Visual attention and active vision

Kootstra, Geert Willem

Part I

Natural Systems

Visual Attention in Natural Systems

The human brain is limited in the amount of information that it can process. To deal with this problem, attention plays an important role (Tsotsos, 1997). By selecting specific parts of the sensory input for processing, the system deals with its limitations. This is especially true for the visual modality. Visual attention is therefore an important element in visual processing. By actively focussing attention on a specific part of the visual field, uninteresting information can be disregarded, while interesting information is further investigated. In this way, the processing capacity of the brain is efficiently deployed (Findlay & Gilchrist, 2003).

This chapter gives an overview of a number of important aspects of visual attention. In Section 2.1, the dichotomy of overt and covert visual attention is explained. Section 2.2 discusses the control of eye movements and the bottom-up and top-down influences. Section 2.3 deals with visual search and basic features that lead to pop outs. The study of visual search has been an inspiration for many visual-attention models, which are discussed in Section 2.4. The bottom-up components of most of these models are based on the contrast of basic features. Section 2.5, however, shows that configural properties, which can emerge from the constellation of basic features, play an important role in visual attention as well. This shows that visual attention is essentially object oriented. The chapter ends by proposing the use of *symmetry* as a configural property to predict bottom-up visual attention. This proposal is motivated in Section 2.6. The concept of symmetry will be used as an important feature to select points of interest in the visual field in the rest of this dissertation.

2.1 Overt and Covert Visual Attention

First of all, it is important to make a difference between *overt* and *covert* visual attention. The most commonly used interpretation of visual attention is the movement of the eyes to attend to a specific location in the visual field. This is called *overt visual attention*. Overt visual attention incorporates all visual attention that involves body movements. One can overtly attend to something by making eye, head, and/or body movements. By a continuous sequence of saccades and fixations, the visual field is inspected. A saccade is a rapid change of the orientation of attention, and a fixations is a short period of stable orientation. *Covert visual attention* on the other hand, is the mental focus on a particular part of the visual field. If you keep your eyes focused on this book, you can mentally focus attention towards the cup of coffee that might be in

the right part of the visual field without making an eye movement. However, this feels a bit awkward and unnatural.

Although visual attention is undeniably possible without making any eye movement, visual attention is only measured in the presented work by recording eye movements. That is, we focus solely on overt visual attention. Whenever the term visual attention is used in this dissertation, it refers to overt visual attention. The saliency methods discussed in this chapter and presented in Chapter 3 make a prediction of locations in an image that are likely to be attended by an eye fixation. Although the performance of the models is measured by comparing the prediction with human eye fixations, it is likely that the same predictions would hold for covert visual attention. It is hypothesized that covert visual attention makes a quick scans of the visual field to find potentially interesting locations, which leads to the execution of eye movements to further investigate these locations (Findlay & Gilchrist, 2003). The representation of some sort of a saliency map seems to precede the deployment of attention, either overt or covert. Moreover it is argued that covert visual attention is overemphasized in psychophysical studies and rarely occurs outside of experimental setups when people can freely move their eyes (Findlay & Gilchrist, 2003).

2.2 Control of Eye Movements

Human visual attention is controlled top down as well as bottom up. Top-down or endogenous influences are driven by internal information that is not present in the stimulus, such as the task, prior experience, knowledge, and interests. These influences are personal and differ greatly among individuals. Bottom-up or exogenous control, on the other hand, is driven by information in the stimulus. Some properties of the stimulus attract attention without top-down knowledge. Because the stimulus is the driving force, the bottom-up influences are more universal and differ less among individuals. Where the bottom-up influences are the result of early visual processes, the top-down influences involve higher-order processes including memory processes. Although in the literature, the relative role of the influences is debated, the general consensus is that both play a role in the guidance of eye movements. We give a short overview of top-down and bottom-up control and neural correlates involved in the control of eye movements.

2.2.1 Top-Down and Bottom-Up Control of Eye Movements

Yarbus (1967) was one of the first to show that the task has a strong influence on eye movements. Depending on the instructions given to the participants before the experiment, the eye-movement patterns when viewing a painting differed greatly. Tsotsos (1990) argued that the attentional mechanism exploits knowledge of the specific problem that needs to be solved to constrain search. Supporting this argument, Rothkopf, Ballard, & Hayhoe (2007) showed that participants occupied with a task did not pay attention to salient objects that were irrelevant to the task. However, when the task was finished, the salient objects did attract attention. A very striking example of this phenomenon was demonstrated by Simons & Chabris (1999). While occupied with the task to count the number of ball passes in a basketball video fragment, participants completely failed to notice the highly salient stimulus of a man in a gorilla suit entering the scene. Henderson, Brockmole, Castelhano, & Mack (2007) showed that eye movements during a visual-search task could not be predicted on the basis of bottom-up information only. Task knowledge has been demonstrated to influence early visual processing in the human brain. Area V1, for instance, shows increased activation during a contrast-discrimination task (Huk & Heeger, 2000).

Context has been shown to have an influence on visual attention as well. An object that is taken out of its normal environment and displayed in an unusual environment attracts much more attention than it normally does (De Graef, Christiaens, & d'Ydewalle, 1990). Scene context also provides a top-down bias on the search for a target (Neider & Zelinsky, 2006). Chun & Jiang (1998) showed that humans build a memory for visual context that guides visual attention to find a target in a search display. Furthermore, depending on the context, humans fixate on different objects in the environment (Rothkopf et al., 2007).

Long-term motor memory also influences eye movements. Noton & Stark (1971a,b) proposed the scanpath theory, stating that a fixed fixation pattern is elicited based on a visual representation of the observed object. Additionally, it has been demonstrated that a spatial memory of the scene is built up while viewing and that this representation is used to guide eye movements (Henderson & Castelhano, 2005; Karn & Hayhoe, 2000). Similarly, the nature of eye fixations changes when there is an abrupt change in a dynamic scene, which is thought to be caused by the spatial memory that is disrupted (Carmi & Itti, 2006a).

Finally, semantic information can influence eye movements. Objects that are semantically related to the task or to other objects of interest, are more likely to attract attention. This semantic priming facilitates the attention to relevant objects (Odekar, Hallowell, Kruse, Moates, & Lee, 2009).

On the other hand, there is also evidence for the influence of bottom-up guidance. Theeuwes (1991, 1994), for instance, showed that attention was captured by a salient distractor in a search task. Even after extended practice, the irrelevant stimulus influenced the eye movements, and complete top-down guidance was ruled out (Theeuwes, 1992). Also for more complex photographic stimuli, overt attention is attracted towards contrast-manipulated parts of the images (Einhäuser, Rutishauser, Frady, Nadler, Köning, & Koch, 2006). Since the contrast enhancement did not change the meaning of the stimulus, this must be a bottom-up effect on attention. Mannan, Ruddock, & Wooding (1995), concluded that eye movements made during brief presentation of photographic images are a reaction to the spatial features of the image and not to the content of the image.

Eye movements are thus controlled by both top-down and bottom-up influences. In experiments by Van Zoest & Donk (2004); Van Zoest, Donk, & Theeuwes (2004), evidence for both mechanisms is found. The fast eye movements were stimulus driven whereas the slower eye movements were goal driven. Whereas eye movements were biased towards the contrast enhanced parts of the image in a free-viewing condition, Einhäuser, Rutishauser, & Koch (2008) showed the eye movements are strongly goal-driven when a task was given to the participants. According to Treue (2003), visual attention is a result of the combination of bottom-up stimulus features and top-down attentional modulation, in order to favor potentially relevant information. Wolfe, Butcher, Lee, & Hyle (2003) showed that both bottom-up and top-down guidance is present in visual search. They showed that the top-down guidance can be based on information about the task, and on priming by preceding targets.

