$

$$M_h(c) = \frac{1}{|X|_c} \sum_x p(c|I_x) \cdot x_h, \quad (16)$$

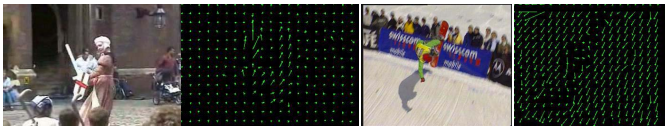
where x_h is the x-coordinate of the pixel x, and $|X|_c = \sum_x p(c|I_x)$. The vertical variance $V_v(c)$ is similarly defined. The spatial variance of a component c is $V(c) = V_h(c) + V_v(c)$. We normalized $\{V(c)\}_c$ to the range $[0, 1]$ ($V(c) \leftarrow (V(c) - \min_c V(c))/(\max_c V(c) - \min_c V(c))$). Finally, the color spatial-distribution feature $f_s(x, I)$ is defined as a weighted sum:

$$f_s(x, I) \propto \sum_c p(c|I_x) \cdot (1 - V(c)). \quad (17)$$

The feature map $f_s(\cdot, I)$ is also normalized to the range $[0, 1]$. Fig. 6 (b) shows color spatial-distribution feature maps of several example images. The salient objects are well covered by this global feature. Note that the spatial variance of the color at the image corners or boundaries may also be small because the image is cropped from the whole scene. To reduce this artifact, a center-weighted, spatial-variance feature is defined as:

$$f_s(x, I) \propto \sum_c p(c|I_x) \cdot (1 - V(c)) \cdot (1 - D(c)), \quad (18)$$

where $D(c) = \sum_x p(c|I_x) d_x$ is a weight which assigns less importance to colors nearby image boundaries and is also normalized to $[0, 1]$, similar to $V(c)$. d_x is the distance from pixel x to the image center. As shown in Fig. 6 (c), center-weighted, color spatial variance shows a better prediction of the saliency of each color.



(a) (b)

Fig. 8. Motion map. (a) the salient object with a large motion. (b) the background with a large motion.

To verify the effectiveness of this global feature, we plot the color spatial-variance versus average saliency probability curve on the image set $\mathcal{A}$, as shown in Fig. 7. Obviously, the smaller a color variance is, the higher the probability the color belongs to the salient object.

3.2 Dynamic Salient Feature

3.2.1 Motion salient features

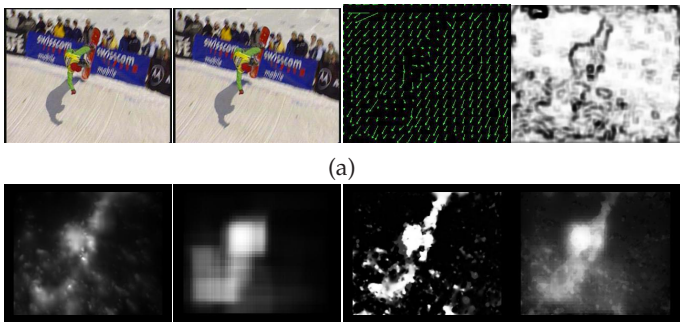
The motion field and the features derived from it are useful to induce visual attention. For example, large motion and motion contract are supposed to induce visual attention in [27] [24], and a constant velocity motion model is assumed for the salient object in [30]. Motion magnitude is a possible cue, but may not be sufficient. For example, in Fig. 8(a), the region with larger motion magnitude includes the salient object. In contrast, the region with smaller motion magnitude includes the salient object in Fig. 8(b). In this paper, we view the motion field as an image and define the local, regional and global salient features from it.

We compute the motion field M using the SIFT flow [35]. It can be observed that the motion fields have some special properties for the salient feature computation. For example, the motion fields from the salient object tend to be consistent because the regions from the salient object are inclined to have a similar motion, and the motion fields in the regions of object boundaries are usually disordered. To measure this consistency, motion variance $V(x, M)$ in a small rectangle surrounding x is computed, and a weight is assigned to each pixel as follows:

$$W(x, M) = \exp(-\epsilon_c ||V(x, M)||^2), \quad (19)$$

where $V(x, M)$ is computed on a 2D motion vector from a window (5×5 in this paper) centered at x and $\epsilon_c = 0.2$. As in the first row of Fig.9, the motion from the surrounding region of pixel x is more cluttered, and the weight of pixel x is smaller.

Compared with the salient features defined for a single image, all the local, regional, and global salient features are defined similarly on weighted 2D motion vectors, including the motion magnitude and the motion direction. In the following, we present the formulation and only highlight the difference from the image.



(a)

(b)

Fig. 9. Motion salient features. From left to right: (a) two adjacent images, the motion field, the motion weight map, (b) the local, regional, global and the combined motion salient features.

Multi-scale contrast of weighted motion field is defined on weighted motion vectors as follows:

$$f_{Mc}(x,M)=\sum_{l=1}^{L}\sum_{x'\in N(x)}W_x^l W_{x'}^l||M^l(x)-M^l(x')||^2, \quad (20)$$

where M^l is the lth-level motion in the pyramid, and W_x^l is the weight at pixel x. We also test the multi-scale contrast on motion magnitude or motion direction. They do not outperform the feature on the 2D motion vector because neighborhood pixels may have the same motion magnitude but different directions, and the salient feature from orientation does not perform well especially when the motion magnitude is small.

Center-surround histogram of weighted motion field captures the statistic difference of motion field in a regional extension. We compute the histogram of motion vectors where horizontal and vertical motion are both normalized and used. The regional salient feature is defined as:

$$f_{Mh}(x,M)\propto\sum_{\{x'|x\in R_M^*(x')\}}w_{xx'}W_{x'}\chi^2(R_M^*(x'),R_{MS}^*(x')), \quad (21)$$

where R_M^* has the largest center-surround histogram distance on motion vectors, $w_{xx'}$ is the weight for the spatial distance, and $W_{x'}$ is the weight of pixel x'.

