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 (IAIS)
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 paper 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

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

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
Permission to make digital/hard copy of all or part of this material without fee for personal or classroom use provided that the copies are not made or distributed for profit or commercial advantage, the ACM copyright/server notice, the title of the publication, and its date appear, and notice is given that copying is by permission of the ACM, Inc. To copy otherwise, to republish, to post on servers, or to redistribute to lists requires prior specific permission and/or a fee.
© 2010 ACM 0000-0000/2010/0000-0001 $5.00

character hidden in the audience, these effects not only keep our interest alive, 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 like 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 which 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 5-10 years, the computational power enabled implementations of computational attention system that are useful in practical applications, causing an increasing in-

terest 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 paper, we provide a survey of computational visual attention systems and their applications. The 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 neuro-scientific background knowledge about human visual attention. For psychologists and neuro-biologists, 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 which 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 pre-processing method might be called attentional, because it focuses subsequent processing to parts of the data which 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 paper, we focus on approaches which are motivated by human visual attention (see sec. 2.2.1 for a definition).

The structure of the paper is as follows. In section 2, we introduce the concepts of human visual attention and present the psychological theories and models which 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 paper.

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 sec. 2.1. After that, section 2.2 introduces the concepts of visual attention. Finally, we present in sec. 2.3 the most important psychological theories and models of visual attention which form the basis for most current computational systems.

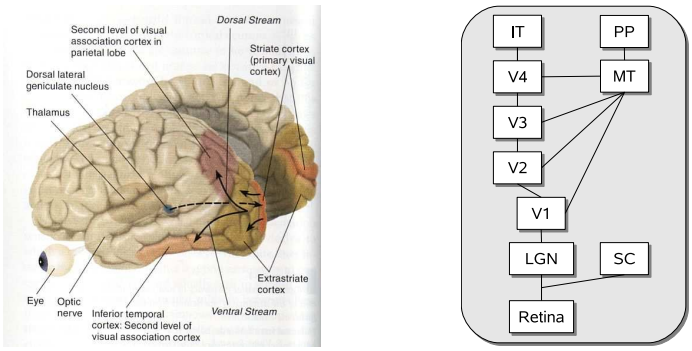


Fig. 1. Left: Visual areas and pathways in the human brain (Fig. from http://philosophy.hku.hk/courses/cogsci/ncc.php). Right: some of the known connections of visual areas in the cortex (Fig. adapted from [Palmer 1999]).

2.1 The Human Visual System

Here, we start with providing a very rough overview of the human visual system (cf. Fig. 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 *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 – V4, infero temporal cortex (IT), the middle temporal area (MT or V5) and the posterior parietal cortex (PP). Although there 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 decades 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 *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, e.g. [Pashler 1997; Styles 1997; 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, that means the observer is "blind" for this change.

The reason why people are nevertheless effective in every-day 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*. A definition is given for example 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.

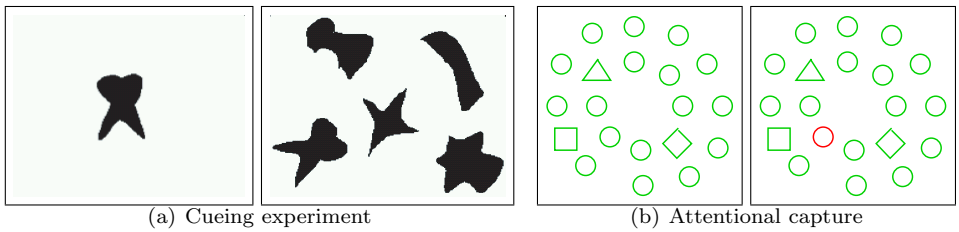
2.2.3 *The unit of attention.* During the last decades, there has been a long debate about the units of attention, that means about the target our attentional focus is directed to. Do we attend to *spatial locations*, to *features*, or to *objects*?

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 only a single unit of attention. Humans are able to attend simultaneously to multiple regions of interest, usually between 4 and 5 regions. This 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. Beside *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 pro-



(a) Cueing experiment (b) Attentional capture

Fig. 2. (a) Cueing experiment: a cue (left) is presented for 200 ms. Then, human subjects have to search for the cued shape in a search array (right) (Fig. 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 65 ms (885 vs 950 ms) [Theeuwes 2004]. That means, the color pop-out "captures" the attention independent of the task (Fig. adapted from [Theeuwes 2004]).

cess. 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" which 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] (cf. Fig. 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 slow down 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 neuro-physiological 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* (cf. Fig. 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 which 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.

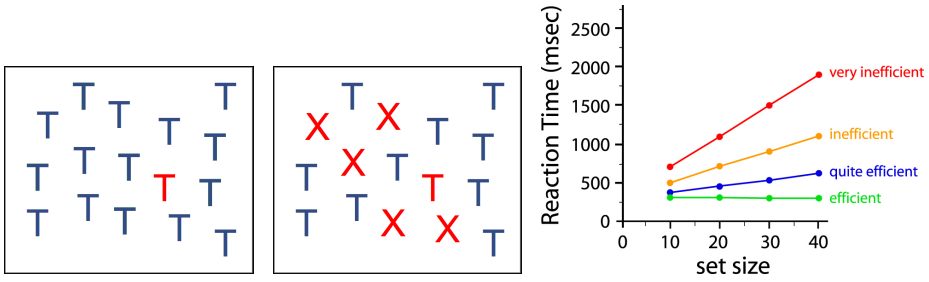
2.2.5 *Visual Search and Pop-out Effect.* An important tool in research on visual attention is *visual search* [Neisser 1967; Styles 1997; Wolfe 1998a]. The general question of visual search is: given a target and a test image, is there an instance of the target in the test image? We perform visual search all the time in every-day life. For example, finding a friend in a crowd is such a visual search task. Tsotsos has proven that the problem of *unbounded visual search* is so complex that it in practice is unsolvable in acceptable time[1] [Tsotsos 1987; 1990]. 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 (cf. Fig. 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, i.e., there is

[1]The problem is *NP-complete*, i.e., 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 X's and blue T's) 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 (Fig. adapted from [Wolfe 1998a]).