Although it is clear that both influences play a role, the focus of the dissertation is on the bottom-up influences. Mainly the role of the stimulus in the guidance of eye movements is studied, specifically the visual features that can be used to predict human eye fixations. This gives insight in the inherent properties of the stimulus that attract attention.

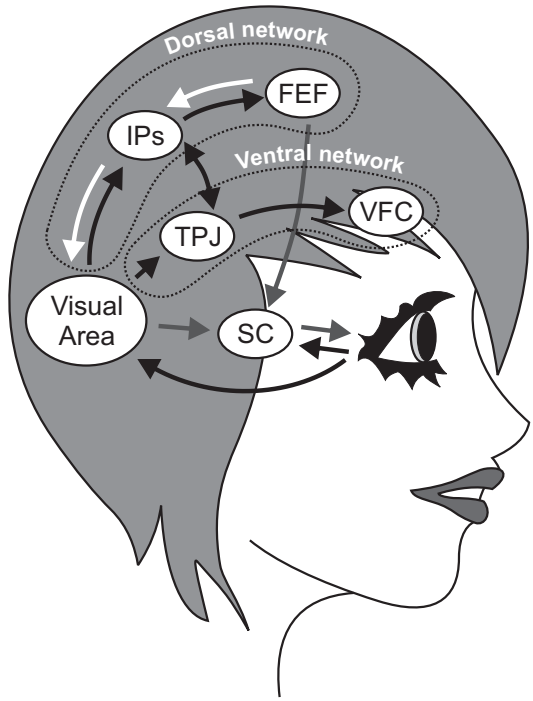

Figure 2.1: Top-down and bottom-up control of visual attention and eye movements in the human brain. The superior colliculus (SC), located in the midbrain, plays a central role. This brain area receives input directly from the retina, and from other brain areas, most notably the visual areas in the visual cortex and the frontal eye fields (FEF) in the premotor cortex. The SC projects down to areas in the midbrain and brainstem, where the retinotopic representation of a target is transformed to motor commands. From the visual area, there are two distinct networks, the dorsal frontoparietal network, which is involved in top-down control, and the ventral frontoparietal network, which is involved in bottom-up control. The main areas involved in controlling eye movements in the ventral network are along the intraparietal sulcus (IPs) and FEF. The dorsal network is only present in the right hemisphere, and consists of the temporoparietal junction (TPJ) and the ventral frontal cortex (VFC). Bottom-up and top-down integration takes place by interactions between the two networks (Corbetta & Shulman, 2002). Note that the figure gives an extremely simplified, conceptual view.

2.2.2 Neural Correlates of Eye-Movement Control

Figure 2.1 gives the most important areas in the brain that are involved in visual attention and eye movements according to Corbetta & Shulman (2002). Central in the system is the superior colliculus (SC), which is located in the midbrain. The SC receives input from many brain areas, but most notably from the visual area and from the frontal eye fields (FEF). Furthermore, it receives input directly from the retina. In the SC, the target of an eye movement is retinotopically represented. The area projects its output to the ocular-motor pathway in the midbrain and brainstem, where the retinotopic representation is transformed into motor commands.

A fast, reflexive control mechanism is thought to consist of a pathway from the retina via the SC to the ocular-motor pathway or alternatively from the retina via the visual area to the SC to the ocular-motor pathway. This control mechanism is thought to be involved in *smooth pursuit* and the *optokinetic reflex*, to keep an object in the focus of attention.

Two distinct pathways are involved in *shifting* the focus of visual attention, the *dorsal frontoparietal network* and the *ventral frontoparietal network* (Corbetta & Shulman, 2002). The dorsal frontoparietal network is concerned with top-down control. It mainly consists of the area along the intraparietal sulcus (IPs) and of the FEF. The network plays a role in spatial and featural selection depending on contextual or task-related information. The FEF play a role during task switching. From single-cell recordings in monkeys, it seems that different areas along the IPs are specialized in specific features. The dorsal network furthermore sends top-down control signals to the visual cortex, which modulates the visual processing depending on task and context. In general, the network links relevant sensory representations to relevant motor actions, which are sent down to the SC via the FEF.

The ventral frontoparietal network is involved in bottom-up control. This network is activated when there are unexpected or salient stimuli. The network is mainly located in the right hemisphere, and consists of the temporoparietal junction (TPJ) and the ventral frontal cortex (VFC).

There are interactions between the two networks. Top-down processing is influenced by the saliency of the stimuli and bottom-up processes are modulated by contextual and task-related knowledge. According to Corbetta & Shulman (2002), the ventral network interupts ongoing cognitive processes of the dorsal network and reorients attention to

the spatial locations of salient stimuli when unexpected and salient stimuli are present.

2.3 Visual Search

This section discusses insights about human visual attention from visual-search experiments. A set of basic features are discussed that result in a pop-out effect. This effect is a clear demonstration of the bottom-up components of human visual attention. It has led to models of visual search, which have been an inspiration for the visual-attention models that will be discussed in Section 2.4. In later sections, we argue that basic features are not the only features of importance for human visual attention.

Figure 2.2 shows examples of displays in which participants have to find the *odd-one-out*, that is, a single item that differs from the other items. The pop-out effect in *feature search* is illustrated in Figure 2.2a. The three displays contain a *singleton pop-out* meaning that search for the target, the odd-one, is very efficient. In every display, there is one object that differs from the others in a single unique feature, either intensity, color, or orientation. The reaction times for detecting the target are very little affected by the set size (Treisman & Gelade, 1980), that is, they are unaffected by the number of distracting objects in the display. Reaction times for large set sizes are similar to those for smaller set sizes. Figure 2.2b can be used to experience this. The pop-out is detected quickly and without effort. Only one item can be the object of overt or covert visual attention, and if attention would be needed to detect the target, the reaction times would depend on the set size. Since this is not the case, it suggests that the processing of the visual information in feature search is done in parallel and preattentively (Treisman, 1985).

However, in *conjunctive search*, the target is not defined by a single feature, but by a conjunction of two or more features (see Figure 2.2c). The tilted red bar is the target among distractors which are either vertical red bars or tilted blue bars. Neither the single feature color nor the feature orientation uniquely defines the target. In conjunction search, the target does not pop out and search is inefficient. Here, the reaction times do depend on the set size. This shows that the information processing in conjunction search is serial and needs visual attention to integrate the features (Treisman & Gelade, 1980).

