2

Background on Visual Attention

Visual attention is, as mentioned in the introduction, the selective process that enables us to act effectively in our complex environment. The term *attention* is common in everyday language and familiar to everyone. Nevertheless — or even therefore — it is necessary to clarify and define the term properly. Since visual attention is a concept of human perception, it is important to understand the underlying visual processing in the brain and to know about the psychophysical and neuro-biological findings in this field.

In this chapter, we first describe what we understand by attention and which concepts are important in this field (section 2.1). Then, we discuss the neural processes that underlie visual processing and attention in the human brain (section 2.2). Next, we introduce in section 2.3 several psychophysical models of visual attention that form the basis for many current computer models of attention and finally, we bridge the gap between biology and models by discussing which neuro-biological correlates exist for current attention models in psychology and computer science (cf. section 2.4). We conclude this chapter with a discussion in section 2.5.

2.1 Concepts of Visual Attention

In this section, we discuss several concepts of visual attention. First, we define the term attention, then we introduce the concepts of overt versus covert attention as well as of bottom-up versus top-down attention, and finally, we elaborate on visual search, its efficiency, pop-out effects, and search asymmetries.

2.1.1 What Is Attention?

The concept of selective attention refers to a fact that was already mentioned by Aristoteles: "It's not possible to perceive two things in one and the same indivisible time". Although we usually have the impression to retain a rich

S. Frintrop: VOCUS: A Visual Attention System..., LNAI 3899, pp. 7-31, 2006.

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 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. One convincing experiment on this topic is described in [Simons and Levin, 1998]: an experimenter approaches a pedestrian to ask for directions. During their conversation, two people carrying a door pass between the experimenter and the pedestrian and during that interruption, the first experimenter is replaced by a second experimenter. Even though subjects engaged in an interaction with both the first and the second experimenter and the second person was also wearing different clothing, 50% of the subjects did not notice the person 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 perform two or more tasks at the same time, or the ability to remain alert for long periods or time), for the purposes of this work, attention shall refer exclusively to perceptual selectivity.

If attention is needed to perform higher tasks in the human brain, and there are mechanisms that perform the attentional selection, this yields to a dichotomy of visual perception: one part is responsible for selecting the region of interest, the other one investigates the selected regions further [Neisser, 1967]. The mechanisms involved in the first task are called *pre-attentive* whereas the mechanisms operating on the selected data are called *attentive*. At which point this separation actually takes place is subject of the *early selection, late selection debate* which is discussed in [Pashler, 1997].

2.1.2 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 truth. As early as in 1890, William James posited that we are able to attend to peripheral locations of interest without moving our eyes [James, 1890]; this is referred to as *covert attention*. This mechanism should be well known to each of us when we "look out of the corner of our eyes".

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 [Findlay and Gilchrist, 2001] found that in tasks like reading and complex object search, *saccades* (rapid eye movements) 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 with a higher resolution. That covert and overt attention are not independent was shown by Deubel and Schneider [Deubel and Schneider, 1996]: it is not possible to attend to one location while moving the eyes to a different one.

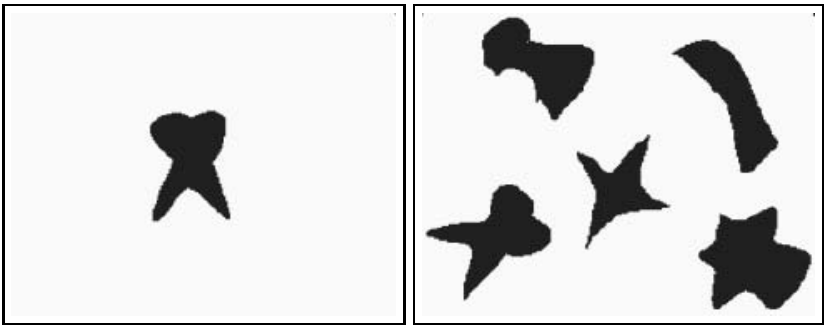
An advantage of covert attention is that it is independent of motor commands. Neither the eyes nor the head have to be moved to concentrate on a certain scene region. Therefore, the process is much faster than overt attention. Nevertheless, many experiments on visual attention investigate mainly overt attention since this can be easily measured with eye trackers. Covert attention is more difficult to investigate. Posner [Posner, 1980] proposes several methods to analyze covert attention: psychological investigations include the measuring of the reaction time to detect a target, neuro-biological methods include for example the measurement of the evoked potential amplitude or of changes in firing rates of single cells.

2.1.3 Bottom-Up Versus Top-Down Attention

Shifting the focus of attention can be initiated by two general categories of factors: *bottom-up factors* and *top-down factors* [Desimone and Duncan, 1995]. Bottom-up factors are derived solely from the conspicuousness of regions in a visual scene, for example by strong contrasts. 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 the "mental state" of the subject, that means by information from "higher" brain areas such as knowledge, expectations and current goals [Corbetta and Shulman, 2002]. That means, car holders are more likely to see the petrol stations in a street whereas bikers notice if there are cycle tracks. And if looking for a yellow highlighter on your desk, yellow regions attract the view more easily than other regions. Only parts of top-down processing are investigated by now, usually the parts concerning the knowledge about a target to be found. Other top-down influences like motivations, expectations, and emotions are much more difficult to control and to analyze and therefore much less is known on these aspects.

In psychophysics, top-down influences are often investigated by so called *cuing experiments*. In these experiments, a "cue" directs the attention to the target. Cues may have different characteristics: they may indicate *where* the



(a) Cue (b) Search array

Fig. 2.1. Cuing experiment: **(a)** a cue is presented for 200 ms. Thereafter, human subjects had to search for the cued shape in a search array **(b)**. The reaction time is usually faster when the cue matches the target exactly than when the cue was rotated (Fig. reprinted with permission from [Vickery et al., 2005]. ©2005 The Association for Research in Vision and Ophthalmology (ARVO))

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. 2.1). 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]. Other terms for top-down attention are *endogenous* [Posner, 1980], *voluntary* [Jonides, 1981], or *centrally cued* attention.

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 [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 text, regardless of how engrossed in the topic you were. This effect is called *attentional capture* (cf. Fig. 2.2).

Bottom-up attention is much better investigated. One reason is that the data-driven stimuli are easier controlled than the mental state that includes knowledge and expectations. Even less is known on the interaction of both processes.