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 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" (cf. Fig. 3 (c)).

The concept of efficient search has been 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 nowadays referred to as *pop-out effect*, according to the subjective impression that the target leaps out of the display to grab attention (cf. Fig. 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 (cf. Fig. 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] (cf. sec. 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, that means the features which are early and pre-attentively 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*, that 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 (cf. Fig. 4). An explanation is

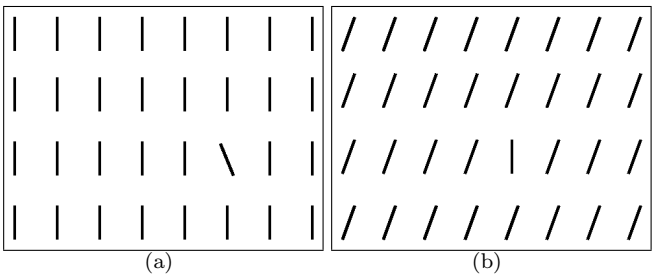


Fig. 4. Search asymmetries: it is easier to detect a tilted line among vertical distractors (a) than vice versa (b)

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 neuro-physiological 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 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 which area fulfills which task, the opinions diverge. We review several findings here.

Posner and Petersen [1990] describe three major functions concerning attention: first, orienting of attention, second, target detection, and third, 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, e.g., 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.

To sum up, 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, there exists a wide variety of theories and models on visual attention. Their objective is to explain and better understand human perception. Here, we introduce the theories and models which 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 has been one of the most influential theories in the field of visual attention. The theory was first introduced in 1980 [Treisman and Gelade] 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 (cf. sec. 2.2.5). An overview of the theory is found in [Treisman 1993].

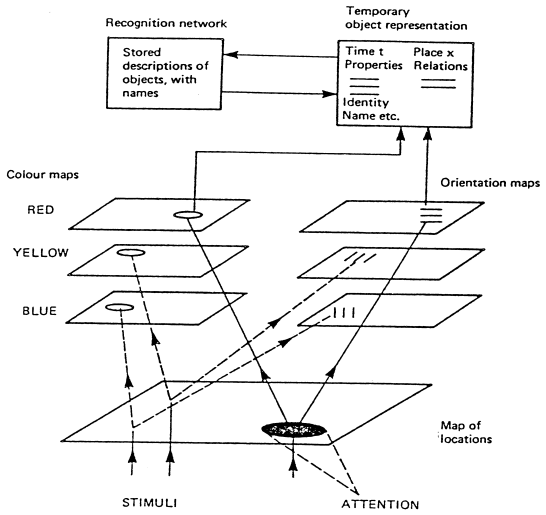


Fig. 5. Model of the *Feature Integration Theory (FIT)* (Fig. reprinted with permission from [Treisman and Gormican 1988] © 1988 American Psychological Association (APA)).

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 (cf. Fig. 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 non-target 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.* Beside 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

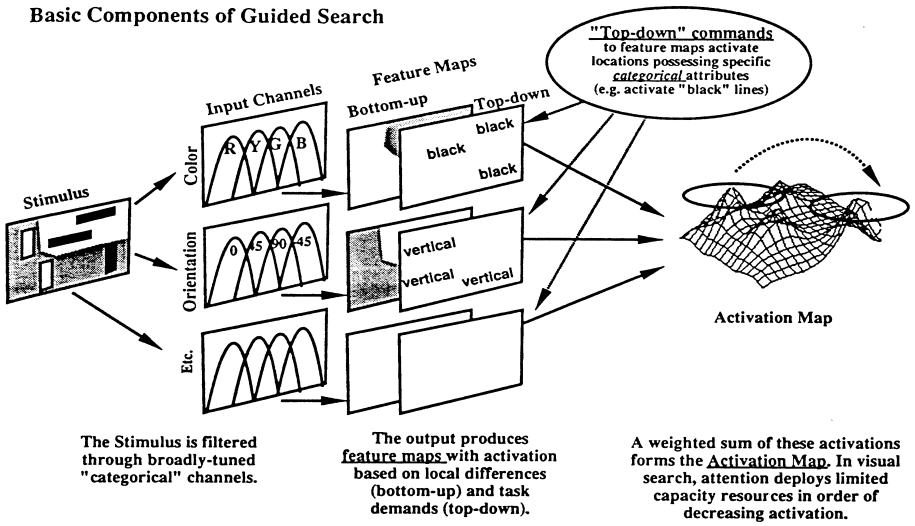


Fig. 6. The *Guided Search model* of Wolfe (Fig. reprinted with permission from [Wolfe 1994] ©1994 Psychonomic Society).

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 been also 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 which 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.

2.3.3 *Other theories and models.* Beside 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, that 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 models 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 a SDT experiment, one or several search displays are presented briefly and masked afterwards. 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 which have been used to predict human performance for detection and localization of targets have been presented in [Palmer et al. 1993; Verghese 2001; 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 non-attentional system provides setting information, for example, on the *gist* — the abstract meaning of a scene, e.g., beach scene, city scene, etc. — 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 and therefore, many computational attention systems have been built during the last three decades (mainly during the last 5-10 years). The systems which are considered here have in common that they built on the psychological and neurobiological concepts and theories which have been presented in the previous section. In contrast to the models described in sec. 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

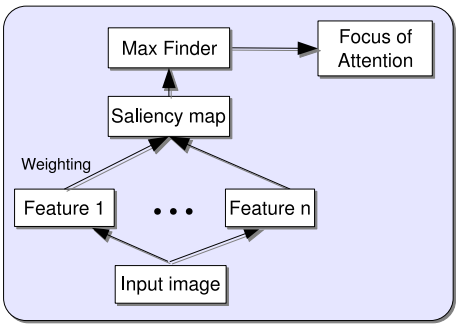


Fig. 7. General structure of most bottom-up attention systems.