Features that result in a pop out in a single-feature search task are so called *basic*

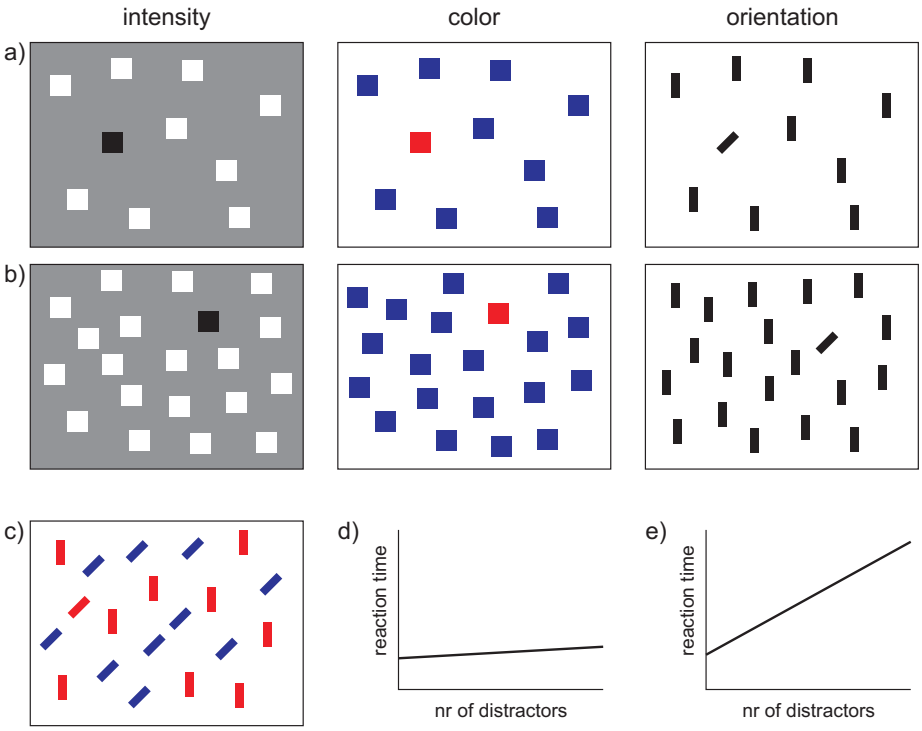

Figure 2.2: The pop-out effect. a) Three examples of feature-search display where the target is clearly visible, because it differs in one unique feature. As can be appreciated by looking at (b), human reaction times in finding the target are hardly influenced by the number of distractors, that is, the slope of the reaction time x set size curve is near zero (see also the fictitious graph d)). Search for this pop-out is efficient. In conjunction search (c), the target does not differ from the distractors in one unique feature, but has rather a unique combination of two features (here: tilted and red). In contrast to feature search, the reaction times in conjunction search do depend on the number of distractors, and search is inefficient (e).

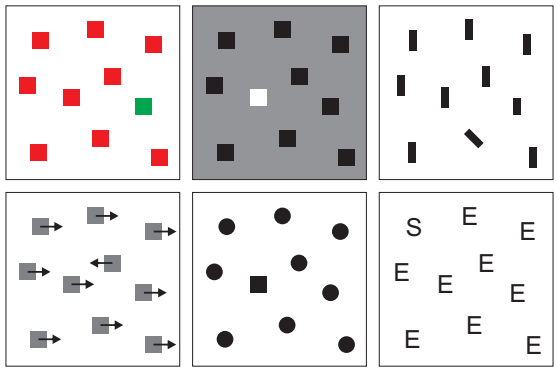

Figure 2.3: A list of basic features. From left to right, top to bottom: color, intensity, orientation, motion, and two examples of shape. The pop out of the target depends on the feature difference at the center and at its surround.

features. The contrast-saliency model of Itti, Koch, & Niebur (1998), built upon the Feature-Integration Theory (Treisman & Gelade, 1980) utilizes tree basic features which are known to result in a pop-out: intensity, color, and orientation. However, there are more singletons known. We discuss some of these basic features.

2.3.1 Basic Features

A component that draws attention is one that has some basic feature properties that are sufficiently different from its surrounding components. The strength of the bottom-up attraction is based on the magnitude of the difference between features in the center and the surround of the receptive field. Wolfe (1998, 1994) lists a number of these basic features:

Color Targets with a color that is sufficiently different from the distractors are efficiently found (e.g., Nagy & Sanchez, 1990; Farmer & Taylor, 1980). When the distractors are heterogeneous, search is efficient only when the target color is linearly separable from the distractor colors (Bauer, Jolicoeur, & Cowan, 1996; D'Zmura, 1991) in chromatic color space. If the distractors have a variety of different colors, search can be inefficient. The target is more easily found when it

contrasts with its local neighborhood (Nothdurft, 1991). There is a search asymmetries. Search for a non-prototypical color among prototypical colors is easier than the reverse, that is, it is easier to find magenta among red distractors, than to find red among magenta (Treisman & Gormican, 1988). Wolfe (1994) uses four basic colors: red, green, blue, and yellow. Such search asymmetries are found in many of the other features. It is usually easier to find an uncommon item among more common items.

Intensity Similarly to color, targets with a sufficiently different intensity or brightness pop out (Bauer et al., 1996). Also for intensity, targets are more easily found when the target is linearly separable from the distractors, that is, a middle gray is more easily found among darker distractors than among darker and brighter distractors.

Orientation Although humans can distinguish lines that differ by 1-2°, orientation pop-out is roughly present when the line differ by $>15°$ (Foster & Ward, 1991). Also in orientation, the local contrast of the item plays a role in the efficiency of visual search (Nothdurft, 1991). Surprisingly, a 50° target is more easily found among -10° distractors, than among -50° distractors, even though the angular difference is smaller (Wolfe & Friedman-Hill, 1992). This can be explained by distracting symmetries emerging from the configuration of the items.

Motion The detection of moving targets among stationary distractors is highly efficient (e.g., Dick, Ullman, & Sagi, 1987). The reverse is more difficult. Also targets whose speed and direction differ sufficiently from the local neighboring distractors pop out (Royden, Wolfe, Konstantinova, & Hildreth, 1996). Although the detection of a slow moving target among faster moving distractors is less efficient than finding a fast moving targets among slow moving items.

Shape In contrast to the other basic features, the dimensionality of shape is difficult to establish. Where the color space is two dimensional, intensity and orientation are one-dimensional, and motion is two-dimensional, the dimensionality of the shape space is unclear. In the literature, different definitions of shape are used. Theeuwes (1992) showed that humans can efficiently find squares among circles and vice versa. Similarly, Treisman & Gormican (1988) used curvature, and found that curved lines are efficiently found among straight lines. Julesz

(1984) used line terminations, and let participants search for 'S' (two terminators) among 'E's (three terminators). It must be noted that both stimuli also differ in angularity. Julesz also proposed intersections as basic features.

In pop-out experiments, saliency is always defined in terms of center-surround calculations or contrast of basic features. The features in the center are compared with the features in the surroundings. An item is salient, or conspicuous when one or more of its basic features are different from the features of neighboring items. This dissertation, however, proposes that there are other features that attract bottom-up attention as well. The results of Chapter 3 shows that good predictions of the locations of human eye fixations can also be made using symmetry. In the case of symmetry, the center is not compared to its surroundings, but the complete local pattern is analyzed. However, before discussing the role of symmetry in vision, a number of visual-attention models is discussed in the next section that utilizes the center-surround contrast of basic features.

2.3.2 Models of Visual Search

There are two influential models that explain and predict human behavior in visual-search experiments, the *Feature-Integration Theory* and the *Guided-Search model*. Although these models are not directly used in the work described in this thesis, they give interesting hypotheses of the mechanisms underlying visual search.

2.3.2.1 Feature-Integration Theory

According to the *Feature-Integration Theory* of Treisman & Gelade (1980), several basic visual features are processed in parallel by the visual system. The center-surround contrasts of the basic features are represented in separate feature maps. The feature maps are then integrated into an overall saliency map. An implementation of the theory, the saliency model of Itti et al. (1998), is presented in Appendix A. In their model, three basic features are utilized: intensity, color, and orientation.