Spatial-distribution of weighted motion field captures the global distribution of the motion field in an image. There are usually several different prominent motions in one frame, such as the motions from the background, object or disturbs. Similar to the spatial distribution of color, the wider a motion is distributed in the image, the less possibly a salient object corresponds to this motion. To get the spatial-distribution, these motion vectors, in which each vector is weighted by W_x, are first clustered into several GMMs. The spatial variance $V_M(m)$ of each Gaussian component m is computed similar to the static feature, and the final spatial distribution feature is defined as:

$$f_{Ms}(x,M)\propto\sum_m W_x p(m|M_x)\cdot(1-V_M(m)). \quad (22)$$

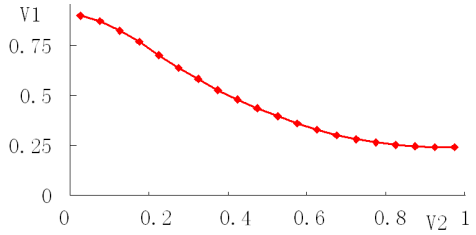


Fig. 10. Histogram distance of appearance features (x-coordinate) vs. average saliency ratio (y-coordinate).

(a) (b) (c) (d) (e)

Fig. 11. Appearance coherent feature. (a) and (c) are an image pair, (b) and (d) are the corresponding static salient features, and (e) is the appearance coherent feature.

Usually, there are fewer components for motion than for color because there are not many independent moving regions in one image. We use $3\sim5$ motion components in this paper.

3.2.2 *Appearance coherent feature*

It is observed that salient objects from two consecutive frames probably have similar appearance features. The observation is verified on our labeled image pairs. We first compute the color histogram $R_t(x)$ in the labeled rectangle on the current image I_t, and the bth bin of $R_t(x)$ is computed as: $R_t(x)^b=\sum_{x'\in R_t}f(x',I_t)\delta(I_{x'}=b)$, where $f(x',I_t)$ is set to 1 if x' is in the labeled rectangle and 0 otherwise. Second, we randomly select one rectangle with the same size in the previous image I_{t-1}, and compute the color histogram of $R_{t-1}(x')$ similarly. Third, we compute the saliency ratio V_1 and the χ^2 distance between the color histograms $R_t(x)$ and $R_{t-1}(x')$ as two variables:

$$V_1=\frac{\sum_{x'}f(x',I)\delta(x'\in R_t)}{\sum_{x'}\delta(x'\in R_t)}\cdot\frac{\sum_{x'}f(x',I_{t-1})\delta(x'\in R_{t-1})}{\sum_{x'}\delta(x'\in R_{t-1})},$$
$$V_2=\chi^2(R_t(x),R_{t-1}(x')), \quad (23)$$

where $f(x,I)$ and $f(x',I_{t-1})$ come from the labeled ground truths. We then create a statistic of the relationship between the two variables V_1 and V_2, as shown in Fig. 10.

To integrate the appearance coherence into the energy defined on a 2D graph, we try to penalize the pixels that are identified to be in the salient object by static salient features but with a big color histogram difference. First, we compute the weighted color histogram $R_t(x)$ from a $N\times N$ patch surrounding pixel x, and the bth bin of the color histogram $R_t(x)$ is computed as: $R_t(x)^b=\sum_{x'\in R_t}f(x',I)\delta(I_{x'}=b)$, where $f(x',I)$ is the static salient feature defined on a single image. Second, we search the patch $R_{t-1}(x^*)$ in image I_{t-1} to satisfy:

$x^* = \arg\max_{x'} \chi^2(R_t(x), R_{t-1}(x'))$, where $x' \in N(x)$ and $N(x)$ are the set of the neighboring pixels of x, and $R_{t-1}(x')$ is computed similar to $R_t(x)$. Finally, the appearance coherent feature is computed as:

$$f(x, I_t, I_{t-1}) \propto \frac{f(x, I_t) + f(x^*, I_{t-1})}{2} \exp(-\chi^2(R_t(x), R_{t-1}(x^*))), \quad (24)$$

where $f(x, I_t)$ and $f(x^*, I_{t-1})$ are the static salient features from I_t and I_{t-1}. Fig. 11 gives an example of the appearance coherent feature.

4 Evaluation

4.1 Data Set

4.1.1 Image data set

We have collected a very large image database with 130,099 high quality images from a variety of sources, mostly from image forums and image search engines. Then we manually selected 60,000+ images, each of which contains a salient object or a distinctive foreground object. To test the performance, we further selected 20,840 images that contain a clear, unambiguous object of interest, which is helpful for building the ground truth. In the selection process, we excluded any image containing a very large salient object so that the performance of detection can be more accurately evaluated.

Fig. 2 gives some example images, and each image contains an unambiguous salient object. These salient objects differ in category, color, shape, size etc. In other words, there is no more prior knowledge or constraint on these objects except that they are the most salient. This image database is different from the UIUC Cars dataset, or the PASCAL VOC 2006 dataset, where images containing a specific category of objects are collected together. As clarified in the above section, we do not judge whether an object exists or discriminates from multiple objects. Specifically, we aim to locate the salient object, with the assumption that one salient object exists in the given image.

4.1.2 Sequential image data set

We collected a video database with 2,000+ video segments from a variety of sources, e.g., video sharing web sites. Further, we selected 100 video segments that include salient object sequences, such as racing car, long jump, kids sequences and so on. Example images from these video segments are shown in Fig. 14. Each video segment contains about $100 \sim 500$ frames with the same salient objects. We also label the ground truth by hand for parameter learning and result evaluation, and 30,000+ image pairs are collected for labeling. One trait of these image pairs is that the image quality is not as good as the image quality from the above image data set, because all images are taken from video segments on web sites. Another trait is that the salient objects are much smaller and the average of the ratios between the sizes of salient objects and images is about 0.1 , which results in that the salient object detection tasks in those video segments are very challenging.