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.

We start with a description of the general structure of typical computational attention systems (sec. 3.1). Then, we continue with a more detailed investigation of the characteristics of different approaches. Connectionist versus filter models are distinguished (sec. 3.2), the choice of different feature channels is discussed (sec. 3.3), and the integration of top-down cues is introduced (sec. 3.4). Finally, we provide a chronological overview of important computational attention systems (sec. 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 like 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 *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 (cf. Sec. 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 (cf. Fig. 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 sec. 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, i.e., 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; 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 human visual attention almost always top-down effects participate and guide our attention according to the current situation. These effects will be discussed in sec. 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 which 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, that means by single units which 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 like a GPU.

Since the focus of attention (FOA) is usually not on a single point but on a region (we call it *MSR (most salient region)*), 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, also the local or global surrounding of the region is considered [Navalpakkam et al. 2005; Frintrop et al. 2005]. The vector can be used to match the region to previously seen regions, e.g., 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 (cf. sec. 2.3.1).

To obtain a trajectory of image regions which mimics a human search trajectory, most common is a method 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 sur-

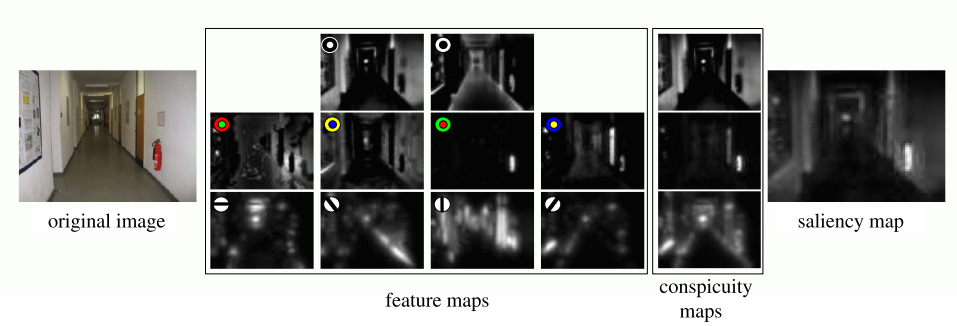


Fig. 8. Feature, conspicuity and saliency map(s) for an example image computed with the attention system VOCUS [Frintrop 2005]. 1st row: intensity maps (on-off and off-on). 2nd row: color maps (green, blue, red, yellow). 3rd row: orientation maps (0°, 45°, 90°, 135°). The feature map 'red' is weighted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

rounding 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, i.e., not all items in a search display are marked after they have been considered. That means, IOR works probably 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 which 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 sec. 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 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 for instance in [Olshausen et al. 1993; Postma 1994; Tsotsos et al. 1995; Baluja and Pomerleau 1997; 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 and 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 for instance in [Milanese 1993; Itti et al. 1998; Backer et al. 2001; Sun and Fisher 2003; Heidemann et al. 2004; Hamker 2005; 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; 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 above 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, for example 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 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, there exist different kinds of top-down influences. 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, but also 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 neuro-biological 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, that 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, i.e., 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, that 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 which can be integrated into attention models is the gist, i.e., 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 rule [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 not only knowledge about a target but also 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 non-salient 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 will be discussed in sec. 3.6.

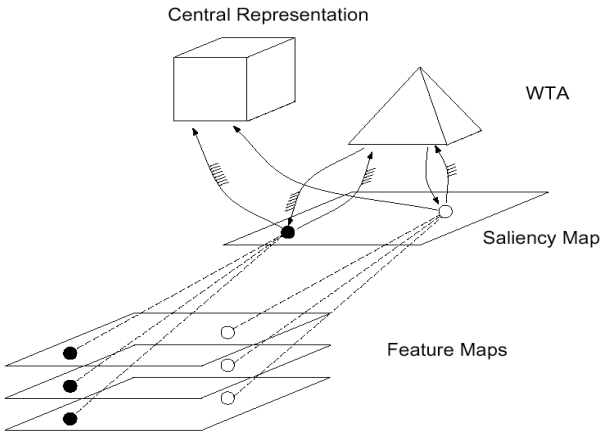


Fig. 9. The Koch-Ullman model. Different features are computed in parallel and their conspicuities are represented in several *feature maps*. A central *saliency map* combines the saliencies of the features and a *winner take all network (WTA)* determines the most salient location. This region is routed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

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 implementations and for many current computational attention systems. An important contribution of their work is the WTA network (see Fig. 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 are 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* which consists of a pyramidal architecture with an *inhibitory beam* (see Fig. 10). It is also possible to consider target-specific top-down cues by either inhibiting all regions with features different

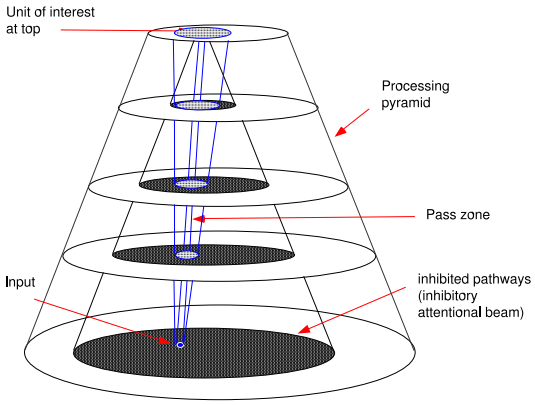


Fig. 10. The *inhibitory attentional beam* of Tsotsos et al. The selection process requires two traversals of the pyramid: first, the input traverses the pyramid in a feedforward manner (pass zone). Second, the hierarchy of WTA processes is activated in a top-down manner to localize the strongest item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

from the target features or regions of a specified location. The model has been implemented for several features, for example luminance, orientation, or 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 recently, it was 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)* (Fig. 11), a derivative of the Koch-Ullman model, that is steadily kept up to date by the group 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 also 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.