The theory explains human behavior in feature and conjunction search. In feature search, the target contrasts with its surrounding, and is therefore represented as salient in one of the feature maps, and thus in the saliency map. Search for the target is then very efficient. In conjunctive search, on the other hand, the target is not uniquely

defined in one of the feature maps. The feature maps will therefore not contain a single salient location, but multiple less salient ones. After integrating the feature maps, the saliency map will contain multiple salient locations, making search for the target inefficient and dependent on the set size.

2.3.2.2 Guided-Search Model

The *Guided-Search* model of Wolfe (1994, 2007) seeks to explain and predict human behavior in visual-search experiments. The model consists of two stages. In the first stage, visual information from all locations in the visual field is processed in parallel. However, only limited visual information can be used in this stage, which is made up of the basic features. In the second stage, more complex processing can be performed, but limited to only one or a few locations at a time.

Feature representations of the stimulus are formed in the early stages of the model. In the version of the model presented in (Wolfe, 1994), the features color and orientation are used, but any of the basic features could be used. The stimulus is filtered through broadly-tuned filters. In the orientation domain, that means that the stimulus is represented by the orientation channels that represent how *steep*, *shallow*, *left*, and *right* the stimulus is at a given location. This orientation representation in the model better explains human psychophysical data that shows that 30° and -30° orientations are more similar than expected on basis of their angular difference (Wolfe & Friedman-Hill, 1992). Similarly, in the color domain, broadly-tuned channels for *red*, *yellow*, *green*, and *blue* are used.

Activations are calculated both bottom up and top down. The bottom-up activations are based on center-surround differences in the different feature channels. The activation of an item is calculated by comparing the item to the 5×5 array of neighbors surrounding the item, with near neighbors having a stronger influence than more distant ones. The differences are thresholded using a *preattentive just noticeable difference* threshold. This threshold is inspired by psychophysical data showing that small differences in color and orientation are not noticeable preattentively (Foster & Ward, 1991; Nagy & Sanchez, 1990). Next, the differences are multiplied by the strength of the response of the broadly-tuned channel to the central item. Thus resulting in a higher activation for a difference when the item itself is prototypical. This accounts for the visual-search asymmetries that have been discussed in Section 2.3.1. Finally, the bottom-up

activation of an item has a ceiling threshold.

The bottom-up part of the model accounts for odd-one-out experiments. However, if the task is to search a specific item, top-down knowledge needs to be incorporated. In order to do so, all items in the display are examined to determine which channel categories of the target are unique. These are assigned more weight. Next, the weighted response of each channel to the target is compared to its response to the distractors. The channel with the greatest positive difference per feature is selected. The response of the selected channels for all features are combined to create an activation map.

2.3.3 Basic Features Revisited

The earlier-mentioned basic features result in a pop-out effect, that is, they result in a flat curve of the reaction time as a function of the dislay size. There can be other, more complex features that do not result in a flat slope, but in an ascending curve with a steep slope. However, as pointed out by Townsend (1990), one cannot simply infer from a steep curve that processing is serial and attentive, since such a profile can also result from a limited-capacity parallel model that shares limited computational resources over the presented items. More items will then result in fewer resources per item, and thus in a slower reaction time (Wolfe, 1998). Moreover, it is difficult to define when a curve is flat and when it is steep, because there is a continuum of different slopes in the reported visual search studies. A hard separation between parallel and serial mechanisms in search can therefore not be made through experiments alone.

A continuum of search slopes is also found in feature search as a function of the difference between target and distractor. Whereas search for a green target among red distrators results in zero slope, search for a green target among yellowish green distractors results in a steep slope, although the involved mechanisms are unlikely to have changed (Nagy & Sanchez, 1990). Similarly, the slopes can become steep when the distractors are sufficiently heterogeneous (Bauer et al., 1996).

Also in conjunction search a variety of different reaction time $\times$ set size slopes are found. Some studies show that conjunction search does not always result in steep slopes, but can also result in flat slopes when the target is known and the features of the target are highly distinctive from that of the distractors (Theeuwes & Kooi, 1994; Wolfe, Cave, & Franzel, 1989; Wolfe, 1992). Also Treisman & Sato (1990) found that the slopes are flat for known targets, and steep for unknown targets. To account for

these findings, top-down components are integrated in the Feature-Integration Theory (Treisman & Gelade, 1980; Treisman & Sato, 1990) and the Guided-Search model (Wolfe, 2007, 1994; Wolfe et al., 1989).

This suggests that it is impossible to separate the mechanisms involved in serial and parallel processes. Wolfe (1998) therefore proposes to speak of *efficient* and *inefficient* processes. Search for a line of one orientation among distracting lines of sufficiently different orientation, for instance, is efficient, while search for an 'L' shape among 'T' shapes is inefficient (Egeth & Dagenbach, 1991).

2.4 Models of Visual Attention

In the previous section, two existing visual-search models are discussed. These models predict human reaction times in visual search. In this section, an overview of visual attention models is given that have been proposed in the literature to predict the locations of human eye fixations. Some of these models are purely stimulus-driven, while others combine bottom-up and top-down influences. Although this dissertation mainly deals with bottom-up models, a short overview of top-down models is given as well.

2.4.1 Saliency to Model Bottom-Up Visual Attention

Many predictive models of human visual attention are so-called *saliency models*. A saliency model is a computational model that determines the conspicuous parts of an image based on specific image features. The results of a saliency model, a saliency map, can be compared with actual human eye fixations to determine how well the model correlates with the human data. A good correlation is a strong suggestion that the used image features play a role in the guidance of human eye movements. Not only does it give more insight in visual processing in natural systems, but the model can also be used to predict overt visual attention and to improve computer and machine vision systems.

Most existing saliency models are inspired by the pop-out effects discussed in Section 2.3 and use feature contrasts to determine the saliency at different points in an image. The basic features of a certain location are compared with those of the local neighborhood. The location is determined to be salient when it contrasts with its

surroundings. The most influential saliency model based on contrast is the saliency model of Itti, Koch, & Niebur (1998), and will be referred to as the *contrast-saliency model* hereinafter. This model is based on the Feature-Integration Theory (see Section 2.3.1)and calculates the saliency at every point in an image using contrast in three different feature channels: intensity, color, and orientation. Since this model is compared to the saliency model proposed in Chapter 3, which is based on symmetry, the contrast-saliency model is described in detail in Appendix A. The contrast-saliency model has been found to predict human behavior correctly in feature and conjunction search (Itti & Koch, 2000). Moreover, it predicts human eye fixations well above change levels (Parkhurst, Law, & Niebur, 2002; Ouerhani, von Wartburg, Hügli, & Müri, 2004; Kootstra & Schomaker, submitted). Later versions of the model also include contrast in dynamic features like *flicker* and *motion* (Itti, Dhavale, & Pighin, 2003). These features are found to be good predictors of human gaze in video clips (Carmi & Itti, 2006b)

Other models of visual attention are also based on contrast. The Guided-Search model of Wolfe (1994), mentioned in Section 2.3.1, for instance, is based on feature contrasts using center-surround differences. The model explains and predicts many results in visual-search experiments. Unfortunately, in contrast with the model of Itti et al. (1998), it cannot predict human eye fixations on complex photographic images, but only on search displays containing simple synthetic stimuli.