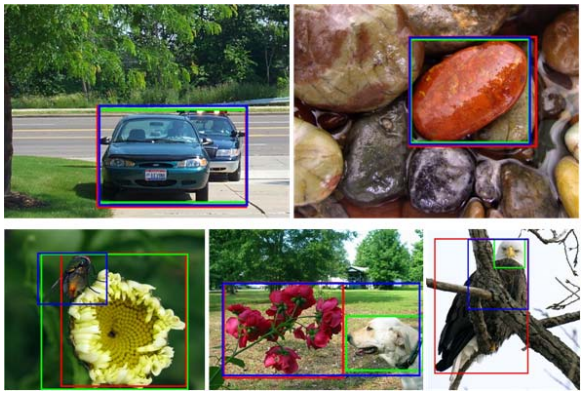


Fig. 12. Labeled images from 3 users. Top: two consistent labeling examples. Bottom: three inconsistent labeling examples.

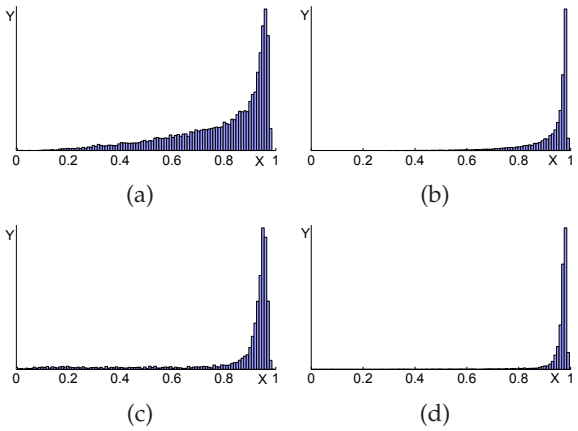
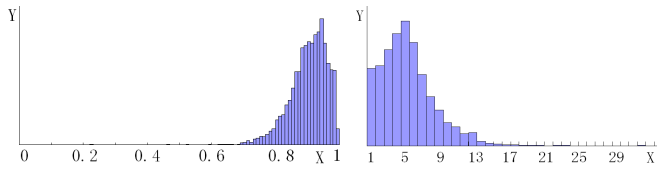


Fig. 13. Labeling consistency for image data set. (a) (b) $C_{0.9}$ (agreed by all 3 users) and $C_{0.5}$ on image set $\mathcal{A}$. (c) (d) $C_{0.9}$ (agreed by at least 8 of 9 users) and $C_{0.5}$ on image set $\mathcal{B}$.

4.2 Ground Truth Construction

For labeling the ground truth, we ask the user to draw a bounding rectangle to specify a salient object. Our detection algorithm also outputs a rectangle around the salient object. As addressed in [43], one advantage is that it is much easier to provide ground truth annotation for bounding boxes than e.g. for pixel-wise segmentations. At the same time, the rectangle representation of the salient object satisfies many applications, such as adaptive image display on small devices and image collage. We still represent the salient object piecewise as A_t in the problem formulation, and we will transform the final binary result to a bounding rectangle for further evaluation and applications where this strategy can avoid cutting off the spindly edge of the salient object.

Fig. 14. Sample images from experimental video segments, and our detection results on these images.



(a) (b)

Fig. 15. Labeling continuity for video data set. (a) is the histogram of C, and (b) is the histogram of maximal boundary distance. both are on two adjacent labeled rectangles R_{t-1} and R_t.

4.2.1 Labeling consistency in image data set.

People may have different ideas about what a salient object in an image is. To address the problem of "what is the most likely salient object in a given image", we take a voting strategy by labeling a "ground truth" salient object in the image by multiple users. In this paper, we focus on the case of a single salient object in an image. For each image to be labeled, we ask the user to draw a rectangle that encloses the most salient object in the image according to his/her own understanding. The rectangles labeled by different users usually are not the same. To reduce the labeling inconsistency, we vote a "ground truth" labeling from the rectangles drawn by multiple users.

In the first stage, we asked three users to label all 20,840 images individually. On average, each user took 10∼20 seconds to draw a rectangle on an image. The whole process took about three weeks. Then, for each labeled image, we compute a saliency probability map $G = \{g_x | g_x \in [0, 1]\}$ of the salient object using the three user labeled rectangles:

$$g_x = \frac{1}{M} \sum_{m=1}^{M} a_x^m, \tag{25}$$

where M is the number of users and $A^m = \{a_x^m\}$ is the binary mask labeled by the mth user. Fig. 12 shows two highly consistent examples and three inconsistent examples. The inconsistent labeling is due to multiple disjointed foreground objects for the first two examples at the bottom row. The last example at the bottom row shows that an object has hierarchical parts that are of interest. We call this image set $\mathcal{A}$. In this paper, we focus on consistent labeling of a single salient object for each image.

To measure the labeling consistency, we compute statistics C_t for each image:

$$C_t = \frac{\sum_{x \in \{g_x > t\}} g_x}{\sum_x g_x}. \tag{26}$$

C_t is the percentage of pixels whose saliency probabilities are above a given threshold t. For example, $C_{0.5}$ is the percentage of the pixels agreed on by at least half of the users. $C_{0.9} \approx 1$ means that the image is consistently labeled by all the users. Fig. 13 (a) and 13 (b) show the histograms of $C_{0.9}$ and $C_{0.5}$ on the image set $\mathcal{A}$. As we can see, the labeled results are quite consistent, e.g., 92% of the labeling results are consistent between at least two users (Fig. 13 (b)) and 63% of the labeling results are highly consistent among all three users (Fig. 13 (a)).