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

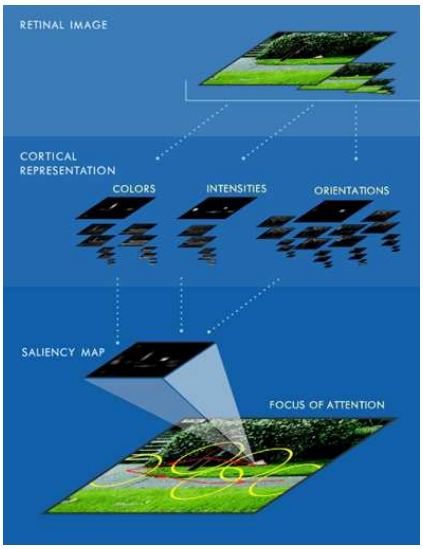


Fig. 11. Model of the *Neuromorphic Vision Toolkit (NVT)* by Itti et al. For each input image, image pyramids are computed to enable processing on different scales. Several feature channels investigate feature-dependent conspicuity independently. These are fused to a saliency map and a winner take all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

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: e.g., 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 like 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** lays special emphasis on closely mimicking the neural processes in the human visual cortex [Hamker 2005; 2006]. 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, that 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 by so called *match detection units*.

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** has several aspects which make it well suitable for applications in computer vision and robotics. The top-down part enables an easy, user-friendly search for target objects [Frintrop 2005]. The system is largely robust to illumination and viewpoint changes and it is real-time capable (50 ms per frame for a 400×300 pixel image on a 2.8 GHz 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 pre-knowledge influence the processing. Easiest is the evaluation on simple, artificial scenes containing pop-outs, 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 (cf. [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. Also Tatler et al. [2005] discovered that features like contrast, orientation energy, and chromaticity all differ between fixated and non-fixated 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 are less so. It may be also 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, as 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; 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. Also 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 non-fixated locations with signal detection and information theoretic techniques. In [Tatler et al. 2005], they state that "contrast and edge information was more strongly discriminatory than luminance or chromaticity". In [Baddeley and Tatler 2006], they 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 correlates between fixations and contrast were simply artefacts of their correlates 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 like 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 either 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 like image matching (sec. 4.1). The second, mid-level category considers attention as a front-end for high-level tasks as object recognition (sec. 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, i.e., to guide object manipulation or human-robot interaction (sec. 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 non-focused 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**, i.e., finding correspondences between two or more images which 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 which determine the feature saliencies of the ROI and its surrounding (cf. sec. 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 very powerful also 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, i.e., 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 results often 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].

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

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 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 was presented in [Siagian and Itti 2009]. Additionally 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.

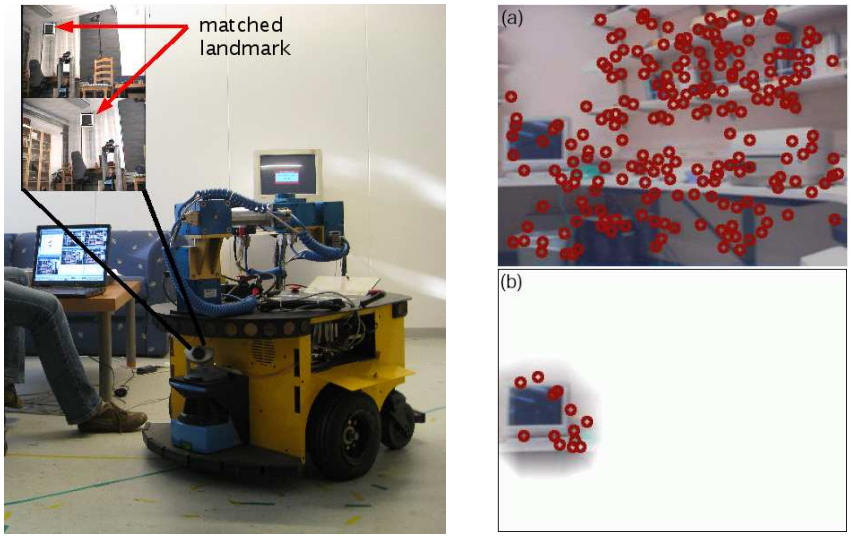
In the above examples, a map of the environment is initially known. Usually, it is obtained in a training phase. A more difficult task is **simultaneous localization and mapping (SLAM)** in which a robot initially does not know anything about its environment and has to build a map and localize itself inside the map at the same time. This topic was up to now rarely investigated in combination with visual attention. Frintrop et al. investigated the combination of visual attention and SLAM [Frintrop and Jensfelt 2008]. 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 Fig. 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



(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 which 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 (Fig. reprinted with permission from [Walther 2006]).

[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] combine an attention system with an object recognizer based on SIFT features [Lowe 2004] and show that the recognition results are improved by the attentional front-end (see Fig. 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. Non-salient objects are not detected. For some object classes like 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 above 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

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

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 towards 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 present 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 well from other objects in an object data base. A similar approach pursue Pessoa and Exel [1999].