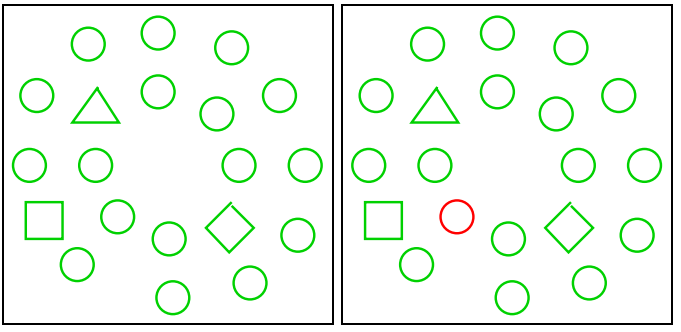


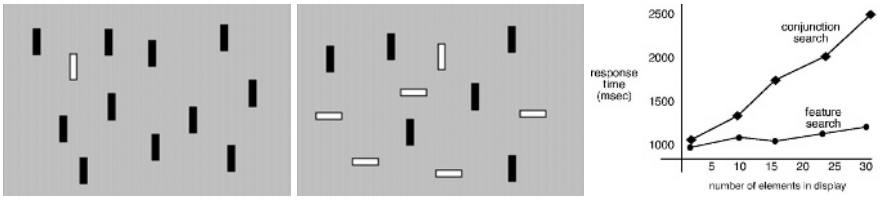
Fig. 2.2. Attentional capture: in both displays, human subjects had to search for the diamond. Although they knew that color was unimportant in this search task, the red circle in the right display slowed down the search about 65 ms (885 vs 950 ms) [Theeuwes, 2004]. That means, the color pop-out "captures" the attention independent of the task (Fig. adapted from [Theeuwes, 2004])

2.1.4 Visual Search and Pop-Out

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. Finding your friend in a crowd as discussed in the introductory example of this monograph is such a visual search task. In psychophysical experiments, the scene for a visual search task is usually an artificial composition of several items with different features such as color, orientation, shape, or size (cf. Fig. 2.2). The computational complexity of visual search has been investigated in [Tsotsos, 1990, Tsotsos, 2001]. *Unbounded visual search* (no target is given or it cannot be used to optimize search — for example, if the command is to find the odd-man-out) is proven to be *NP-complete*[1]. This is due to the fact that all subsets of pixels must be considered to find the target in a worst case. In contrast, the *bounded visual search* (the target is explicitly known in advance) requires linear time. Also, psychological experiments on visual search with known targets report that the search performance has linear time complexity 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, one measure of the *efficiency* of visual search is the *reaction time* or *response time (RT)* that a subject needs to detect the target. The RT is measured, for example, by pressing one button

[1] Problems that are *NP-complete* belong 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) Parallel search task (feature search) for white target among black distractors (pop-out effect).

(b) Serial search task (conjunction search) for a white vertical target among black vertical and white horizontal distractors.

(c) The time subjects need to detect the targets of images (a) and (b).

Fig. 2.3. The reaction time (RT) a subject needs to detect a target depends on the complexity of the search task **(c)**. If the target differs only in one feature from the distractors, the search time is almost constant with respect to the number of elements in the display (feature search); the target seems to pop out of the scene **(a)**. If the target is defined by a conjunction of features (conjunction search), the reaction time increases linearly with the number of distractors **(b)** (Fig. from [URL, 01])

if the target was detected and another if it is not present in the scene or by reporting a detail of the target. The efficiency is represented as a function that relates RT to the number of *distractors* (the elements that differ from the target) (cf. Fig. 2.3 (c)).

The searches vary in their efficiency: the flatter the slope of the function, the more efficient the search. Two extremes hereby are *serial* and *parallel* search. Parallel search means that the slope is near zero, i.e., there is no significant variation in reaction time if the number of distractors changes and a target is found immediately without the need to perform several shifts of attention. This effect occurs when the target differs in exactly one feature from the distractors, therefore the search is also called *feature search*. 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. 2.3 (a)). Scenes with pop-outs are sometimes also referred to as *odd-man-out* scenes, one example is the well known black sheep in a white herd. Parallel 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, there is parallel search but no pop-out effect.

Serial search instead occurs if the reaction time increases with the number of distractors. This is usually the case in *conjunction search tasks* in which the target is defined by several features, for example, finding a white, vertical line among white, horizontal and black, vertical ones (cf. Fig. 2.3 (b)). The strict separability of serial and parallel search is doubted nowadays. Experiments by Wolfe indicate that the increase in reaction time seems to be a continuum [Wolfe, 1998b].

There has been a multitude of experiments on visual search and many settings have been designed to find out which features enable parallel search and which do not. There have been several quite interesting experiments not only showing that there is parallel search for red among green or vertical among horizontal items, but also 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]. An interesting experiment was done by Jonides [Jonides and Gleitman, 1972]: the search for an O among letters is fast if subjects are told to search for the "zero" and slow if they are told to search for the letter "O" although the same setting was used in both experiments. This indicates that the pure semantic meaning of the element already influences visual search. Interesting is also that the search for a novel element among familiar ones is parallel [Wang et al., 1994]. This is an important effect that helps humans to ignore known things and focus processing on the new, most informative, sensory data.

The idea behind all these experiments is to find out the *basic features* of human perception, that means the features which are early and pre-attentively processed in the human brain. Testing the efficiency of visual search helps to investigate this since parallel 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 controversial. There appear to be about a dozen [Wolfe, 1998a]. In [Treisman and Gormican, 1988] the following features are named: colors, different levels of contrast (intensity), line curvature, line tilt (orientation) or misalignment, terminators, closure, direction of movement, stereoscopic disparity (depth) and quantitative values like length and number or proximity. Several findings indicate that basic features may also be learned. For example, Neisser mentions that finding special letters in a text is much more difficult for young children and illiterates than for people able to read [Neisser, 1967]. Anyone who has ever played the computer game Tetris for quite some time might also know this: after some time of playing, one seems to see the Tetris blocks everywhere in the environment[2]. Features not meeting the parallel search criterion are, for

[2] Annotation of J. Hertzberg: "This works also for Tangram!"