In the second stage, we randomly selected 5,000 highly consistent images (i.e., $C_{0.9} > 0.8$) from the image set $\mathcal{A}$. Then, we asked nine different users to label the salient object rectangle. Fig. 13 (c) and 13 (d) show the histograms of $C_{0.9}$ and $C_{0.5}$ on these images. Compared with the image set $\mathcal{A}$, this set of images has less ambiguity of what the salient object is. We call these images as image set $\mathcal{B}$.

After the above two-stage labeling process, the salient object in our image database is defined based on the "majority agreement" of multiple users and represented as a saliency probability map. The whole labeled image database are publicly available[1].

4.2.2 Labeling continuity in sequential image data set.

For image pairs from video segments, people may have less disputation about what the salient object is because the motion helps to address the salient object. We ask only one user to label these sequential images, and it takes about two weeks to label all image pairs. In most cases, the movement of the salient object is smooth, which means the labeled rectangles from two adjacent frames is also continuous. To describe the labeling continuity, we compute the statistic: $C = \frac{Region\ area(R_{t-1} \cap R_t)}{Region\ area(R_{t-1} \cup R_t)}$, where R_{t-1} and R_t are two labeled rectangles for two adjacent frames. Fig. 15 (a) shows the histogram about C on these image pairs. We also get statistics on the maximal boundary distance to describe the labeling continuity for sequential images, and the histogram is shown in Fig. 15 (b). We find that the maximal boundary distance is less than 10 pixels for 95% of image pairs, and this is also used as a reference when we define the appearance coherent features.

1. http://research.microsoft.com/jiansun/SalientObject/salient_object.htm

4.3 Evaluation Criteria

With the labeled probability map G, for any detected salient object mask A, we define region-based and boundary-based measurements. We use the precision, recall, and F-measure for region-based measurement. Precision/Recall is the ratio of a correctly detected salient region to the detected / "ground truth" salient region:

$$Precision = \sum_x g_x a_x / \sum_x a_x, Recall = \sum_x g_x a_x / \sum_x g_x. \quad (27)$$

F-measure is the weighted harmonic mean of precision and recall, with a non-negative α :

$$F_\alpha = \frac{(1+\alpha) \times Precision \times Recall}{\alpha \times Precision + Recall}. \quad (28)$$

We set $\alpha = 0.5$ following [44]. The F-measure is an overall performance measurement.

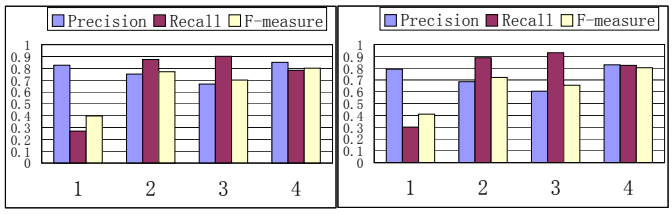
For the boundary-based measurement, we use boundary displacement error (BDE) [45], which measures the average displacement error of the corresponding boundaries of two rectangles. The displacement is averaged over the different users.

4.4 Implementation of CRF Learning and Inference

For image data set, we randomly select 2,000 images from image set $\mathcal{A}$ and 1,000 images from image set $\mathcal{B}$ to construct a training set, which are excluded from the testing phase. For sequential image data set, we randomly select 20 video segments with 5,000+ image pairs to construct a training set, and use others for testing. We do many different splits in terms of a training/test dataset, and we find that the different splits almost do not effect the evaluation results. The key factor is the amount of training data, and the parameter learning algorithm can converge well when the amount of training data is more than 2,000.

Because the ground truth of salient objects are labeled by rectangles, this strategy lacks the precise alignment between object boundaries and labeled rectangles. Instead of learning the parameter of the pairwise feature [46], we normalize the sum of λ_k by experience, eg. $\sum_k \lambda_k = 1$. Furthermore, we observe that the pixels from the boundaries of labeled rectangles are less believable, because the surrounding rectangle may label some pixels near the boundaries as the salient object by mistake. To reduce this effect, we use a Gaussian function to give the weight of pixels when we compute $\Delta\lambda_k$ in (9). This strategy helps to speed up the convergence of the learning algorithm.

We use the tree-reweighted belief propagation to infer the labeling because it is used for CRF learning. We find that there are small differences between the learned parameters if we use different algorithms to compute the marginal distribution $p(a_x^n | I^n; \vec{\lambda})$. To output a rectangle for the evaluation, we exhaustively search for a smallest rectangle containing at least 95% salient pixels in the binary label map produced by the CRF model.



(a) image set $\mathcal{A}$ (b) image set $\mathcal{B}$

Fig. 16. Evaluation of salient object features. 1. multi-scale contrast. 2. center-surround histogram. 3. color spatial distribution. 4. combination of all features.

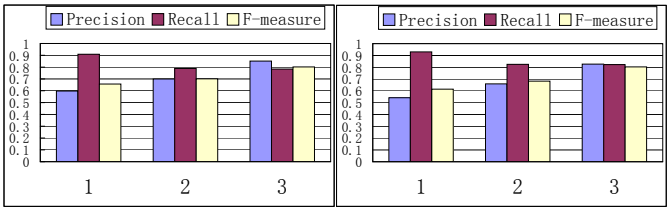
Fig. 17. Examples of salient features. From left to right: input image, multi-scale contrast, center-surround histogram, color spatial distribution, and binary salient mask by CRF.