Other models for the prediction of fixations on complex photographic images are largely based on contrast as well. The saliency model of Le Meur, Le Callet, Barba, Thoreau, & Francois (2004) for instance is based on different types of contrast calculations all throughout the model. They found correlations of the model's prediction with human data that are slightly higher than using the model of Itti et al. (Le Meur, Le Callet, Barba, & Thoreau, 2006). Parkhurst & Niebur (2004) use contrast in texture to predict human gaze. Center-surround differences in the distribution of features are used by Bruce & Tsotsos (2009) to model human visual search. A similar approach is taken by Gao, Mahadevan, & Vasconcelos (2008), who compared histograms of filter responses at the center and at the surround. Privitera & Stark (2000, 1998) tested a number of simpler contrast-saliency operators, like center-surround operators in intensity, orientation and the Michaelson contrast, and found these operators to predict human fixations to some extent. However, the performance of the operators fluctuated over the different images and could not account for fixation sequences (Stark & Privitera, 1997).

Privitera & Stark (2000) also tested a symmetry operator. Although their symmetry operator was relatively simple, it could predict human gaze to some extent. This strengthened us in the idea that there is more to bottom-up visual attention then center-surround contrast and that symmetry might play a role in visual attention as well. We elaborate on this in Section 2.6.

2.4.2 Top-Down Models of Visual Attention

In this thesis, we focus on bottom-up models of visual attention. We are mainly interested in the feature aspects of the stimulus that attract visual attention. As discussed earlier, human visual attention clearly has a top-down component as well. A number of *hybrid saliency models* exist that incorporate bottom-up and top-down control. For most of these models, the bottom-up processes calculate a saliency map based on features in the image, usually very similar to the contrast-saliency model (Itti et al., 1998). The top-down processes are usually modeled either as a target-specific feature channel added to the saliency model to incorporate goal-directed control, or as an a posteriori modulation of the bottom-up saliency map to bias the saliency map towards specific locations based on top-down knowledge.

The VOCUS model of Frintrop (2006) models top-down influences by adding a separate goal-directed map to the contrast-saliency model (Itti et al., 1998). In this map, the similarity of the search target with the image is given for different locations at different scales. In a similar fashion, Zelinsky, Zhang, Yu, Chen, & Samaras (2006) incorporated top-down target search in the contrast-saliency model. Target guidance was modeled by a coarse-to-fine comparison of the target features to the stimulus features (Rao, Zelinsky, Hayhoe, & Ballard, 2002). The attentional model of Schill, Umkehrer, Beinlich, Krieger, & Zetzsche (2001) plans its next fixation at the location that is expected to maximize the information gain about the scene or object under observation. The information gain is estimated based on previously learned knowledge about scenes and the current representation of the observation. Thus, information is gathered that is most informative for disambiguating the stimulus.

Top-down modulation of bottom-up activations is used in the model of Navalpakkam & Itti (2006a,b, 2005) to increase the attention to specific image features. They furthermore used object models to facilitate the top-down attention to objects. Also the Guided Search model uses top-down information about the target to select specific

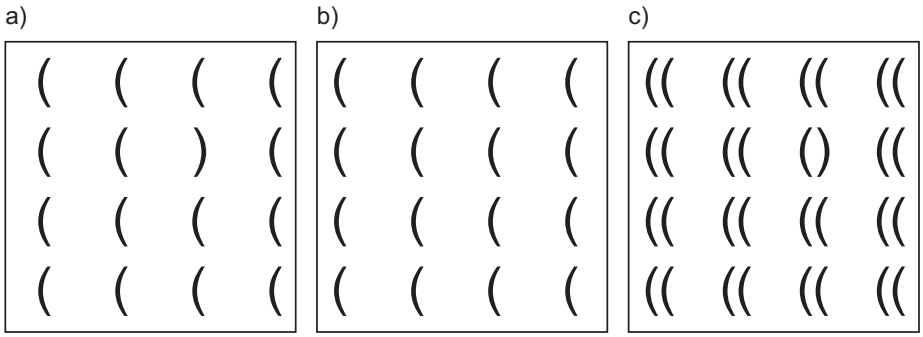

Figure 2.4: Configural superiority. Search for the rightward curved line segment among 15 distractors in (a) takes longer than among 31 distrators in (c). Although one would intuitively think that the addition of the 16 identical items in (b) would lead to less efficient search, it actually leads to more efficient search. This can be explained by the fact that the addition of the item leads to the preception of form, and the addition of the configural features symmetry and closure. The configural features are superior over the basic feature curvature. (Adapted with modifications from Pomerantz, 2006)

image features for bottom-up processing (Wolfe, 2007, 1994). Torralba, Oliva, Castelhano, & Henderson (2006) learned the most likely locations to find certain objects based on a database with labeled images. This knowledge is applied to bias the saliency maps to these locations and thereby decrease search times.

2.5 Beyond Basic Features: Configural Features

An interesting finding in visual-search studies is that the reaction times can also decrease as a function of the set size. An example of this is shown in Figure 2.4 (Pomerantz, 2006). If the left display (a) is shown, humans are relatively slow in detecting the target among the 16 items. However, if another 16 identical items, shown in (b), are added, visual search is suddenly more effective (c). Humans are much faster in finding the target in (c) than in (a), despite the increase in set size. Instead of the addition of the identical items leading to less efficient search due to crowding or masking effects, it leads to more efficient search. This looks counterintuitive, since the added items are

all identical and therefore carry no information. However, the reason for this is that the addition of (b) results in the perception of *configurations*, constellations of basic features. The display in (c) is perceived by humans as containing 16 figures instead of 32 basic features. Moreover, the addition of items leads to *emerging features*. Where (a) only contains the basic feature *curvature*, (c) also contains the *configural features symmetry* and *closure* (these features are explained in Section 5.2). The pop-out effect of the features symmetry and closure is stronger than that of curvature. There is a *configural-superiority effect* (Pomerantz, Sager, & Stoever, 1977). The figures are stronger attractors than basic features.

The example in Figure 2.4 shows that there is more to bottom-up visual attention than the pop out of basic features. Although previous sections demonstrated that visual attention is sometimes feature oriented, these results show that visual attention can also be strongly object oriented. Configural features like symmetry and closure can be stronger visual attractors than the contrasts of basic features. The presence of configural features in an image strongly suggest that a figure or object is present at that location. In order to predict the attention to objects, these higher-order features need to be detected. We motivate in the next section that the detection of local symmetry is a good candidate for predicting object-based visual attention.

2.6 Symmetry in Vision

Most research on visual attention has dealt with either low-level basic features, or with high-level top-down control. However, Wang, Kristjansson, & Nakayama (2005) showed that visual processes on an intermediate level of visual analysis can also account for visual search. They demonstated that processes related to perceptual organization play a role in visual attention as well. These processes account for configural features such as symmetry and closure. The influences of intermediate processes involved in perceptual organization are underexplored (Wang et al., 2005). This thesis aims to fill this gap in the current literature on visual attention. This dissertation focuses on the role of symmetry in visual attention. A saliency model based on local symmetry is proposed for the prediction of human eye movements in Chapter 3.