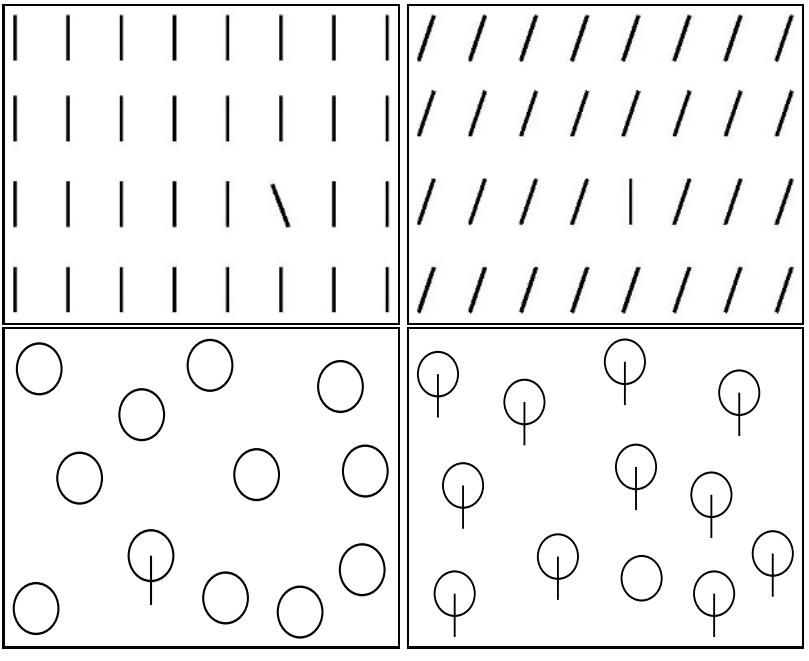


Fig. 2.4. Search asymmetries: it is easier to detect a tilted line among vertical distractors than vice versa (top) and to find a circle with a line among circles than vice versa (bottom)

example, line arrangements like intersection, juncture, and angles, topological properties like connectedness and containment, and relational properties like height-to-width ratio.

An important aspect in visual search tasks are *search asymmetries*, that means a search for stimulus A among distractors B produces different results from a search for B among As. An example is that finding a tilted line among vertical distractors is easier than vice versa (cf. Fig. 2.4). An explanation is proposed in [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 high interest: if it is clear which feature detectors are there in the human brain, it might be adequate to focus on the computation of these features and unnecessary to compute more.

2.2 The Neurobiology of Vision and Attention

Since visual attention is a concept of human perception, it is worth to regard the human visual system in more detail to get an insight into the nature of this concept. In this section, we first introduce the basic mechanisms that are involved in the processing of the visual information (section 2.2.1). Thereafter, we mention in section 2.2.2 the processes involved in assigning visual attention to regions of interest. While being far from an exhaustive explanation of the mechanisms in the human brain, we focus on describing the parts that are necessary for understanding the visual processing involved in selective attention. Further literature on this topic can be found, for example, in [Palmer, 1999, Kandel et al., 1996] and [Zeki, 1993].

2.2.1 The Human Visual System

Before going into the details of the mechanisms involved in the processing of visual information, let us briefly summarize the whole process in a few sentences [Palmer, 1999] (cf. Fig. 2.5): The light that achieves the eye is projected onto the retina and from there the optic nerve transmits the visual information 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*. 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). In the following, we discuss the processing in detail.

The Eye

The light that enters the eye through the *pupil* passes through the *lens*, travels through the clear *vitreous humor* that fills the central chamber of the eye and finally reaches the *retina* at the back of the eye (cf. Fig. 2.6, left). The retina is a light-sensitive surface and is densely covered with over 100 million photosensitive cells. The task of the photoreceptors is to change the electromagnetic energy of photons into neural activity that is needed as input by neurons.

There are two categories of photoreceptor cells in the retina: *rods* and *cones*. The rods are more numerous, about 120 million, and are more sensitive to light than the cones. However, they are not sensitive to color. The cones (about 8 million) provide the eye's color sensitivity: among the cones,

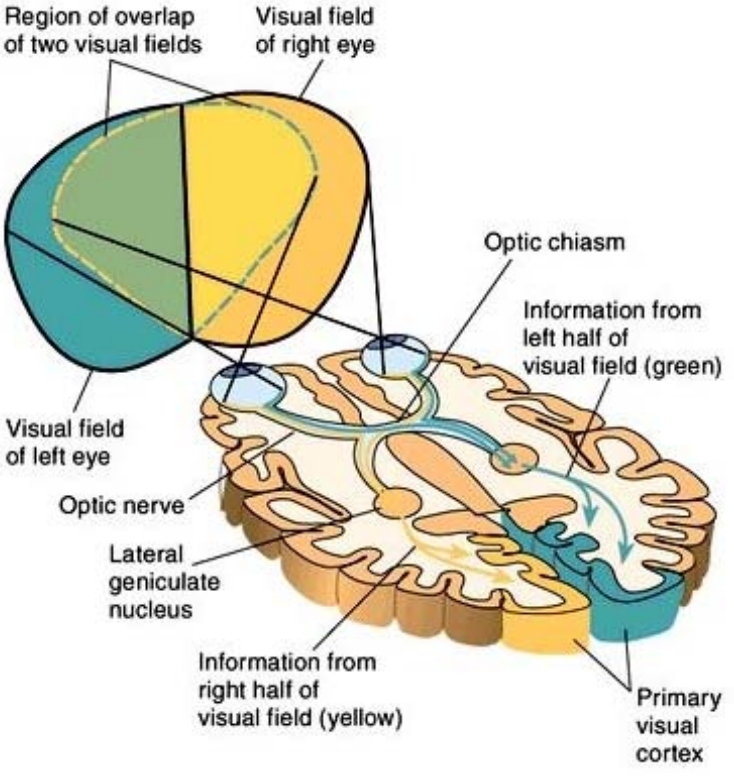
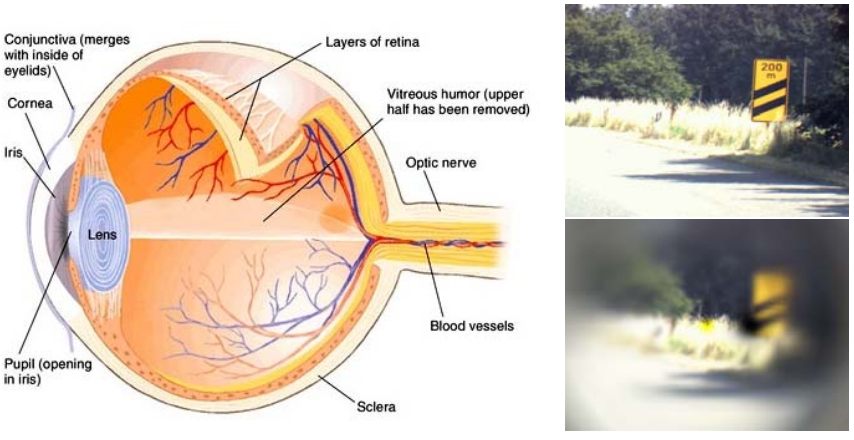


