CHAPTER 2
BACKGROUNDS OF VISUAL ATTENTION

2.1 Introduction

Nature is fond of arranging things in twos, and the visual system is no exception: starting from two eyes, the world is split into two (left and right) visual half-fields, information from each being sent to two separate half-brains. Millions of years of evolution in the presence of a ubiquitous visual horizon has taught the brain to further split the world into upper and lower visual half-fields (and to pay extra attention to the lower one). From the retina, visual information also begins a divergence into two parallel processing streams, one (roughly) for motion and spatial judgments, and one for shape and colour; these are variously named as the where/what, or magno/parvo, or action/perception, or dorsal/ventral pathways [122]. These parallel streams provide the top-level structural organization of this thesis, which contains two main parts: the first addresses visual attention, how do we know where to look? and the second addresses visual object categorization, or how do we know what we're looking at?

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 life. Nevertheless 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, first a description of attention and the concepts important in this field is given in section 2.2. Next in section 2.3, an introduction is given to several psychophysical models of visual attention that form the basis for many current computer models of attention. In section 2.4 correlations between neurobiology and the current attention systems in psychology and computer science is dealt, followed by a summary in section 2.5.
2.2 Concepts of Visual Attention

Several concepts of visual attention are discussed in this section. First a definition for the term ‘attention’ followed by the concepts of ‘overt’ versus ‘covert’ attention, ‘bottom-up’ versus ‘top-down’ attention and finally on ‘visual search’, its efficiency, ‘pop-out’ effects and ‘search asymmetries’.

2.2.1 What is Attention?

The concept of selective attention refers to a fact that was already mentioned by Aristoteles: “It is not possible to perceive two things in one and the same indivisible time”. Although the impression to retain a rich representation of the visual world and those large changes to the environment will attract one’s attention, various experiments reveal that the ability to detect changes is usually highly over estimated. 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. The other regions than the attended one are usually ignored. For example, in experiments on ‘change blindness’ [105] [100], it is proved that significant changes in a scene remains unnoticed, that means the observer is “blind” for the significant change. One convincing experiment on this topic is described in [106], an experimenter approaches a pedestrian to ask for directions. During their conversation, two people carrying a door pass/go between the experimenter and the pedestrian and during that interruption; the first experimenter is replaced by the second experimenter. Even though subjects engaged in an interaction with both the first and 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 [18] “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, for the purpose 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 [86]. 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 to the **early selection, late selection** debate which is discussed in [95].

2.2.2 **Covert versus Overt Attention**

Usually, directing the focus of attention to a region of interest is associated with eye movements referred as **overt attention**. However, this is only half of the truth. As early as in 1890, William James speculated that human visual system can attend to peripheral locations of interest without moving the eyes is referred to as **covert attention**. There is evidence that simple manipulation tasks can be performed without overt attention [55]. On the other hand, there are cases in which an eye movement is not preceded by covert attention. Findlay and Gilchrist [27] found that in tasks like reading and complex object searches, 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. The covert and overt attention is not independent as shown by Deubel and Schneider [23] as 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 [97] proposes several methods to analyze covert attention. Psychological investigations include the measuring of the reaction time to detect a target. Neuro-biological methods make measurement of the evoked potential amplitude or of changes in firing rates of single cells.

2.2.3 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. Bottom-up factors are derived solely from the conspicuousness of regions in a visual scene, for example by strong contrasts. Besides bottom-up attention, this attentional mechanism is also called exogenous, automatic, and reflexive or peripherally cued [25].

On the other hand, top-down attention is driven by the “mental state” of the subject, which comprises of information from ‘higher” brain areas such as knowledge, expectations and current goals [19]. That means, if looking for a stop sign post, all red circular regions will be attracted more easily than other regions. Only parts of top-down processing are investigated 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 are 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 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”) [124] [134]. 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.
Figure 2.1: Attentional Capture is demonstrated in [105], the image depicts that the red sheep captures the attention among the white sheep’s.