Although contrast has been the dominant feature for saliency models, a clear deficiency in current contrast-based saliency models is illustrated in Figure 2.5. The figure shows examples of images containing symmetrical objects that were used in the eye-tracking

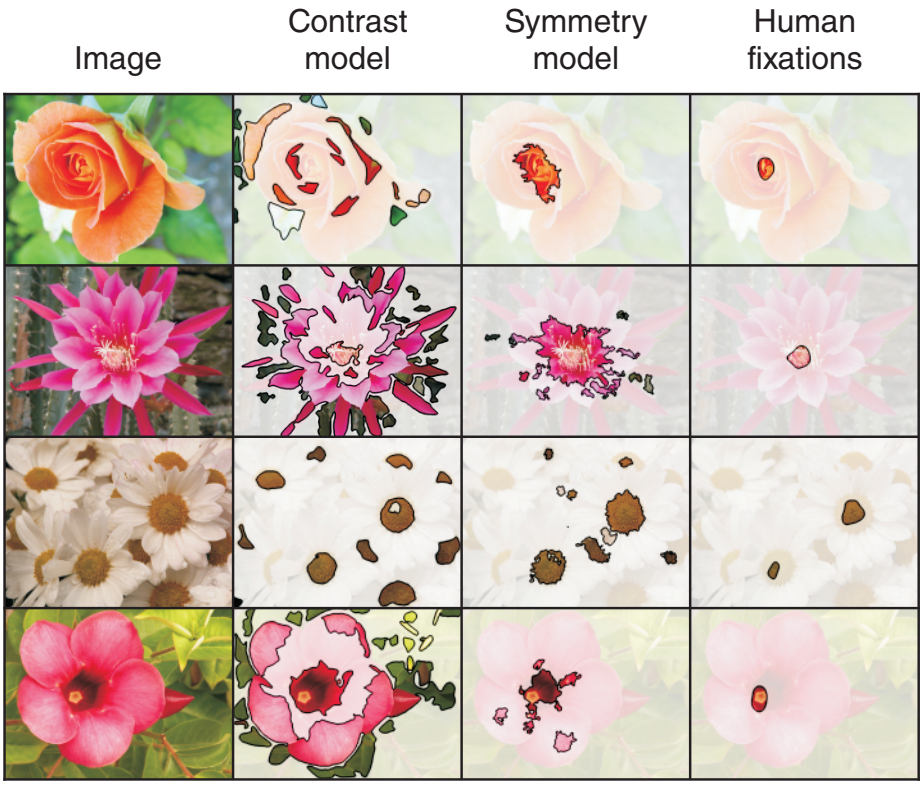

Figure 2.5: Examples of images containing symmetrical objects. The second column shows the contrast-saliency maps, the third column gives the symmetry-saliency maps, and the human-fixation density maps are shown in the last column. The preference of humans to fixate on the center-of-symmetry of the flowers is correctly reproduced by the symmetry model, whereas the contrast model puts focus on the edges of the forms. The regions of the image that are highlighted are the parts of the maps above 50% of its maximum value.

experiment that is discussed in Chapter 3. The majority of eye fixations of the participants are concentrated at the center of the symmetrical objects (see last column). The response of the contrast-saliency model shown in the second column, however, shows much more spread over the whole image, and no particular concentration on the center

of the objects. To the contrary, fixations are predicted at the border of the objects, where the contrast with the background is high. The symmetry-saliency model, on the other hand, much more specifically predicts eye fixations in the center of the objects (see third column). Although contrast has been shown to predict human gaze, Figure 2.5 shows that the predictions do not always correspond with human behavior. The next chapter shows that not only for these images, but more generally for a wide variety of photographic images, the symmetry-saliency model better correlates with the human eye-fixation data than the contrast-saliency model.

The observation that humans pay attention to the symmetrical center of objects motivated me to explore the use of symmetry in visual attention. In the remainder of this section, symmetry and its role in vision are discussed.

2.6.1 Types of Symmetry

Figure 2.6a displays three different types of symmetry: *mirror*, *rotational*, and *translational* symmetry. A pattern that contains mirror or reflectional symmetry has a symmetry axis. Mirroring the pattern in that symmetry axis will results in the same pattern. A rotationally symmetrical pattern can be rotated over an angle $\leq 180^\circ$ and remain identically. Finally, in a translationally symmetrical pattern, a part of the pattern is repeated without mirroring. In this dissertation, the focus is on mirror symmetry exclusively. Therefore, all occurrences of the word *symmetry* in the text refer to *mirror symmetry*.

Different types of mirror symmetry can be distinguished. First of all, the orientation of the symmetry axis can vary, as can be appreciated in Figure 2.6b. In left-right symmetry, the symmetry axis is vertically oriented. This type of symmetry is therefore called *vertical symmetry*. Likewise, the axis is horizontally oriented in *horizontal symmetry*. All other orientations of the axis are associated with *oblique symmetry*. Besides the orientation of the axis, the number of symmetry axes can vary (see Figure 2.6c). The frontal view of a human face, for instance, has one symmetry axis, whereas a book has two, and a sunflower has many axes of symmetry.

2.6.2 Symmetry in Our Visual Environment

We regularly experience visual symmetries in our daily lives (see (Hargittai & Hargittai, 2009) for many nice photographic examples). Most living things, for instance, have

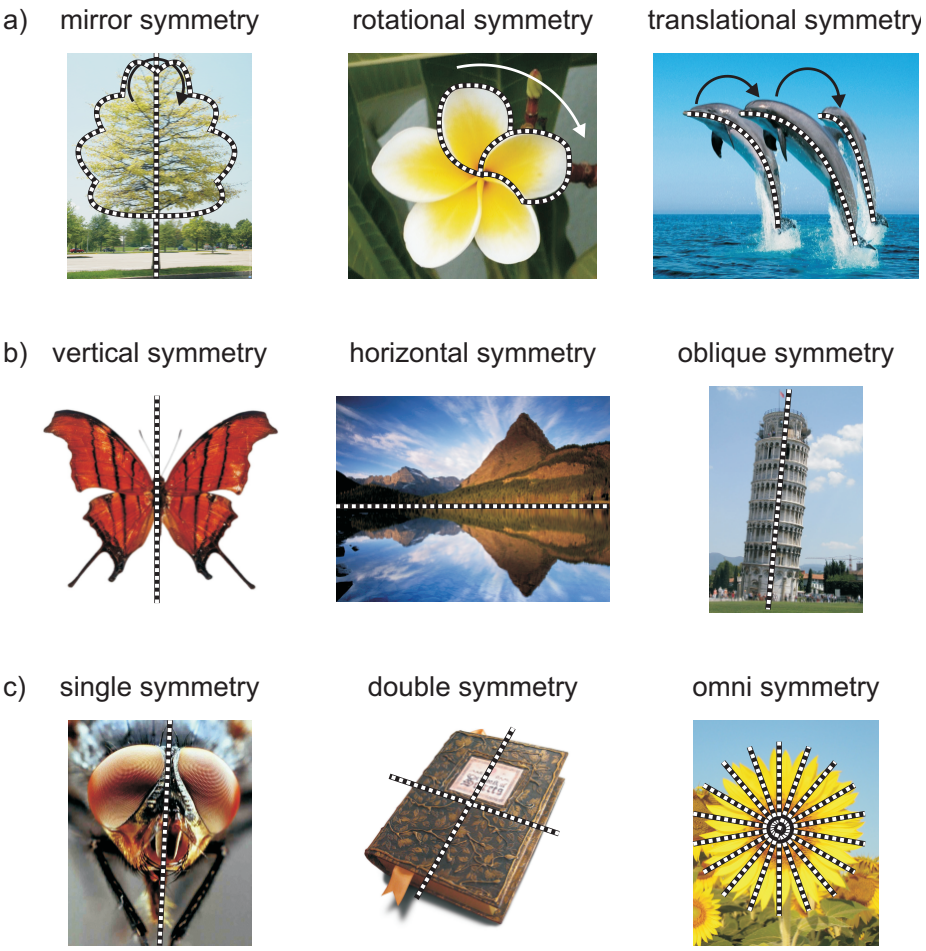

Figure 2.6: Different types of symmetry: a) shows the three types of symmetry, mirror symmetry, rotational symmetry, and translational symmetry. b) In mirror symmetry, the axis of symmetry can be oriented vertically, horizontally, or obliquely. c) Patterns with different numbers of symmetry axes. Omni symmetry is sometimes referred to as radial symmetry. In this thesis, mirror symmetry with an arbitrary number of symmetry axes and orientations is employed. (Source: `http://commons.wikimedia.org`).