4.3 Attention Systems for Guiding Robot Action

A robot which 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**, i.e., 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; 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, i.e., 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 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 to 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 disabled children. The wheelchair has a display as 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 Fig. 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

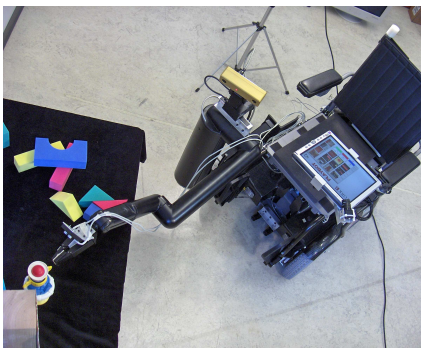


Fig. 13. PlayBot: a visually guided robotic wheelchair for disabled children. The selective tuning model of visual attention supports the detection of objects of interest (Fig. reprinted with permission from http://www.cse.yorku.ca/∼playbot).

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 to 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 paper gives a broad overview over 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 in the brain, resulting often in psychophysical theories or models. Neuro-biologists take a view directly into the brain with new techniques like 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 neuro-biological findings to improve their attention models and neuro-biologists 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 paper, 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, certainly open questions remain. A still largely unexplored area is the investigation of visual perception in dynamic scenes (but see e.g. [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 like prior knowledge, motivations, and emotions influence the visual system and are worth being investigated further. Interesting are also questions like "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 multi-processors 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 CPU's [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]. Interesting is also the investigation of how the concepts of attention apply to other sensors than cameras, e.g. 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 US. The European efforts are part of the emphasis on cognitive systems whereas the US 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 pre-selecting attention system which determines in advance the regions of highest potential interest in the sensor data.

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

REFERENCES

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

Aloimonos, Y., Weiss, I., and Bandopadhay, A. 1988. Active vision. *International Journal of Computer Vision (IJCV) 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. *Journal of Experimental Psychology: Human Perception and Performance 26,* 2, 834–846.

Aziz, M. Z. and Mertsching, B. 2007. Pop-out and IOR in static scenes with region based visual attention. In *ICVS Workshop on Computational Attention and Applications (WCAA 2007)*. Applied Computer Science Group, Bielefeld University, Germany, Bielefeld, Germany.

Backer, G. 2004. Modellierung visueller Aufmerksamkeit im Computer-Sehen: Ein zweistufiges Selektionsmodell für 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. on Pattern Analysis and Machine Intelligence (PAMI) 23(12)*, 1415–1429.

Bacon, W. and Egeth, H. 1994. Overriding stimulus-driven attentional capture. *Perception & 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 Research 46*, 2824–2833.

Balkenius, C. 2000. Attention, habituation and conditioning: Towards a computational model. *Cognitive Science Quarterly 1,* 2, 171–214.

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

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 Research 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, Chapter 11.

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 (April), 529 – 534.

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

Björkman, M. and Eklundh, J.-O. 2007. Vision in the real world: Finding, attending and recognizing objects. *Int'l Journal of Imaging Systems and Technology 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, 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 *Proc. of the Int'l Joint Conference on Artifical Intelligence (IJCAI 99)*. Stockholm, Sweden, 1146–1151.

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

Bruce, N. D. B. and Tsotsos, J. K. 2005b. Saliency based on information maximization. In *Proc. of Neural Information Processing Systems (NIPS)*. Vancouver, Canada.

Bundesen, C. 1990. A theory of visual attention. *Psychological Review 97*, 523–547.

Bundesen, C. 1998. A computational theory of visual attention. *Philosophical Transactions of the Royal Society of London, Series B 353*, 1271–1281.

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

Bur, A., Wurtz, P., Müri, R., and Hügli, H. 2007. Motion integration in visual attention models for predicting simple dynamic scenes. In *Human Vision and Electronic Imaging XII. Proceedings of SPIE*, B. E. Rogowitz and S. J. Pappas, Thrasyvoulos N.and Daly, Eds. Vol. 6492.

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

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

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

Cave, K. R. 1999. The FeatureGate model of visual selection. *Psychological Research 62*, 182–194.

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

Cherry, E. C. 1953. Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America 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 Information Processing-Letters and Reviews 2,* 1.

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

Clark, J. J. and Ferrier, N. J. 1988. Modal control of an attentive vision system. In *Proc. of the 2nd International Conference on Computer Vision.* Tampa, Florida, US.

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

Clark, J. J. and Ferrier, N. J. 1992. Attentive visual servoing. In *An Introduction to Active Vision*, A. Blake and A. Yuille, Eds. MIT Press, Cambridge Massachusetts, Chapter 10.

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

Corbetta, M. 1990. Frontoparietal cortical networks for directing attention and the eye to visual locations: Identical, independent, or overlapping neural systems? *Proc. 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. *Nature Reviews 3,* 3, 201–215.

Dankers, A., Barnes, N., and Zelinsky, A. 2007. A reactive vision system: Active-dynamic saliency. In *Proc. of the 5th Int'l Conf. on Computer Vision Systems (ICVS 2007).* Applied Computer Science Group, Bielefeld University, Bielefeld, Germany.

Desimone, R. and Duncan, J. 1995. Neural mechanisms of selective visual attention. *Annual Reviews of Neuroscience 18*, 193–222.

Deubel, H. and Schneider, W. X. 1996. Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research 36,* 12, 1827–1837.

Draper, B. A. and Lionelle, A. 2005. Evaluation of selective attention under similarity transformations. *Journal of Computer Vision and Image Understanding (CVIU), Special Issue on Attention and Performance 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. *Journal of Experimental Psychology 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. *Perception & Psychophysics 62,* 3, 425–451.