As per Theeuwas [110], 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 a emergency bell, one tends to attend that irrespective of whatever other job he is involved. This effect is called attentional capture as shown in Figure 2.1 where the red sheep captures the attention irrespective of all the background information. Bottom-up attention is much better investigated in the literature than the top-down influences. One reason is that the data-driven stimuli are easy to control than the mental state that includes knowledge and expectations. Even much less is known on the interaction of both processes.

2.2.4 Visual Search and Pop-Out

An important tool in research on visual attention is visual search [86] [132] [136]. 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? 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. The computational complexity of visual search has been investigated in [115] [117] and proved that an unbounded visual search (no target is given or it cannot be used to optimize search) is a Nondeterministic Polynomial complete problem. 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 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). The searches vary in their efficiency: the flatter the slope of the function, the more efficient the search. Two extreme cases 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. In case of serial search the reaction time increases with number of distractors. Already in the 11th century AD, Ibn Al-Haytham found that “some of the particular properties of visible objects appear at the moment when sight glances at the object occur, 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. Scenes with pop-outs are sometimes also referred to as odd-man-out scenes; one example is the well known red sheep in a white herd.

Parallel search is often but not always accompanied by pop-out [135]. Usually, pop-out effects only occur when the distractors are homogenous, 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. The strict separability of serial and parallel search is doubted nowadays [132]. Experiments by Wolfe indicate that the increase in reaction time seems to be continuous.
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, and for the face of another race among the faces of the same race as the test subject. An interesting experiment was subjects were told to search for the “zero” and were relatively 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 also influences visual search. Interesting is also that the search for a novel element among familiar ones is parallel [128]. This is an important effect that helps humans to ignore known things and focus processing on the new, most informative, sensory data.

Figure 2.2: Search asymmetries shown here depicts searching a tilted line among vertical distractors is easier than a vertical line among tilted line.

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 homogenous [65]. 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 [136], few among them are colors, intensity (different levels of contrast), line curvature, orientation (line tilt) or misalignment, terminators, closure, direction of movement, stereoscopic disparity (depth) and orientation quantitative values like length and number of proximity. Several findings indicate that basic features
may also be learned. For example, Neisser [86] mentions that finding special letters in a text is much more difficult for young children and illiterates than for people able to read.

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 A's. An example is that finding a tilted line among vertical distractors is easier than vice versa as shown in Figure 2.2. An explanation is proposed in [114] where 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 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, left and right diagonal, and for luminance there exist separate detectors for darker and lighter contrasts [112]. 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 others.

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 understand human perception. Here, two important psychophysical models are discussed. One is Feature Integration Theory of Treisman (in section 2.3.1) and the second one is the Guided Search Model of Wolfe (in section 2.3.2). In section 2.3.3 several other models are discussed.

2.3.1 Treisman’s Feature Integration Theory

The Feature Integration Theory (FIT) of Treisman [112] 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 requires focused attention”. Information from the resulting feature maps named as
topographical maps that highlight saliencies according to the respective feature is
collected in a master map of location. This map specifies where things are in the
display but not what they are. Scanning serially through this map focuses the
attention on the selected scene regions and provides this data for higher level
perception tasks. In FIT, a target is detected easily, fast and in parallel if it differs
from the distractors in exactly one feature and the distractors are homogenous. If it
differs in more than one feature, focal attention is required resulting in serial search.
The information about the target object called object file influences the search task
by inhibiting the feature maps.

2.3.2 Wolfe’s Guides Search

The basis goal of this model is to explain and predict the results of visual
search experiments. There have been many versions of this model [135-137]. The
architecture shares many concepts with FIT. The concept differs by considering only
one map for each feature dimension (like color, orientation). Within each map,
different feature types are represented (like red, green). In case of FIT it is saliency
map, in Wolfe’s it is activation map. The fusion of the feature maps is done by
summing up. For each feature along with bottom-up map a top-down map is also
considered. This leads to failure of identification of target when there are different
feature types.