Figure 2.7: The symmetries possessed by the palace and gardens of Versailles are wonderful examples of the tendency of humans to create symmetrical objects. (Source: `http://commons.wikimedia.org`).

a high degree of symmetry. Many animals display vertical symmetry, that is, frontally viewed, the left and right side are mirror symmetric. This symmetry is even an indication of the fitness of the individual. Human faces with artificially enhanced symmetry, for instance, are judged more attractive than the original faces (Grammer & Thornhill, 1994; Rhodes, Proffitt, Grady, & Sumich, 1998). Facial symmetry is a sign of overall phenotypic quality and developmental health (Thornhill & Gangestad, 1993). Moller & Thornhill (1998) performed a meta-analysis studying the relation between asymmetry and mating success in many animal species, and found a negative relationship, showing that more symmetric individuals are more sexually attractive. Not only animals, also many plants are symmetrical or contain symmetrical parts. Leaves and especially flowers often contain mirror, rotational, and translational symmetries.

Also most man-made objects, like tools and buildings, are symmetrical. The palace and gardens of Versailles near Paris are excellent examples (see Figure 2.7). In general symmetry is preferred over asymmetry in architecture and art (Tyler, 2000).

Symmetry contributes to the *figural goodness* according to Gestalt psychologists. A

symmetrical figure is subjectively experienced as nicer, simpler, and more organized by humans than asymmetrical figures (Palmer, 1991).

This abundance of symmetry in our visual environments plus the tendency of humans to create symmetrical objects and judge symmetry as beautiful suggests that humans are sensitive to symmetry, and that the human visual brain is equipped with symmetry detectors. This is discussed in the next section.

2.6.3 Sensitivity to Symmetry

Due to abundance of symmetry, it is not surprising that humans are sensitive to symmetry (Wagemans, 1997). Humans very rapidly detect symmetrical patterns, especially when the pattern contains multiple axes of symmetry (Palmer & Hemenway, 1978). Evans, Wenderoth, & Cheng (2000) showed that this is even more so when more complex photographic stimuli are used instead of simple line and dot figures. Similarly, recognition performance increases when symmetrical patterns are presented (Royer, 1981). Corballis & Roldan (1975) found that symmetry detection is most efficient for vertical orientation of the symmetry axis, followed by horizontal symmetry, and least efficient for oblique symmetry. Similar results were found by Evans et al. (2000). The improvement in performance might be explained by the intrinsic redundancy present in symmetrical forms, which gives rise to simpler representations (Barlow & Reeves, 1979). Not only humans display this sensitivity to symmetry, it is also found in other animals like doves (Delius & Nowak, 1982) and macaques (Beck, Pinsk, & Kastner, 2005).

The detection of symmetry in figures by humans is highly efficient, where as the detection of repetition, that is translational symmetry, is not Baylis & Driver (1994). The complexity of symmetrical figures has only a small influence on the reaction times, whereas the complexity is highly influential in for figures containing repetition. This shows that symmetry perception is parallel and not done by a pointwise serial matching process. It more over suggests that the perception of symmetry is preattentive, that is, symmetry can be perceived without the need of attention (Wagemans, 1995, 1999).

According to Palmer & Hemenway (1978), symmetry recognition consists of two phases. In the first preattentive phase (50-200ms), a rough symmetry detection takes place in which an estimation of the position and the orientation of the symmetry axis is made. This is followed by an attentive verification phase (2000-4000ms), in which a

closer investigation of the pattern takes place to verify the symmetry of the pattern in detail.

A developmental process in symmetry detection in children has been found (Bornstein & Stiles-Davis, 1984). Four-year-old children can discriminate only vertical symmetry. Five-year-olds can also discriminate horizontal symmetry, and six-year-olds posses the ability to detect oblique symmetrical forms. This corroborates the work of Corballis & Roldan (1975) that vertical symmetry is most efficiently detected. Also Palmer & Hemenway (1978) found fastest detection of vertical symmetry, then horizontal symmetry, and finally oblique symmetry. Fisher, Ferdinandsen, & Bornstein (1981) reported the ability of four-month-old infants to discriminate a vertically symmetrical form from an asymmetrical one. The infants were unable to do discriminate horizontal symmetry. These studies suggest that vertical symmetry detection is innate in humans or at least needs relatively little experience to develop. It can be assumed that the fact that horizontal and oblique symmetries are less frequently occurring visual stimuli is the reason for the worse and later-developed sensitivity to these stimuli.

Symmetry also influences eye movements. Fixations on symmetrical forms are concentrated at the center of the form, or at the crossing points of the symmetry axes (Kaufman & Richards, 1969). In free viewing photographic images, the amount of symmetry is significantly higher at the points of human fixation than on average in the image. This effect is stronger for symmetry than for contrast at the fixation points (see Chapter 3). Similarly, a center-of-gravity effect or global effect is reported, showing the tendency of eye saccades to land at the geometric center of a target object or target configuration (Findlay, 1982; He & Kowler, 1989; Ottes, Van Gisbergen, & Eggermont, 1984). Bindemann, Scheepers, & Burton (2009) showed that the first eye movements to human faces land on the center of gravity of the face independent of the three-dimensional orientation of the face. The subsequent fixations focus on more detailed facial features like the eyes and the nose. The center of gravity of a pattern usually is approximately its center of symmetry, and the center-of-gravity effect can thus be predicted on the basis of local symmetry, with the advantage that there is no need for prior segmentation of the object. Furthermore, for images containing a single axis of symmetry, the fixations are concentrated along this axis, whereas they are more spread out on non-symmetrical images (Locher & Nodine, 1987). In this paper, we also investigate the role of symmetry in guiding eye movements. However, instead of using relatively simple artificial stimuli with only one symmetrical pattern, we presented our participants with complex photographic images with natural and man-made scenes.

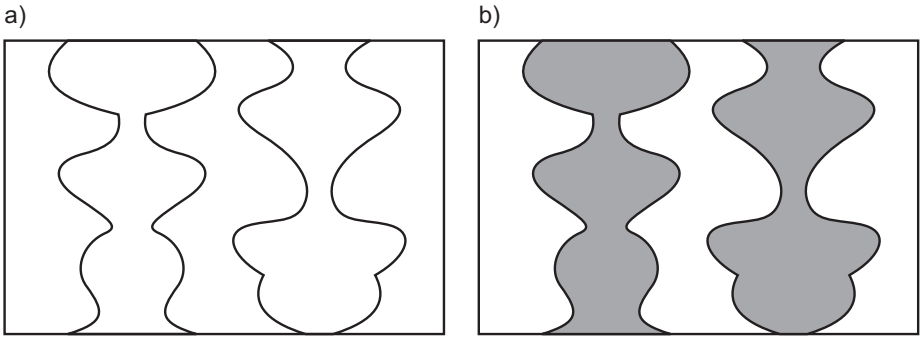

Figure 2.8: Symmetry as a cue for figure-ground segregation. If people are asked to determine what the foreground is and what the background in (a), they generally determine that the symmetrical areas, marked gray in (b), are the foreground. Symmetry is non-accidental. When two contours are symmetrical, they highly likely belong to the same object.