Egeth, H. E. and Yantis, S. 1997. Visual attention: control, representation, and time course. *Annual Review of Psychology 48*, 269–297.

Einhäuser, W., Spain, M., and Perona, P. 2008. Objects predict fixations better than early saliency. *Journal of Vision 8,* 14, 1–26.

Elazary, L. and Itti, L. 2008. Interesting objects are visually salient. *Journal of Vision 8,* 3:3 (Mar), 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. *Perception and Psychophysics 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, Chapter 5, 83–103.

Findlay, J. M. and Walker, R. 1999. A model of saccade generation based on parallel processing and competitive inhibition. *Behavioral and Brain Sciences 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 *Conference on Intelligent Robots and Systems (IROS)*. Beijing, China, 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 *Proc. of the Int'l Workshop on Attention and Performance in Computer Vision (WAPCV)*. Graz, Austria, 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. *Journal of 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. Published 2006 in Lecture Notes in Artificial Intelligence (LNAI), Vol. 3899, Springer Verlag.

Frintrop, S. 2008. The high repeatability of salient regions. In *Proc. of ECCV workshop "Vision in Action: Efficient Strategies for Cognitive Agents in Complex Environments"*.

Frintrop, S., Backer, G., and Rome, E. 2005. Goal-directed search with a top-down modulated computational attention system. In *Proc. of the Annual meeting of the German Association for Pattern Recognition (DAGM)*. Lecture Notes in Computer Science (LNCS). Springer.

Frintrop, S. and Cremers, A. B. 2007. Top-down attention supports visual loop closing. In *Proc. of European Conference on Mobile Robotics (ECMR 2007)*. Freiburg, Germany.

Frintrop, S. and Jensfelt, P. 2008. Attentional landmarks and active gaze control for visual SLAM. *IEEE Trans. on Robotics, Special Issue on Visual SLAM 24,* 5 (Oct).

Frintrop, S. and Kessel, M. 2009. Most salient region tracking. In *Proc. of the IEEE Int'l Conf. on Robotics and Automation (ICRA'09)*. Kobe, Japan.

Frintrop, S., Klodt, M., and Rome, E. 2007. A real-time visual attention system using integral images. In *Proc. of the 5th Int'l Conf. on Computer Vision Systems (ICVS)*. Bielefeld, Germany.

Frintrop, S., Königs, A., Hoeller, F., and Schulz, D. 2010. A component-based approach to visual person tracking from a mobile platform. In *accepted for the International Journal of Social Robotics*.

Frintrop, S., Nüchter, A., Surmann, H., and Hertzberg, J. 2004. Saliency-based object recognition in 3D data. In *Proc. of the Int'l Conf. on Intelligent Robots and Systems (IROS)*. Conference: Sendai, Japan, 2167 – 2172.

Frintrop, S., Rome, E., Nüchter, A., and Surmann, H. 2005. A bimodal laser-based attention system. *J. of Computer Vision and Image Understanding (CVIU), Special Issue on Attention and Performance in Computer Vision 100,* 1-2 (Oct-Nov), 124–151.

Fritz, G., Seifert, C., and Paletta, L. 2004. Attentive object detection using an information theoretic saliency measure. In *Proc. of the 2nd Int'l Workshop on Attention and Performance in Computational Vision (WAPCV)*, L. Paletta, J. K. Tsotsos, E. Rome, and G. W. Humphreys, Eds. Conference: Prague, Czech Republic, 136–143.

Fritz, J. B., Elhilali, M., David, S. V., and Shamma, S. A. 2007. Auditory attention - focusing the searchlight on sound. *Current Opinion in Neurobiology 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 colour vision. *Nature Reviews Neuroscience 4*, 563–572.

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

Giesbrecht, B., Wodorff, M., Song, A., and Mangun, G. 2003. Neural mechanisms of top-down 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. *Journal of Computer Vision and Image Understanding (CVIU), Special Issue on Attention and Performance 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. *Machine Vision and Applications 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). *Psychological Review 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, 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 Psychology 25*, 43–110.

Itti, L. 2002. Real-time high-performance attention focusing in outdoors color video streams. In *Proc. SPIE Human Vision and Electronic Imaging IV (HVEI)*. San Jose, CA.

Itti, L. 2004. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing 13,* 10 (Oct).

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

Itti, L. and Baldi, P. 2009. Bayesian surprise attracts human attention. *Vision Research 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 *Proc. of the SPIE 48th Annual International Symposium on Optical Science and Technology*. Vol. 5200.

Itti, L. and Koch, C. 2001a. Computational modeling of visual attention. *Nature Reviews Neuroscience 2,* 3 (Mar), 194–203.

Itti, L. and Koch, C. 2001b. Feature combination strategies for saliency-based visual attention systems. *Journal of Electronic Imaging 10,* 1 (Jan), 161–169.

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

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

Johnson, A. and Proctor, R. 2003. *Attention: theory and practice.* Sage Publications.

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, Hillsadale, NJ, 187–203.

Kadir, T. and Brady, M. 2001. Saliency, scale and image description. *Int'l J. of Computer Vision 45,* 2, 83–105.

Kahneman, D. and Treisman, A. 1992. The reviewing of object files: Object-specific integration of information. *Cognitive Psychology 24*, 175–219.

Kandel, E. R., Schwartz, J. H., and Jessell, T. M. 1996. *Essentials of Neural Science and Behavior.* McGraw-Hill/Appleton & Lange.

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. *Human Neurobiology 4,* 4, 219–227.

Kootstra, G., Nederveen, A., and de Boer, B. 2008. Paying attention to symmetry. In *Proc. of the British Machine Vision Conference (BMVC).* Leeds, UK.

Kunar, M., Flusberg, S., and Wolfe, J. 2008. The role of memory and restricted context in repeated visual search. *Perception and Psychophysics 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. Jan.