Fig. 2.5. The *primary visual pathway* in the human brain. The visual information enters the brain at the eye and is transmitted via the optic nerve to the optic chiasm. From here, most of the information is transmitted to the Lateral Geniculate Nucleus (LGN) and then to the Primary Visual Cortex (V1). From V1 the information is transmitted to "higher" brain areas (Fig. from: [URL, 02])

there are three different types of color reception: long-wavelength cones (L-cones) which are sensitive primarily to the red portion of the visible spectrum (64%), middle-wavelength cones (M-cones) sensitive to the green portion (32%), and short-wavelength cones (S-cones) sensitive to the blue portion (2%) (cf. Fig. 2.7 (a)). The cones are much more concentrated in the central yellow spot known as the *macula*. In the center of that region is the *fovea centralis* or briefly just *fovea*, a 0.3 mm diameter rod-free area with very thin, densely packed cones. It is the center of the eye's sharpest vision. This arrangement of cells has the effect that we do not perceive every part of the visual scene with the same resolution, but instead perceive the small region currently fixated in a high resolution and the whole surrounding only diffuse and coarse. An example of a scene as we perceive it is shown in Fig. 2.6, right.



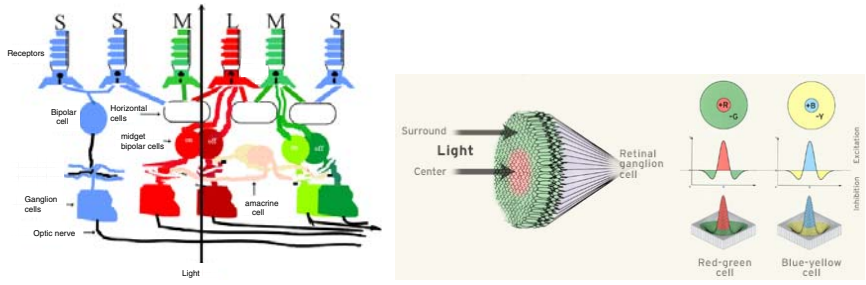
(a) The human eye (b) Perception by retina

Fig. 2.6. (a) the human eye. The incoming light traverses the lens and the vitreous humor and finally reaches the retina. From there the visual information is transmitted via the optic nerve to the brain for further processing (Fig. from: [URL, 02]); **(b)** the cells on the retina are concentrated in one region, the fovea centralis. This is the region with the eye's sharpest vision. Regions in the surrounding are perceived only diffuse as is visualized in this example. The upper image shows the original image, the lower one depicts the scene as perceived by the retina (Images from: [URL, 05])

The photoreceptors are connected via bipolar cells with the ganglion cells (cf. Fig. 2.7 (a)). Whereas photoreceptors and bipolar cells respond by producing graded potentials, the ganglion cells are the first cells which produce spike discharges and so transform the analog signal into a discrete one.

The receptive field of a ganglion cell is circular and separated into two areas: a center area and a surround area (cf. Fig. 2.7 (b)). There are two different types of cells: *on-center cells* which respond excitatorily to light at the center and *off-center cells* which respond inhibitorily to light at the center. The area surrounding the central region always has the opposite characteristic [Palmer, 1999]. There are small ganglion cells (P ganglion cells, parvus = small) and large ones (M ganglion cells, magnus = large). P ganglion cells receive their input just from the cones and are more sensitive to color than to black and white, whereas the M ganglion cells receive input from both rods and cones and are more sensitive to luminance contrasts [Palmer, 1999].

An important question now is: how is the color opponency (red-green and blue-yellow) derived from the outputs of the three-cone system? The red-green contrast is derived from combining the excitatory input from the L-cones and the inhibitory input from the M-cones, essentially subtracting the signals from the L- and M-cones to compute the red-green component of the



(a) The cells on the retina (b) The ganglion cells

Fig. 2.7. (a) there are the three different types of cones in the retina: L-cones ("red"), M-cones ("green") and S-cones ("blue"). They transmit the visual information to the bipolar cells which send it to the ganglion cells (Fig. from: [Kaiser, 1996], copyright ©1996-2004 Peter K. Kaiser); **(b)** the ganglion cells are separated into on-center cells, which respond excitatorily to light at the center and off-center cells, which respond inhibitorily to light at the center. For colors, there is a red-green and a blue-yellow antagonism resulting in red-green, green-red, blue-yellow, and yellow-blue cells (Fig. from: [URL, 03])

stimulus $(L - M)$. The green-red contrast is equally determined by $(M - L)$. The blue-yellow contrast is derived from the excitatory output of S-cones and the inhibitory sum of the M- and L-cones $(S - (L + M))$ and the yellow-blue contrast is determined by the excitatory sum of the M- and L-cones and the inhibitory output of the S-cones $((M+L)-S)$. Finally, the luminance contrast is derived by summing the excitation from all three cone types $(S + M + L)$ (on-off contrast) or by summing their inhibitory output $(-S - M - L)$ (off-on contrast) [Palmer, 1999].

The Optic Chiasm

The axons of the ganglion cells leave the eye via the optic nerve, which leads to the *optic chiasm*. Here, the information from the two eyes is divided and transferred to the two hemispheres of the brain: one half of each eye's information is crossed over to the opposite side of the brain while the other remains on the same side. The effect is, that the left half of the visual field goes to the right half of the brain and vice versa.

From the optic chiasm, two pathways go to each hemisphere: the smaller one goes to the *superior colliculus*, which is e.g. involved in the control of eye movements. The more important pathway goes to the LGN of the thalamus and from there to higher brain areas.