2.3.3 Additional Models

There is wide variety of psychophysical models of visual attention. In zoom
lens model, the scene is investigated by a spotlight with varying lens. In
connectionist models [116], there are large numbers of processing units connected
by inhibitory or excitatory links based on neural networks. In CODE (COntour
DEtector) theory of Visual Attention (CTVA), it integrates contour detector for
perceptual grouping with the theory of visual attention. In these models, a scene is
processed in parallel and the one which finishes faster is the one selected and is
named as race model. Another interesting model a triadic architecture [108] consists
of three parts: first, a low-level vision produces proto-objects rapidly and in parallel.
Second, a limited capacity attentional system forms structures in stable object
representations. Finally, a non-attentional system provides setting information which
influences the selection.

### 2.4 Biology To Computational Model Correlations

In this section, the extent to which the concepts usually used by psychological and computational models of attention are supported by neuro-biological evidence are discussed. Some of these (computational) concepts will be introduced later in the next chapter but its neuro-biological evidence for the formation of such mechanism is discussed here.

#### 2.4.1 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 implemented by most models. However, the distinction into several pathways for color, 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 [136], the biological pathways are argued to be limited to three or four [94]. Interestingly, three of the suggested neural pathways [Appendix A.1.6] usually coincide each with one psychological feature channel, namely motion, color, and depth, whereas there are several psychological feature channels for processing. Since, newer findings suggest that there is no separate pathway but many cells are responsible for edge detection as well as for color processing, the psychological feature channels seem to correspond to several brain areas.

#### 2.4.2 Center-Surround Mechanisms

The centre-surround mechanisms that are used in most computational models
of attention and in several psychological ones determine the feature contrast regarding intensity or colour. These have their neuro-physiological correlation on many different places in the brain. Later, cells in the brain continue in responding to contrasts with these mechanisms.

2.4.3 Color Perception

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 opponents red-green and blue-yellow.

Psychological models often use a three-color or double opponency approach. 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. People generally use directly the RGB input. However, some convert the image to a different color space and use in their model. Many systems consider the red-green and blue-yellow opponency component also.

2.4.4 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 \cite{32,53,82} believed in such a saliency map, whereas others declined this view. Recently \cite{85}, 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. 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. Hence consideration of saliency map as a combination of neurons will enhance the behavioral nature of the computational model rather than
considering it as a map with same behavior.

2.4.5 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. In [25], 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 the superior parietal lobule (SPL), the frontal eye fields (FEF), and the supplementary eye field (SEF). The areas in the inferior parietal lobule (IPL), the lateral prefrontal cortex in the region of the middle frontal gyrus (MFG), and the anterior cingulated cortex [Appendix A.5 –A.7] are used less consistently. 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 weigh the features with target-specific weights. Some influence only the feature dimensions. Some influence the processing not until the saliency map. These 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. Consideration of semantic cues can influence the processing of the region under investigations, which has greater influence in reducing the computational tasks involved in the model.

2.5 Summary

In this chapter, the background information that is important in the field of visual attention is reviewed. Here, several concepts that are relevant in the field has been introduced, such as 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. Then, several psychological models of visual attention, starting from the Feature Integration Theory by Treisman and Guided Search model by Wolfe are discussed. Finally, a discussion of how these mechanisms of current attention models correlate with the biological process is given.

The discussed literature in this chapter shows that the research on visual attention is a highly interdisciplinary field. The different disciplines attack the problem from different perspectives, for example, psychologists regard the brain as a black box and experiments are desired to investigate. 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 results 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 techniques to build improved systems for computer vision related applications.

In the past, 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 have started to implement their models computationally to verify if the behavior of the system on example scenes equals human perception. These findings help to improve the understanding of the mechanisms and eventually lead to improved attention systems. As the theories on visual attention proceeds further, the better the computational systems and its applications in computer vision and robotics.