4.5 Salient Object Detection from a Single Image

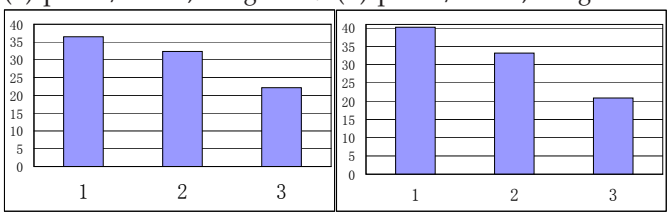
4.5.1 *Effectiveness of features and CRF learning*

To evaluate the effectiveness of each salient object feature, we trained four CRFs: three CRFs with individual features and one CRF with all three features. Fig. 16 shows the precision, recall, and F-measure of these CRFs on the image sets $\mathcal{A}$ and $\mathcal{B}$. As can be seen, the multi-scale contrast feature has a high precision but a very low recall. The reason is that the inner homogenous region of a salient object has low contrast. The center-surround histogram has the best overall performance (on F-measure) among all individual features. This regional feature is able to detect the whole salient object, although the background region may contain some errors. The color spatial-distribution has slightly lower precision but has the highest recall. Later, we will discuss that for attention detection, recall rate is not as important as precision. It demonstrates the strength and weakness of the global feature. After CRF learning, the CRF with all three features produces the best result, as shown in the last bars in Fig. 16. The best linear weights we learnt are: $\vec{\lambda} = \{0.24, 0.54, 0.22\}$.

Fig. 17 shows the feature maps and labeling results of several examples. Each feature has its own strengths and limitations. By combining all features with the pairwise feature, the CRF successfully locates the most salient object.

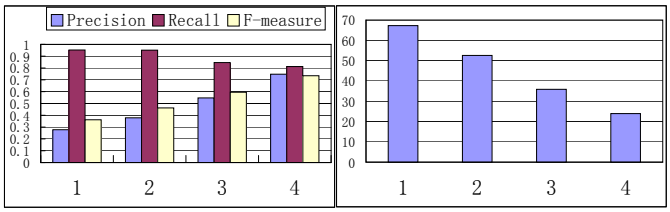


(a) preci./recall, image set $\mathcal{A}$ (b) preci./recall, image set $\mathcal{B}$



(c) BDE, image set $\mathcal{A}$ (d) BDE, image set $\mathcal{B}$

Fig. 18. Comparison of different algorithms. (a-b) and (c-d) are region-based (precision, recall, and F-measure) and boundary-based (BDE - boundary displacement error) evaluations. 1. FG. 2. SM. 3. our approach.



(a) preci./recall (b) BDE

Fig. 19. Comparison on a small object (object/image ratio $\in [0, 0.25]$) dataset from image set $\mathcal{A}$. 1. a rectangle centered at the image center and with 0.6 object/image ratio. 2. FG. 3. SM. 4. our approach.

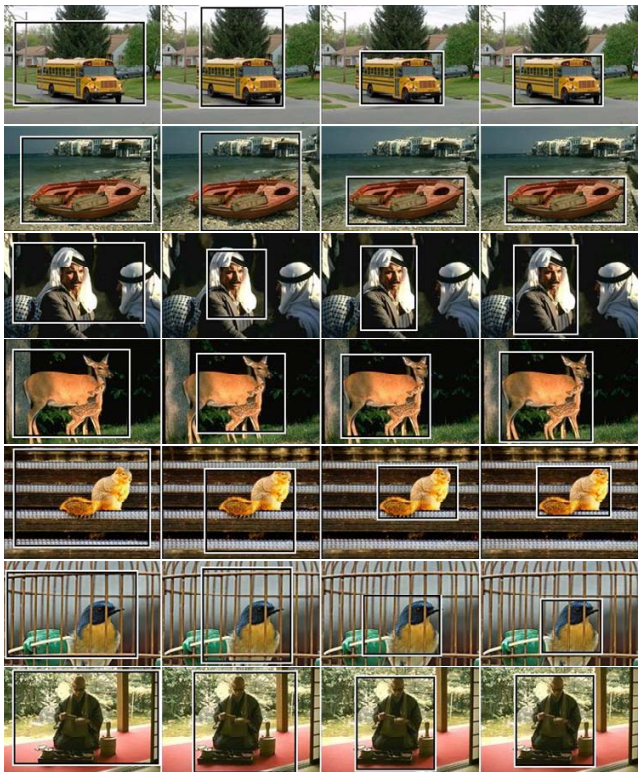


Fig. 20. Comparison of different algorithms. From left to right: FG, SM, our approach, and ground-truth.

4.5.2 Comparison with other approaches

We compare our algorithm with two leading approaches. One is the contrast and fuzzy growing based method [39], which we call "FG". This approach directly outputs a rectangle. Another approach is based on the salient model presented in [13](We use a matlab implementation from http://www.saliencytoolbox.net), and we call it "SM". Because the output of Itti's salient model is a saliency map, we convert the saliency map to a rectangle containing 95% of the fixation points, which are determined by the winner-take-all algorithm [13]. We also resolve the rectangles directly through complete searching by maximizing $\sum_{x \in R}(1 - F(x)) + \sum_{x \notin R} F(x)$, where R is the resolve rectangle and $F(x) \in [0, 1]$ is the normalized saliency map. This method can be applied on our saliency map and Itti's saliency map, but the results do not outperform the corresponding results using the current method.

Fig. 18 shows the evaluation results of three algorithms on both image sets $\mathcal{A}$ and $\mathcal{B}$. On image set $\mathcal{A}$, our approach reduces by 42% and 34% the overall error rates on F-measure, and 39% and 31% boundary displacement errors (BDEs), compared with FG and SM. Similarly, 49% and 38% overall error rates on F-measure, and 48% and 37% BDEs are reduced on the image set $\mathcal{B}$.