The Lateral Geniculate Nucleus (LGN)

The *Lateral Geniculate Nucleus (LGN)* consists of six main layers composed of cells that have center-surround receptive fields similar to those of retinal ganglion cells but larger and with a stronger surround. Four of the LGN layers consist of relatively small cells, the *parvocellular cells*, the other two of larger cells, the *magnocellular cells*. The parvocellular cells process mainly the information from the P-cells of the retina and are highly sensitive to color, especially to red-green contrasts [Gegenfurtner, 2003], whereas the magnocellular cells transmit information from the M-cells of the retina and are highly sensitive to luminance contrasts. Below those six layers lie the koniocellular sub layers, which respond mainly to blue-yellow contrasts [Gegenfurtner, 2003]. From the LGN, the visual information is transmitted to the *primary visual cortex* at the very back of the brain.

The Primary Visual Cortex (V1)

The *primary visual cortex* is with some 200 million cells the largest cortical area in primates and is also one of the best-investigated areas of the brain. It is known by many different names. Besides the primary visual cortex, the most common ones are *V1* (the abbreviated form) and the *striate cortex* (due to its striped appearance).

V1 is essentially a direct map of the field of vision, organized spatially in the same fashion as the retina itself. In other words, any two adjacent areas of the primary visual cortex contain information about two adjacent areas of the retinal ganglion cells. However, V1 is not exactly a point-to-point map of the visual field. Although spatial relationships are preserved, the densest part of the retina, the fovea, takes up a much smaller percentage (1%) of the visual field than its representation in the primary visual cortex (25%).

The primary visual cortex contains six major layers, giving it a striped appearance. The cells in V1 can be classified into three types: *simple cells, complex cells*, and *hypercomplex cells*. As the ganglion cells, the simple cells have an excitatory and an inhibitory region. Most of the simple cells have an elongated structure and, therefore, are orientation selective, that means, they fire most rapidly when exposed to a line or edge of a particular direction [Palmer, 1999]. Complex cells take input from many simple cells. They have larger receptive fields than the simple cells and obtain responses from every part of the receptive field. Furthermore, they are highly nonlinear and sensitive to moving lines or edges. Hypercomplex cells, in turn, receive as input the signals from complex cells. These neurons are capable of detecting lines of a certain length or lines that end in a particular area.

The Extrastriate Cortex and the Visual Pathways

From the primary visual cortex, a large collection of neurons sends information to higher brain areas. These areas are collectively called *extrastriate cortex*,

in opposite to the striped architecture of V1. The areas belonging to the extrastriate cortex are V2, V3, V4, the infero-temporal cortex (IT), the middle temporal area (MT or V5) and the posterior-parietal cortex (PP). The notation V1 to V5 comes from the former belief that the visual processing would be serial.

On the extrastriate areas, much less in known than on V1. It was not before the 1980's that these areas were examined in detail since at this time the advent of functional imaging methodologies has opened the way for closer examination of cortical areas in the intact human brain. One of the most important findings was that the processing of the visual information is not serial — that means the information is not transmitted from one area to the next — but highly parallel. Recently, 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 functional separation already started in the retina with the M-cells and P-cells and results in several pathways leading to different brain areas in the extrastriate cortex. The statements on the number of existing pathways differ: the most common belief is that there are three main pathways, one color pathway, one form pathway, and one motion pathway which is also responsible for depth processing [Kandel et al., 1996]. Other researchers mention four pathways by separating the motion pathway into one motion and one depth pathway [Livingstone and Hubel, 1987, Palmer, 1999] whereas some mention one color, one motion and two form pathways [Zeki, 1993]. The reason for this discordance is that firstly the pathways are not completely isolated and secondly the investigation of the extrastriate cortex has only started several years ago and its functionality is still not completely understood.

The color and form pathways result from the P-cells of the retina and the parvocellular cells of the LGN, go through V1, V2, and V4 and end finally in IT, the area where the recognition of objects takes place. In other words, IT is concerned with the question of "what" is in a scene. Therefore, the color and form pathway together are also called the *what pathway*. Other names are the *P pathway* or *ventral stream* because of its location on the ventral part of the body. The motion (and depth) pathway result from the M-cells of the retina and the magnocellular cells of the LGN, go through V1, V2, V3, MT (V5), and the parieto occipale area (PO) and end finally in PP, responsible for the processing of motion and depth. 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 is considered to lie dorsally. The distinction into "where" and "what" pathway traces back to [Ungerleider and Mishkin, 1982]; a visualization of these pathways is shown in Fig. 2.8.

Newer findings even propose that there is much less segregation of feature computations than suggested by these different pathways. It is indicated that luminance and color are not separated but there is a continuum of cells,

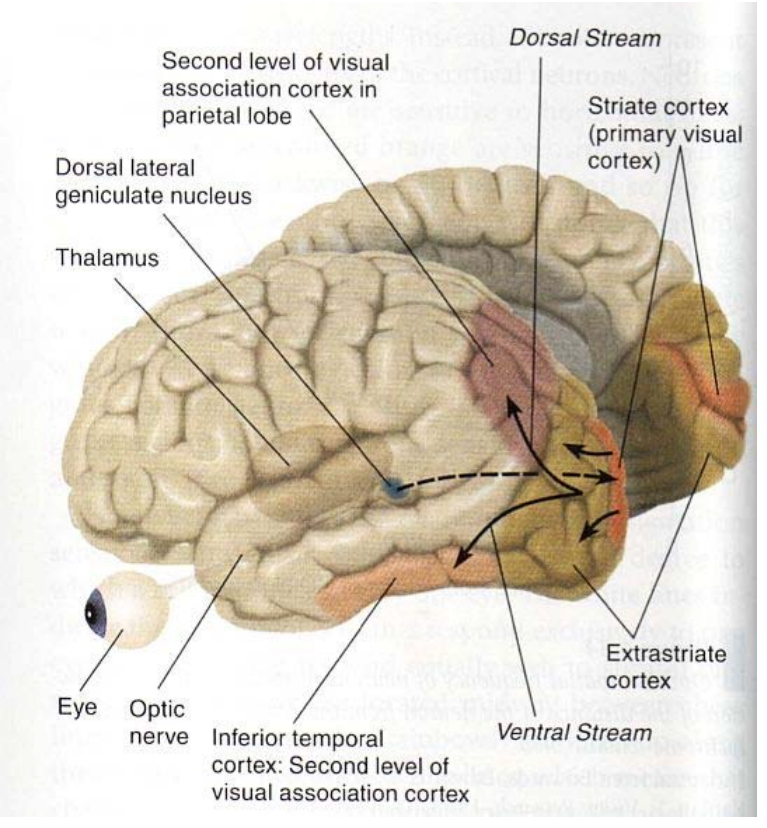


Fig. 2.8. The visual processing is divided functionally: the *ventral stream* leads to the inferior temporal cortex (IT) where object recognition takes place ("what" pathway) whereas the *dorsal stream* leads to the parietal lobe where motion and depth are processed ("where" pathway) (Fig. from: [URL, 04])