Lee, K., Buxton, H., and Feng, J. 2003. Selective attention for cue-guided search using a spiking neural network. In *Proc. of the Int'l Workshop on Attention and Performance in Computer Vision (WAPCV).* Graz, Austria, 55–62.

Levin, D. 1996. Classifying faces by race: the structure of face categories. *Journal of Experimental Psychology: Learning, Memory, & Recognition 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 attentional 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. *Journal of Neuroscience 7,* 11, 3416–3468.

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

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *Int'l J. of Computer Vision (IJCV) 60,* 2, 91–110.

Maki, A., Nordlund, P., and Eklundh, J.-O. 2000. Attentional scene segmentation: Integrating depth and motion. *Computer Vision and Image Understanding (CVIU) 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 (NY).

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 *Proc. of Int'l Conf. on Intelligent Robots and Systems (IROS).* IEEE, 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 NAVIS. 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 *Proc. SPIE 46 Annual Int'l Symposium on Optical Science and Technology.* Vol. 4479. 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 *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '94)*. Conference: Seattle, 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 *IEEE Proc. of the Int'l Conf. on Robotics and Automation (ICRA '05)*. Conference: Barcelona, Spain, 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. Hove, UK: Lawrence Erlbaum Associated Ltd., 83–104.

Muhl, C., Nagai, Y., and Sagerer, G. 2007. On constructing a communicative space in HRI. In *Proc. of the 30th German Conference on Artificial Intelligence (KI 2007)*, J. Hertzberg, M. Beetz, and R. Englert, Eds. Springer, Osnabrück, Germany.

Nagai, Y. 2009. From bottom-up visual attention to robot action learning. In *IEEE 8th Int'l Conf. on Development and Learning.*

Nakayama, K. and Mackeben, M. 1989. Sustained and transient components of focal visual attention. *Vision Research 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 *Proc. of the Conf. on Computer Vision and Pattern Recognition (CVPR).*

Navalpakkam, V. and Itti, L. 2006b. Top-down attention selection is fine-grained. *Journal of Vision 6,* 11 (Oct), 1180–1193.

Navalpakkam, V., Rebesco, J., and Itti, L. 2004. Modeling the influence of knowledge of the target and distractors on visual search. *Journal of Vision 4,* 8, 690.

Navalpakkam, V., Rebesco, J., and Itti, L. 2005. Modeling the influence of task on attention. *Vision Research 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. *Robotics and Autonomous Systems 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, 233–239.

Ogawa, T. and Komatsu, H. 2004. Target selection in area V4 during a multidimensional visual search task. *Journal of Neuroscience 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, Chapter 41, 251–257.

Oliva, A., Torralba, A., Castelhano, M. S., and Henderson, J. M. 2003. Top-down control of visual attention in object detection. In *Int'l Conf. on Image Processing (ICIP)*. Barcelona, Spain, 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. *Journal of Neuroscience 13,* 11 (November), 4700–4719.

Olshausen, B. A. and Field, D. J. 2005. How close are we to understanding V1? *Neural Computation 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.

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 *Proc. of European Conference on Mobile Robotics (ECMR 2005)*. Ancona, Italy, 8–13.

Ouerhani, N. and Hügli, H. 2000. Computing visual attention from scene depth. In *Proc. of Int'l Conf. on Pattern Recognition (ICPR 2000)*. Vol. 1. IEEE Computer Society Press, 375–378.

Ouerhani, N., Jost, T., Bur, A., and Hügli, H. 2006. Cue normalization schemes in saliency-based visual attention models. In *Proc. Int'l Cognitive Vision Workshop*. Graz, Austria.

Ouerhani, N., von Wartburg, R., Hügli, H., and Müri, R. 2004. Empirical validation of the saliency-based model of visual attention. *Electronic Letters on Computer Vision and Image Analysis 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. Human perception and performance 19,* 1, 108–130.

Palmer, S. E. 1999. *Vision Science, Photons to Phenomenology.* The 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 Research 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 *Proc. of the International Work-Conference on Artificial and Natural Neural Networks (IWANN '99)*, J. Mira and J. Sachez-Andres, Eds. Lecture Notes in Computer Science (LNCS), vol. 1606. Springer, Alicante, Spain, 850–859.

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

Peters, R. J. and Itti, L. 2008. Applying computational tools to predict gaze direction in interactive visual environments. *ACM Trans. on Applied Perception 5,* 2, Article 8.

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 Psychology 22*, 273–341.

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

Posner, M. I. 1980. Orienting of attention. *Quarterly Journal of Experimental Psychology 32*, 3–25.

Posner, M. I. and Petersen, S. E. 1990. The attentional system of the human brain. *Annual Review of Neuroscience 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.

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 *Proc. Workshop on Biologically Motivated Computer Vision (BMCV)*. Vol. 2525. Springer Verlag, Lecture Notes in Computer Science (LNCS).

Ramström, O. and Christensen, H. I. 2004. Object based visual attention: Searching for objects defined by size. In *Proc. of Int'l Workshop on Attention and Performance in Computational*

Vision (WAPCV), L. Paletta, J. K. Tsotsos, E. Rome, and G. W. Humphreys, Eds. Conference: Prague, Czech Republic, 9–16.

Rao, R., Zelinsky, G., Hayhoe, M., and Ballard, D. 2002. Eye movements in iconic visual search. *Vision Research 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. *International Journal of Robotics Research*. (in press).

Rauschenberger, R. 2003. Attentional capture by auto- and allo-cues. *Psychonomic Bulletin & Review 10,* 4 (Dec.), 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. *Psychological Science 8*, 368–373.