Notice that as shown in Fig. 16 and 18, the individual features (center-surround histogram and color spatial-distribution), FG, and SM all have higher recall rates than our final approach. In fact, recall rate is not much of a useful measure in attention detection. For example, a 100% recall rate can be achieved by simply selecting the whole image. So an algorithm trying to achieve a high recall rate tends to select as large an salient region as possible, sacrificing the precision rate. The key objective of salient object detection should be to locate the position of a salient object as accurately as possible, i.e. with high precision. However, for images with a large salient object, a high precision is also not too difficult to achieve. Again, for example, for an image with a salient object occupying 80% of the image area, just selecting the whole image as the salient area will give 80% precision with 100% recall rate. So the real challenge for salient object detection is to achieve high precision on small salient objects. To construct such a challenging data set, we select a small object subset with object/image ratio in the range $[0, 0.25]$ from the image set $\mathcal{A}$. The results on this small object dataset are shown in Fig. 19, where we also show the performance of a rectangle fixed at the image center with 0.6 object/image ratio. Notice that both this center rectangle and FG achieve high recall rate but with very low precision and large BDE. Our method is significantly better than FG and SM in both

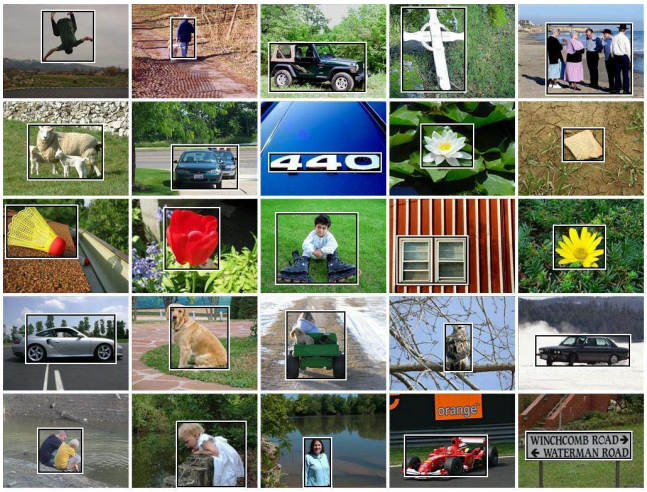


Fig. 21. Our detection result on the images in Fig. 2.

precision (97% and 37% improvement) and BDE (55% and 33% reduction). Fig. 20 shows several examples with ground truth rectangles from one user for a qualitative comparison. We can see that FG and SM approaches tend to produce a larger attention rectangle and our approach is much more precise.

Fig. 21 shows our detection results on the images in Fig. 2. Our results are also publicly available with the whole labeled database.

4.6 Salient Object Detection from Sequential Images

4.6.1 Effectiveness of salient features

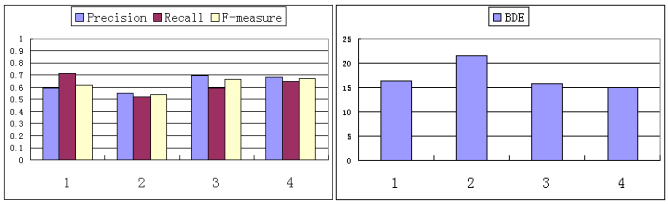
To evaluate the effectiveness of the static and dynamic salient features for sequential images, we set up a "baseline method" model in the CRF framework with the following energy:

$$E(A_t|I_{t-1}, I_t) = \sum_x \sum_{k=1}^{K} \lambda_k F_k(a_x, \cdot) + \sum_{x,x'} S(a_x, a_{x'}, I_t), \tag{29}$$

where $F_k(a_x, \cdot)$ indicates different salient features, and $S(a_x, a_{x'}, I_t)$ is the pairwise feature. We define $F_k(a_x, \cdot)$ from (29) with different features, and train four CRFs as follows:

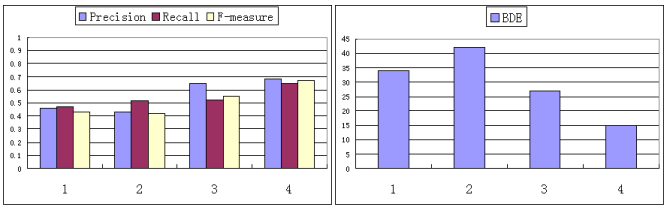
- **C1:** The static salient features from the current image I_t are used: $F_k(a_x, \cdot) = F_k(a_x, I_t)$.
- **C2:** The motion salient features from the motion field M_t are used: $F_k(a_x, \cdot) = F_k(a_x, M_t)$.
- **C3:** The static and motion salient features are both used to detect a salient object, and $F_k(a_x, \cdot)$ is the combination of $F_k(a_x, I_t)$ and $F_l(a_x, M_t)$.
- **C4:** All the static and dynamic salient features are used, and $F_k(a_x, \cdot)$ is the combination of $F_k(a_x, I_t)$, $F_l(a_x, M_t)$ and $F_m(a_x, I_{t-1}, I_t)$. This is the proposed approach for salient object detection.

Fig. 22 shows the precision, recall, F-measure and BDE of these CRFs on 30,000+ image pairs. The salient objects in these image pairs are very small compared with the



(a) precion/recall (b) BDE

Fig. 22. Evaluation of different salient features for sequential images. In (a) and (b), 1)-4) corresponds to C1-4 respectively where different salient features are trained within the CRF framework.



(a) precion/recall (b) BDE

Fig. 23. Comparison with different approaches for sequential images. In (a) and (b), 1)-3) corresponds to D1-3 respectively, 4) is our approach C4.

image database, and that is why the performance of C1 is not as good as the trained CRF in our image database. The approach C3, combining salient features on color and motion, improves $8 \sim 9\%$ on F-measure compared with the approach C1 with only the salient features on color. C3 also improves 23% on F-measure compared with C2 with only the salient features on motion. We can see that C1 outperforms C2 on F-measure. With the appearance coherent features, the proposed approach C4 improves 9% on recall with a little sacrifice on precision, and improves 1% on the overall criteria F-measure.