varying from cells that respond only to luminance, to a few cells that do not respond to luminance at all [Gegenfurtner, 2003]. Furthermore, neurons in the cortex can have a chromatic preference not only for red, green, yellow, or blue, but for any hue [Lennie et al., 1990], and V4, usually claimed to be the "color center" of the brain, processes also many other aspects of spatial vision. Additionally, the form processing is not clearly segregated from the processing of color since most cells that respond to oriented edges respond also to color contrasts. So a more correct view, at least as it is seen currently, is that some cells respond more to one kind of feature than to another one and certain brain areas have a prevalence of processing certain features but a clear segregation does not exist.

Finally, it is worth to mention that although the processing of the visual information was so far described in a feed-forward manner, it is usually bi-directional. Top-down connections from higher brain areas influence the processing and go down as far as LGN. Also lateral connections combine the different areas, for example, there are connections between V4 and MT, showing that the "what" and "where" pathway are not completely separated. The simplification of the last sections shall help to get an impression of the overall concept, but it should not be forgotten that the whole processing is much more complex and not yet completely understood at all.

2.2.2 Attentional Mechanisms in the Human Brain

The mechanisms of selective attention in the human brain still belong to the unsolved problems in the field of research on perception. Perhaps the most prominent outcome of new 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].

The more specific attentional mechanisms are carried out by a network of anatomical areas [Corbetta and Shulman, 2002]. Important areas of this network are PP, SC, the Lateral IntraParietal area (LIP), the Frontal Eye Field (FEF) and the pulvinar. Regarding the question which area fulfills which task, the opinions fall apart. 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 et al. 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 the movements of the eyes are the FEF, an area of the prefrontal cortex, 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], or, in most recent findings, at V4 [Mazer and Gallant, 2003].

Recently, 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 frontal eye fields (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. The results from [Corbetta and Shulman, 2002], which show that two independent but interacting brain areas are associated with bottom-up and top-down attentional mechanisms, support these findings.

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 into attentional processes but the accurate task and behavior of each area as well as the interplay among them still remain an open question.

2.3 Psychophysical Models of Attention

In the field of psychology, there exists a wide variety of models on visual attention. Their objective is to simulate behavioral data and thereby to explain and better understand human perception. There are descriptive models and models that are computationally implemented. The latter ones are especially well suited for comparison with psychophysical data obtained from experiments with humans. A review on computational models with a psychological objective is found in [Heinke and Humphreys, 2004]. In contrast to the models presented in this chapter, the computational systems of the next chapter intend to improve computer vision systems. However, there is an overlap of psychologically and technically motivated models and some of the mentioned approaches might be categorized in this as well as in the next chapter.

Here we describe two psychophysical models in detail because they belong to the best-known models in the field and have the greatest impact on this work. The first, introduced in section 2.3.1, is the *Feature Integration Theory* of Treisman and the second one is the *Guided Search Model* of Wolfe, described in section 2.3.2. In section 2.3.3, we mention several additional models.

2.3.1 Treisman's Feature Integration Theory

The *Feature Integration Theory (FIT)* of Treisman is one of the best known and most accepted theories in the field of visual attention. The theory was first introduced in 1980 [Treisman and Gelade, 1980] but it was steadily modified and adapted to current research findings. An overview of the theory is found in [Treisman, 1993], a model of FIT is depicted in Fig. 2.9.

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

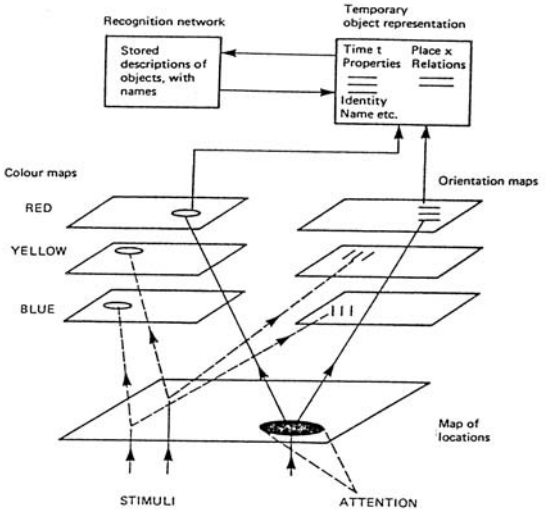


Fig. 2.9. Model of the *Feature Integration Theory (FIT)* of Treisman. Features such as color and orientation are coded automatically, pre-attentively, and in parallel. Each feature dimension consists of several *feature maps* such as red, yellow, and blue for color. The saliencies of the feature maps are coded in the *master map of locations*. When attention is focused on one location in this map, it allows retrieval of the features that are currently active at that location and creates a temporary representation of the object in an *object file* (Fig. reprinted with permission from [Treisman and Gormican, 1988]. ©1988 American Psychological Association (APA))

and Gelade, 1980]. Information from the resulting *feature maps* — topographical maps that highlight saliencies 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.

One of the main statements of the feature integration theory is that a target is detected easily, fast, and in parallel (pop-out) if it differs from the distractors in exactly one feature and the distractors are homogeneous. If it differs in more than one feature (conjunctive search) focal attention is required resulting in serial search. In later work, Treisman pointed out that information about the target object, represented in so called *object files*, , influences the search task by inhibiting the feature maps [Treisman, 1993].

Finally, it may be mentioned that Treisman uses the notation *feature* for intra-dimensional characteristics like red or horizontal and the notation *dimension* for supersets of these features, for example, color or orientation. In other approaches, the term feature is used for the dimensions.

2.3.2 Wolfe's Guides Search

Beside Treisman's Feature Integration Theory, the *Guided Search Model* of Wolfe is among the most important work in the field of psychophysical models of visual attention. Originally, the model was created as an answer to some criticism on early versions of the FIT. During the years, a competition arose between Treisman's and Wolfe's work, resulting in continuously improved versions of the models.