This shows that humans are sensitive to symmetry, and that symmetry influences overt visual attention. In addition, symmetry plays a role in early object segmentation. According to the Gestalt law of Prägnanz (Koffka, 1935; Köhler, 1947), symmetry is one of the principles to find the simplest and most likely interpretation of the sensory input. This hypothesis is supported by the fact that symmetry is a cue for figure-ground segregation. Humans usually see the symmetrical areas of an image as foreground (Driver, Baylis, & Rafal, 1992). If people are asked to determine the objects and the background in a display as shown in Figure 2.8a, the majority will give the answer as shown in Figure 2.8b. The symmetrical parts of the figure are considered object and the non-symmetrical parts are considered background. Also Machilsen, Pauwels, & Wagemans (2009) found support for symmetry as a cue for figure-ground segregation. In their experiments, symmetrical shapes were easier to detect across different noise levels than asymmetrical shapes. This suggests that symmetry can be used for context-free object segmentation. Since visual attention is likely to be object-oriented (Scholl, 2001; Yeshurun, Kimchi, Sha'shoua, & Carmel, 2009), symmetry might play an important role in the bottom-up guidance of eye movements. To gain insight in this topic, Chapter 3 investigates how well eye fixations can be predicted on the basis of local symmetry. Also in Chapter 4 the role of symmetry in visual attention is studied.

2.6.4 *Memory and Representation of Symmetrical Forms*

Deregowski (1971) showed that humans can reproduce patterns that are mirror symmetrical about a vertical axis better than patterns that repeat a sub-pattern in the same orientation. De Kuijer, Deregowski, & McGeorge (2004) also found that symmetrical patterns are reproduced better than asymmetrical ones. Moreover, the symmetrical property is often correctly reproduced, even when the pattern itself is not correctly copied. They furthermore identified that the orientation of the symmetry axis influences the quality of a reproduction. This suggests that the symmetry is exploited in the internal representation of the pattern and that especially vertical symmetry is easily detected and memorized.

Similarly, Attneave (1955) compared the recognition and reproduction of symmetrical and asymmetrical patterns. When the two patterns contained the same number of points, the symmetrical pattern was easier to recognize and reproduce. It is likely that this is due to the intrinsic redundancy in the symmetrical pattern. Although Attneave showed that the exploitation of this redundancy is not perfect, it indicates that there is some perceptual mechanism capable of organizing and encoding the redundant pattern into a simpler and more compact representation.

2.6.5 *Neural Correlates*

In a functional MRI (fMRI) study, Sasaki et al. (2005) presented line-based and dot-based stimuli of various sizes that were either symmetrically or randomly organized. Robust brain activity was found in higher-order regions of the human visual cortex, especially in areas V3a, V4v/d, V7, and the lateral occipital complex (LOC) (see Figure 2.9). The later is also involved in object recognition (Grill-Spector, Kourtzi, & N., 2001) and object-shape representation (Kourtzi & Kanwisher, 2001). Little activity was measured in the lower-order visual parts of the brain. The same fMRI response was also found without attentional control, when participants had a task unrelated to the symmetrical pattern. This confirms the behavioral studies discussed above that suggest that symmetry detection is preattentive and bottom-up. The fMRI response was slightly stronger for vertical symmetry than for horizontal symmetry. Moreover, the response was somewhat stronger for patterns with two axes of symmetry instead of one. This is in accordance with the psychophysical experiments presented earlier as well.

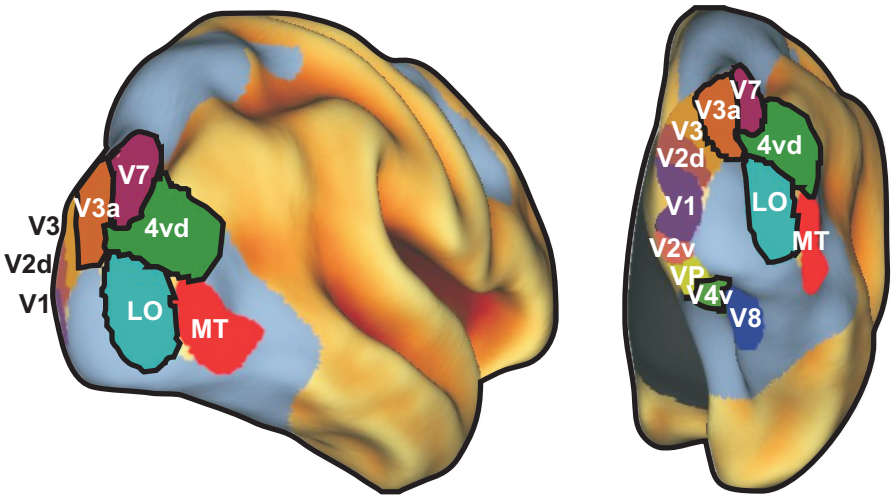

Figure 2.9: Visual brain areas involved in symmetry perception. The images show the right hemisphere of a human brain. The left image gives a view from right back, and the right from full back. V3a, V4d/v, V7, and the lateral occipital complex (LOC) are shown to play a role in symmetry perception (Sasaki et al., 2005).

Symmetry detection probably involves long-range interconnections between cortical orientation filters (Saarinen & Levi, 2000). Levi & Saarinen (2004) studied symmetry detection in humans with amblyopia, also known as lazy eye, and found that amblyopia severely impairs the detection. It is suggested that the loss of symmetry detection is a result of a deficit in the integration of local orientation information over long-range interconnections in the brain.

This dissertation does not look at the neural correlates of symmetry detection. Instead the role of local symmetry to attract human eye fixations is investigated. The aim is to contribute both to the study of bottom-up visual attention and to the study of symmetry perception. However, as pointed out by Beck et al. (2005), more fMRI studies should be done to improve the understanding of symmetry detection by the human brain.

2.7 Conclusion

This chapter discussed overt visual attention in humans as an active method to efficiently view the world. The top-down and bottom-up influences on visual attention are discussed and some studies on visual search are reviewed. From these studies, a number of basic features have been proposed that are involved in bottom-up visual attention. Based on contrast calculations on these basic features, a number of visual-attention models have been proposed in the literature. However, basic features are not the only factors in bottom-up visual attention. Configural features also have a strong influence, such as, notably, symmetry. This was further demonstrated by the correct predictions of human eye fixations based on symmetry when humans view symmetrical objects. The contrast-saliency model fails to predict human gaze in these situations. Psychophysical and neurophysiological studies show that symmetry plays a role in visual processing and that humans are sensitive to symmetry. The results of these studies moreover suggest that symmetry perception is parallel, preattentive, and efficient. This thesis investigates whether symmetry can be used to model human visual attention.

The remainder of this part of the dissertation focuses on the role of symmetry in visual attention. In Chapter 3, a saliency model based on local symmetry for the prediction of human eye fixations is proposed. To test the performance of the model, it is compared with data gathered in an eye-tracking experiment. The results are compared to that of the contrast-saliency model. In Chapter 4, the preattentive perception of symmetry and the role of symmetry in visual attention is further investigated. In a visual-search experiment, a possible pop-out effect of symmetry is investigated and a scene-memory experiment is used to study whether symmetrical objects attract more attention than non-symmetrical objects. Finally, the object-oriented nature of human visual attention is discussed in Chapter 5, as well as the role of symmetry in the bottom-up detection of objects. To investigate the applicability of symmetry for machine vision, symmetry is used in Part II of this dissertation to guide the attention of a robotic system.