4.6.2 Comparison with other approaches

A general CRF model is defined in (29) where different static and dynamic features can be included to learn a detector. We compare our approach with the following approaches:

- **D1:** We use Itti.'s salient model to compute the static saliency map following [47] [24] [27], and the multiple-scale motion contrast from [24] to compute the dynamic saliency map. We also test the motion saliency from [47], where the difference between the pixel's motion and the global motion from the whole image is computed as the motion saliency, and experiments indicate that it does not outperform the multiple-scale motion contrast.
- **D2:** Different from D1, we use the saliency map from the temporal surprise in [11] as the dynamic salient feature. However, the surprise computation using all features is extremely costly on a large number of image sequences. A reasonable simplification is to use only four combined saliency

(a) (b)

Fig. 24. Multiple salient object detection. (a) Two birds are detected at the same time. (b) The toy car is detected first, and using the updated feature maps, the boy is detected second.

Fig. 25. Failure cases. From left to right: FG, SM, our approach, and ground-truth.

maps to compute the temporal surprise. We use the publicly available Bayesian Surprise Matlab toolkit (http://sourceforge.net/projects/surprise-mltk) for the implementation.

- **D3:** Other related work includes the tracking algorithm with a hand-initialization in the first frame, and we report the results of the typical mean-shift tracking algorithm [48].

Fig. 23 shows the comparison of our approach with D1-3. Our approach improves 52% on F-measure and reduces 54% on BDEs compared with D1. We also evaluate the results with only the motion saliency in (29), and find that the multi-scale motion contrast has a very low recall. This is the main reason that motion saliency is not well leveraged in D1. Our approach also improves 59% on F-measure and reduces 64% on BDEs compared with D2. The difference between D1 and D2 is the motion saliency, and we find that video surprise with the goal of eye movements cannot help to locate the small salient object well, and further the video surprise does not strengthen the static saliency much because it is computed based on the static saliency. We find that the collected image sequences are also very challenging for the mean-shift tracking algorithm, because of the following traits: large motion of object and camera, object rotation and appearance change, illumination change, and so on. Our approach improves 20% on F-measure and reduces 45% on BDEs compared with D3. The results imply that the salient features can help visual object tracking for those challenging videos.

5 Discussion

In this section, we discuss the connection and clarify the difference between our approach for salient object detection and other related work.

5.1 Salient object vs. visual saliency

A visual saliency map is computed from multi-scale image features in Itti's model [13] [12], which is one of the most representative works on computational modeling of visual attention. Itti's model and those similar to it are based on the biologically plausible computational models of attention, with a particular emphasis on bottom-up control of attentional deployment. They state as a goal the determination of fixation and eye movements over an image. We summarize the difference between salient object detection and visual saliency computation with Itti's model in the following.

First, a salient object is essentially one important aspect of visual attention, and the goal is to locate the salient object in an image or sequential images, to help on displaying images on a small device, or browsing image collection. It is different from Itti's model that has as a goal the determination of fixation and eye movements over an image. The recent study [42] also analyzes their connection and indicates that Itti's visual saliency model is closely related to the interesting object detection. Second, we propose a different solution to the problem. We adopt a binary mask to indicate a salient object using the salient feature maps which are combined with learned parameters. These feature maps are different from the visual saliency maps or the conspicuous maps in Itti's model that are based on biological theory.

5.2 Salient object detection vs. figure-ground segregation

The figure-ground segregation task is similar to salient object detection as both aim to find the objects, but they are essentially different. The main difference is that our approach detects a salient object automatically, without any prior knowledge about its category, its shape or size and that the conventional figure-ground segregation algorithms require the supposition of the category of objects [19] [20] [21] or user interactions [22] [23]. On the other hand, the visual features adopted for the detection differ greatly. For salient object detection, we propose generic salient features without discrimination of object categories. For figure-ground segregation of an object with a given category, the specific features, for example for cows, may be defined specifically and these features cannot be adopted for other categories. Due to the above differences, the figure-ground segregation algorithm is not comparable to our approach.

6 Conclusion

In this paper, we have presented a supervised approach for salient object detection, which is formulated as a

binary labeling problem using a set of local, regional, and global salient object features. A CRF model was learned and evaluated on a large image database containing 20,000+ well-labeled images by multiple users. We also extend this supervised approach to detect a salient object sequence from sequential images, where dynamic salient features are included to help detect the salient object.

There are several possible remaining issues for further investigation. We plan to experiment with non-rectangular shapes for salient objects, and a non-linear combination of features. In particular, we are extending our single salient object detection framework to detect any number of salient objects including no salient object at all. Fig. 24 shows two initial results. In Fig. 24 (a), our current CRF approach can directly output two disjointed connected components so that we can easily detect them simultaneously. In Fig. 24 (b), we use the inhibition-of-return strategy [13] to detect the salient objects one-by-one. Finally, Fig. 25 shows two failure cases, which demonstrate one of the challenges in the salient object detection — hierarchical salient object detection.

Acknowledgments Tie Liu and Zejian Yuan were supported by a grant from National Natural Science Foundation of China (No.90820017). Zejian Yuan and Nanning Zheng were supported by grants from National basic research program of China (No.2007CB311005) and the National High-Tech Research and Development Plan of China (No.2006AA01Z192).

REFERENCES

[1] A. Santella, M. Agrawala, D. Decarlo, D. Salesin, and M. Cohen, "Gaze-based interaction for semi-automatic photo cropping," in *CHI*, 2006, pp. 771–780.

[2] L. Chen, X. Xie, X. Fan, W. Ma, H. Shang, and H. Zhou, "A visual attention mode for adapting images on small displays," Microsoft Research, Redmond, WA, Tech. Rep., 2002.

[3] L. Itti, "Models of bottom-up and top-down visual attention," Ph.D. dissertation, California Institute of Technology Pasadena, 2000.

[4] C. Rother, L. Bordeaux, Y. Hamadi, and A. Blake, "Autocollage," in *SIGGRAPH*, 2006, pp. 847–852.