The basic goal of the model is to explain and predict the results of visual search experiments. There has 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. According to versions of software, the different versions of the system have been denoted with 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]. 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 minor changes, 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 Fig. 2.10. It shares many concepts with the FIT, but is more detailed in several aspects which are necessary for computer implementations. Alike FIT, it models several feature maps but unlike FIT it does not follow the idea that there are separate maps for each *feature type* (red, green, ...). There is only one map for each *feature dimension* (color, orientation, ...) and within each map, different feature types are represented. However, Wolfe mentions that there is evidence for differences between features. For example, there may be multiple color maps but only one orientation map [Nothdurft, 1993]. The features considered in the implementation are color and orientation.

Comparable to the *master map of location* in FIT, there is an *activation map* in Guided Search in which the feature maps are fused. But in contrast to at least the early versions of FIT, in Guided Search the attentive part profits from the results of the pre-attentive one. The fusion of the feature maps is done by summing up.

Additionally to this bottom-up behavior, the model also considers the influence of top-down information. To realize this, for each feature there is not only a bottom-up but also a top-down map. The latter map selects the feature type which distinguishes the target best from its distractors. This is not necessarily the feature with the highest activation for the target. Note that only one feature type is chosen, that means for an orange target the image regions with red portions are highlighted. It is not considered that a target might have different feature types, for example, 70% red and 30% yellow.

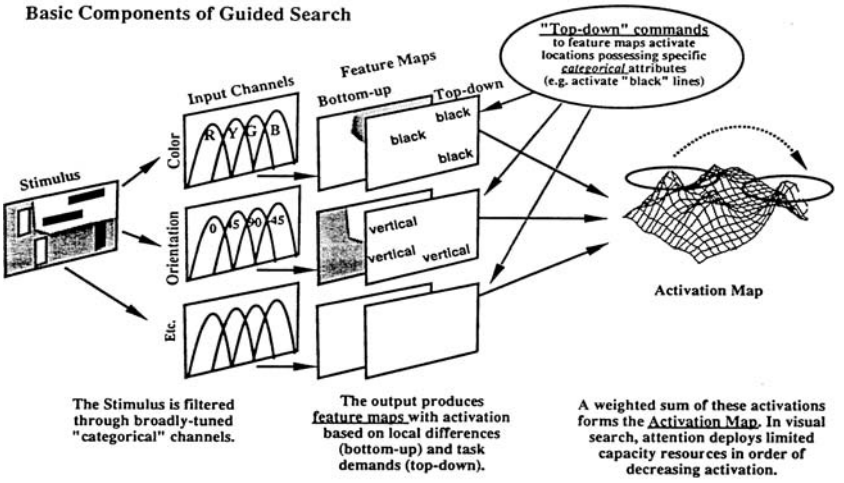


Fig. 2.10. The *Guided Search model* of Wolfe. One map for each feature dimension codes the properties of a scene concerning several feature types. Additionally to these bottom-up maps, top-down maps highlight the regions with task-specific attributes. A weighted sum of these activations forms the *activation map* (Fig. reprinted with permission from [Wolfe, 1994]. ©1994 Psychonomic Society)

2.3.3 Additional Models

Besides the Feature Integration Theory of Treisman and the Guided Search Model of Wolfe, there is a wide variety of psychophysical models on visual attention. The often used metaphor of attention as a *spotlight* comes from the *zoom lens model* [Eriksen and James, 1986]. 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 et al., 2002].

A formal mathematical model is presented in [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 selected is the element that first finishes processing (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, for example, in [Bundesen, 1998].

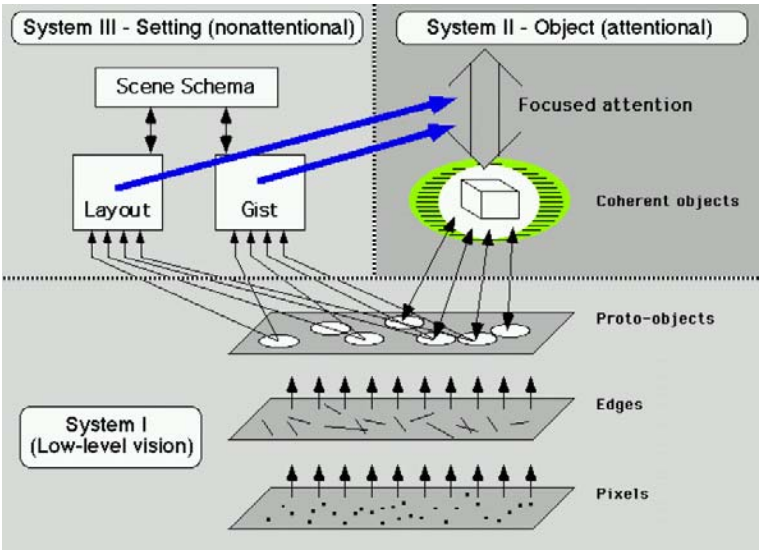


Fig. 2.11. The *triadic architecture* of Rensink suggests that visual perception is carried out via the interaction of three different systems: in the low level system, early level processes produce volatile proto-objects rapidly and in parallel. In system II, focused attention grabs these objects and in system III, setting information guides the attention to various parts of the scene (Fig. reprinted with permission from [Rensink, 2000]. ©2000 Psychological Press)

Recently, an interesting theoretical model has been introduced in [Rensink, 2000, Rensink, 2002]. His *triadic architecture* is very detailed and fits well for simulating it in a computer implementation (cf. Fig. 2.11). This was partially considered in [Navalpakkam et al., 2005]. The architecture consists of three parts: first a low-level vision system which produces *proto-objects* rapidly and in parallel. The proto-objects result from linear and not-linear processing of the input scene and are "quick and dirty" representations of objects or object parts that are limited in space and time.

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 selection of 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. Whereas the first two aspects resemble the traditional approaches of a pre-attentive and an attentive processing stage, the third part of the model is new and seems to be a promising extension of existing models.

2.4 From Biology to Models: Biological Correlates for Attentional Mechanisms

In this section, we will discuss to which extent the concepts usually used by psychological and computational models of attention are supported by neurobiological evidence. Some of these (computational) concepts will be introduced not until the next chapter but we discuss the correlation here since often the neuro-biological evidence forms the basis for these mechanisms.