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

Rosenholtz, R. 2001. Search asymmetries? What search asymmetries? *Perception & Psychophysics 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 *Proc. 2nd Int'l Conf. on Technology and Aging*. Toronto, Canada.

Rothenstein, A. and Tsotsos, J. 2006a. Attention links sensing to recognition. *Image & Vision Computing Journal, Special Issue on Cognitive Vision Systems 26,* 1, 114–126.

Rothenstein, A. and Tsotsos, J. 2006b. Selective tuning: Feature binding through selective attention. In *Proc. of International Conference on Artificial Neural Networks*. Athens, Greece.

Rybak, I., Gusakova, V., Golovan, A., Podladchikova, L., and Shevtsova, N. 1998. A model of attention-guided visual perception and recognition. *Vision Research 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 Transactions on Pattern Analysis and Machine Intelligence (PAMI) 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, New York, NY.

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

Scheier, C. and Egner, S. 1997. Visual attention in a mobile robot. In *Proc. of the IEEE Int'l Symposium on Industrial Electronics*. 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. of Experimental Psychology. Human Perception and Performance 5,* 3, 522–526.

Siagian, C. and Itti, L. 2009. Biologically inspired mobile robot vision localization. *IEEE Transaction on Robotics 25,* 4 (July), 861–873.

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

Styles, E. A. 1997. *The Psychology of Attention*. Psychology Press Ltd, East Sussex, UK.

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

Sun, Y. and Fisher, R. 2003. Object-based visual attention for computer vision. *Artificial Intelligence 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. of 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 Research 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 Research 46*, 1857–1862.

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

Torralba, A. 2003a. Contextual priming for object detection. *International Journal of Computer Vision 53,* 2, 169–191.

Torralba, A. 2003b. Modeling global scene factors in attention. *Journal of Optical Society of America A. Special Issue on Bayesian and Statistical Approaches to Vision 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 Psychology 12*, 97–136.

Treisman, A. M. and Gormican, S. 1988. Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review 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 Research 1225*, 119–132.

Tsotsos, J. K. 1987. A 'complexity level' analysis of vision. In *Proc. of International Conference on Computer Vision: Human and Machine Vision Workshop*. London, England.

Tsotsos, J. K. 1990. Analyzing vision at the complexity level. *Behavioral and Brain Sciences 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, 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. *Artificial Intelligence 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. *Journal of Computer Vision and Image Understanding (CVIU), Special Issue on Attention and Performance 100,* 1-2, 3–40.

Tsotsos, J. K., Verghese, G., Stevenson, S., Black, M., Metaxas, D., Culhane, S., Dickinson, S., Jenkin, M., Jepson, A., Milios, E., Nuflo, F., Ye, Y., and Mann, R. 1998. PLAYBOT: A visually-guided robot to assist physically disabled children in play. *Image and Vision Computing 16, Special Issue on Vision for the Disabled*, 275–292.

Tuytelaars, T. and Mikolajczyk, K. 2007. Local invariant feature detectors: A survey. *Foundations and Trens in Computer Graphics and 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? *Journal of experimental psychology. General 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. *Journal of Vision 5,* 1, 81–92. doi:10.1167/5.1.8.

Vijayakumar, S., Conradt, J., Shibata, T., and Schaal, S. 2001. Overt visual attention for a humanoid robot. In *Proc. International Conference on Intelligence in Robotics and Autonomous Systems (IROS 2001)*. Hawaii, 2332–2337.

Vincent, B. T., Troscianko, T., and Gilchrist, I. D. 2007. Investigating a space-variant weighted salience account of visual selection. *Vision Research 47*, 1809–1820.

Viola, P. and Jones, M. J. 2004. Robust real-time face detection. *International Journal of Computer Vision (IJCV) 57,* 2 (May), 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 *Proc. of Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Walther, D. and Koch, C. 2007. Attention in hierarchical models of object recognition. *Computational Neuroscience: Theoretical insights into brain funciton, Progress in Brain research 165*, 57–78.

Wells, A. and Matthews, G. 1994. *Attention and Emotion: A Clinical Perspective*. Psychology Press.

Wolfe, J. M. 1994. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review 1,* 2, 202–238.

Wolfe, J. M. 1998a. Visual search. In *Attention*, H. Pashler, Ed. Hove, U.K.: Psychology Press, 13–74.

Wolfe, J. M. 1998b. What can 1,000,000 trials tell us about visual search? *Psychological Science 9,* 1, 33–39.

Wolfe, J. M. 2001a. Asymmetries in visual search: An introduction. *Perception & Psychophysics 63,* 3, 381–389.

Wolfe, J. M. 2001b. Guided search 4.0: A guided search model that does not require memory for rejected distractors. *Journal of Vision, Abstracts of the 2001 VSS Meeting 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, New York, NY, Chapter 8.

Wolfe, J. M., Cave, K., and Franzel, S. 1989. Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance 15*, 419–433.

Wolfe, J. M. and Gancarz, G. 1996. Guided search 3.0: Basic and clinical applications of vision science. Dordrecht, Netherlands: Kluwer Academic. 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 Research 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? *Nature Reviews Neuroscience 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 *Proc. of Int'l Conf. on Intelligent Robots and Systems (IROS)*.

Xu, T., Pototschnig, T., Kühnlenz, K., and Buss, M. 2009. A high-speed multi-GPU implementation of bottom-up attention using CUDA. In *Proc. of the International Conference on Robotics and Automation, (ICRA)*.

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. *Nature Neuroscience 5*, 995–1002.

Yantis, S. and Serences, J. T. 2003. Cortical mechanisms of space-based and object-based attentional control. *Current Opinion in Neurobiology 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. of Experimental Psychology: Human Perception and Performance 23,* 1, 244–262.

Received February 2007; revised January and July 2008; accepted: November 2008