[5] V. Navalpakkam and L. Itti, "An integrated model of top-down and bottom-up attention for optimizing detection speed," in *CVPR*, 2006, pp. 2049–2056.

[6] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?" in *CVPR*, 2004, pp. 37–44.

[7] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch., "Attentional selection for object recognition - a gentle way," in *Biol. Motivated Comp. Vision*, 2002.

[8] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *NIPS*, 2005, pp. 155–162.

[9] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *NIPS*, 2006, pp. 545–552.

[10] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," in *NIPS*, 2005, pp. 547–554.

[11] ——, "A principled approach to detecting surprising events in video." in *CVPR*, 2005, pp. 631–637.

[12] L. Itti and C. Koch, "Computational modelling of visual attention," *Neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

[13] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. on PAMI*, vol. 20, no. 11, pp. 1254–1259, 1998.

[14] C. Koch and S. Ullman, "Shifts in selection in visual attention: Toward the underlying neural circuitry," *Human Neurobiology*, vol. 4, no. 4, pp. 219–227, 1985.

[15] O.L.Meur, O.L.Callet, D.Barba, and D.Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans. on PAMI*, vol. 28, no. 5, pp. 802–817, 2006.

[16] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. H. Lai, N. Davis, and F. Nuflo, "Modelling visual attention via selective tuning," *Artificial Intelligence*, vol. 78, no. 1-2, pp. 507–545, 1995.

[17] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[18] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *ICCV*, 2009.

[19] A. Levin and Y. Weiss, "Learning to combine bottom-up and top-down segmentation," in *ECCV*, 2006, pp. 581–594.

[20] E. Borenstein, E. Sharon, and S. Ullman, "Combining top-down and bottom-up segmentation," in *Computer Vision and Pattern Recognition Workshop*, 2004.

[21] B. Leibe, K. Mikolajczyk, and B. Schiele, "Segmentation based multi-cue integration for object detection," in *BMVC*, 2006.

[22] Y. Boykov and M. P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images," in *ICCV*, 2001, pp. 105–112.

[23] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *SIGGRAPH*, 2004, pp. 309–314.

[24] R. Carmi and L. Itti, "Visual causes versus correlates of attentional selection in dynamic scenes," *Vision Research*, vol. 46, no. 26, 2006.

[25] Y. Ma and H. Zhang, "A model of motion attention for video skimming," in *ICIP*, 2002, pp. 129–132.

[26] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *ACM Multimedia*, 2006, pp. 815–824.

[27] A. Bur, P.Wurtz, R.M.Miiri, and H.Hugli, "Dynamic visual attention: competitive versus motion priority scheme," in *ICVS*, 2007.

[28] S. Drouin, P. Hbert, and M. Parizeau, "Incremental discovery of object parts in video sequences," in *ICCV*, vol. 2, 2005, pp. 1754–1761.

[29] N. Jojic, J. Winn, and L. Zitnick, "Escaping local minima through hierarchical model selection: Automatic object discovery, segmentation, and tracking in video," in *CVPR*, vol. 1, 2006, pp. 117–124.

[30] D. Liu and T. Chen, "A topic-motion model for unsupervised video object discovery," in *CVPR*, 2007.

[31] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." in *ICML*, 2001, pp. 282–289.

[32] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr., "Interactive image segmentation using an adaptive GMMRF model." in *ECCV*, 2004, pp. 428–441.

[33] Y. Li, J. Sun, and H.-Y. Shum, "Video object cut and paste," in *SIGGRAPH*, 2007, pp. 595–600.

[34] X. Bai, J. Wang, D. Simons, and G. Sapiro, "Video snapcut: Robust video object cutout using localized classifiers," in *SIGGRAPH*, 2009.

[35] C. Liu, J. Yuen, A. B. Torralba, J. Sivic, and W. T. Freeman, "Sift flow: Dense correspondence across different scenes," in *ECCV (3)*, 2008, pp. 28–42.

[36] X. Ren, C. Fowlkes, and J. Malik, "Cue integration for figure/ground labeling," in *NIPS*, 2005, pp. 1121–1128.

[37] V. Kolmogorov, "Convergent tree-reweighted message passing for energy minimization," *IEEE Trans. on PAMI*, vol. 28, no. 10, pp. 1568–1583, 2006.

[38] F. Liu and M. Gleicher, "Region enhanced scale-invariant saliency detection," in *IEEE ICME*, 2006, pp. 1477–1480.

[39] Y.-F. Ma and H.-J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *Proceedings of ICMM*, 2003, pp. 374–381.

[40] F. Porkili, "Integral histogram: A fast way to extract histograms in cartesian spaces." in *CVPR*, 2005, pp. 829– 836.

[41] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection." in *CVPR*, 2005, pp. 886– 893.

[42] L. Elazary and L. Itti, "Interesting objects are visually salient," *Journal of Vision*, vol. 8, pp. 1–15, 2008.

[43] C. H. Lampert, M. B. Blaschko, and T. Hofmann, "Beyond slidingwindows: Object localization by efficient subwindow search," in *CVPR*, 2008, pp. 1–8.

[44] D. R. Martin, C. C. Fowlkes, and J. Malik., "Learning to detect natural image boundaries using local brightness, color, and texture cues." *IEEE Trans. on PAMI*, vol. 26, no. 5, pp. 530–549, 2004.
[45] J. Freixenet, X. Munoz, D. Raba, J. Marti, and X. Cufi, "Yet another survey on image segmentation: Region and boundary information integration." in *ECCV*, 2002, pp. 408–422.
[46] J. Shotton, J.Winn, C.Rother, and A.Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *ECCV*, 2006, pp. 1–15.
[47] F. Liu and M. Gleicher, "Video retargeting: automating pan and scan," in *ACM Multimedia*, 2006, pp. 241–250.
[48] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Analysis Machine Intell.*, vol. 24, pp. 603–619, 2002.