Feature Maps

In rather all psychological and computational models of attention the processing of distinct features is parallelized in separate feature channels. This separation is much stricter than the processing in the human brain suggests: there are different neurons and also different brain areas specialized for the processing of certain features but the whole processing is much more intertwined than posited by most models [Gegenfurtner, 2003]. However, the distinction into several pathways for color, form, motion, and depth coincides to some extent with the distinct feature channels. Even if these pathways may not exist in their pure forms, they nevertheless refer to the bias of certain brain regions.

However, there is usually no one-to-one mapping between the psychological features and the biological pathways. Whereas psychological findings claim that there are about a dozen basic features [Wolfe, 1998a], the biological pathways are argued to be limited to three or four [Palmer, 1999]. Interestingly, three of the suggested neural pathways usually coincide each with one psychological feature channel, namely motion, color, and depth, whereas there are several psychological feature channels concerning form processing: there are feature maps for line curvature, line tilt or misalignment, terminators, and closure. Since newer findings suggest that there is no separate form pathway but many cells are responsible for edge detection as well as for color processing [Gegenfurtner, 2003], the psychological feature channels seem to correspond to several brain areas.

Center-Surround Mechanisms

The center-surround mechanisms that are used in most computational models of attention and in several psychological ones to determine the feature contrast regarding intensity or color have their neuro-physiological correlates on many different places in the brain: already the ganglion cells in the retina are separated into on-center and off-center cells. Later, cells in the LGN, V1, and the extrastriate cortex continue in responding to contrasts with these mechanisms.

Worth to mention is that some computational models combine the center-surround differences for on-off and for off-on instead of computing both [Itti

et al., 1998, Ouerhani et al., 2004]. This is not only contrary to the processing in the human brain, it also leads to problems: several pop-out effects are not achieved and top-down guidance for particular feature types is not possible. We discuss this in detail in section 4.1.1 (page 59).

Color Perception

As mentioned before, the perception and processing of color starts in the retina with different types of photoreceptors. There are three types of receptors with preferences for the colors red, green, and blue. Later, the processing is extended from this trichromatic architecture to the opponent processing with the color opponencies red-green and blue-yellow.

Psychological models often do not touch the question of how the color feature is processed in detail and if they do they usually focus on a three-color or double opponency approach but do not consider both. Computational models usually take RGB images as input. This correlates to the three-cone system in the retina. The further computation of colors differs strongly in different systems. Most use directly the RGB input whereas some first convert the image to a different color space. Most systems consider the red-green and blue-yellow opponency, but often it is not considered that there are separated mechanisms for red-green and for green-red as well as for blue-yellow and yellow-blue. Instead, the computation is combined, what leads to problems in combination with the center-surround mechanisms concerning several pop-out effects and top-down guidance for particular feature types (see section 4.1.1, page 59 for details).

The Saliency Map

Until recently, the opinions on whether there is a "saliency map" in the brain that collects the saliencies of the feature channels and directs the focus of attention were highly controversial. Several groups believed in such a saliency map, whereas others declined this view. Recently, there is increasing evidence that there is a structure in the brain representing a retinotopic saliency map that guides exploratory eye movements and is influenced by bottom-up as well as by top-down cues. As mentioned before, the opinions on which brain area fulfills this part are controversial. Candidates are the FEF [Bichot, 2001], LIP [Gottlieb et al., 1998], SC [Findlay and Walker, 1999], and, most recently, V4 [Mazer and Gallant, 2003].

It remains to mention that the organization of such a neurological "saliency map" is different from the saliency maps in most psychological and computational models. In the brain, this map is rather a collection of neurons, each with its own specialized behavior, than a map with the same behavior for each element.

Bottom-Up Versus Top-Down

Although there is agreement that top-down cues play an important role in the processing of visual information and it is known that there are numerous connections from higher brain areas to the areas of basic processing, the details of these processes are still not known at all. As mentioned before, in [Corbetta and Shulman, 2002] the authors claim that two independent but interacting brain areas are associated with the two attentional mechanisms, which interact during normal human perception. The areas involved in top-down biasing are as per [Kastner and Ungerleider, 2001] the superior parietal lobule (SPL), the frontal eye fields (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. However, which part fulfills which task and how these areas interact is still not known.

Most psychological and computational models focus on bottom-up processing since this part is better investigated and, for the computational systems, easier to realize. Some existing models which include top-down information weight the features with target-specific weights, some influence only the feature dimensions, and some influence the processing not until the saliency map that means they consider those regions that are salient in this map and are also task-relevant. The latter approaches are far from the biological analogue since in the brain top-down cues influence all parts of the processing down to early feature computations. A detailed, well evaluated system including top-down cues does not exist currently.

2.5 Discussion

In this chapter, we have reviewed the background that is important in the field of visual attention. We introduced several notations that are relevant in this field, for example, the distinction of *overt* and *covert attention* as well as *bottom-up* and *top-down influences*. The psychophysical paradigm of visual search was introduced and explained in detail. We furthermore sketched the processing flow of visual information in the human brain and discussed which processes and brain areas are involved in the attentional mechanisms. Then, we have introduced several psychophysical models of visual attention, ahead the Feature Integration Theory by Treisman and the Guided Search model by Wolfe. Finally, we related these topics by discussing which biological processes correlate to which mechanisms in current attention models.

This chapter shows that the research on visual attention is a highly interdisciplinary field. The different disciplines attack the problem from different sides: the psychologists regard the brain as a black box. In various experiments, they investigate human behavior on different tasks and try to conclude from the outcome of the experiments on the content of the black box.

The result are usually psychophysical theories or models. The neuro-biologists instead take a view directly into the brain. With new techniques like functional Magnetic Resonance Imaging (fMRI) it is visualized which brain areas are active under certain conditions. Again another practice is pursued by the computer scientists: they usually take over what they consider useful from psychological and biological findings and combine this with technical methods to build improved systems for computer vision or robotics applications.

In the last years, the different disciplines have highly profited 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 start to implement their models computationally to verify if the behavior of the systems on example scenes equals human perception. These findings help to improve the understanding of the mechanisms and eventually lead to improved attention systems. The further the theory on visual attention proceeds, the better get also the computational systems and the more useful they are in applications in computer vision and robotics.