Gert Kootstra

Visual Attention and Active Vision

From Natural to Artificial Systems

Visual Attention and Active Vision

From Natural to Artificial Systems

This research was generously supported by SR Research.
This research was generously supported by BCN.

ISBN: 978-90-367-4367-9

RIJKSUNIVERSITEIT GRONINGEN

Visual Attention and Active Vision
From Natural to Artificial Systems

Proefschrift

ter verkrijging van het doctoraat in de
Wiskunde en Natuurwetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. F. Zwarts,
in het openbaar te verdedigen op
maandag 17 mei 2010
om 13:15 uur

door

Geert Willem Kootstra

geboren op 23 december 1978
te Boven-Smilde

Promotor:	Prof. dr. L. R. B. Schomaker
Copromotor:	Dr. B. de Boer
Beoordelingscommissie:	Prof. dr. E. O. Postma Prof. dr. J. Wagemans Prof. dr. B. M. ter Haar Romeny

Contents

1

General Introduction

“....perceiving is an act not a response, an act of attention, not a triggered impression, an achievement, not a reflex”

(Gibson, 1979)

With the rise of computers in the 1950's, a new view on human cognition emerged. Psychologists came to think of cognition as computation. In analogy to computers, humans were viewed as information processing systems: the sensors provide input from the external world, which is used to build a symbolic world representation. The mind processes these symbols, which leads to a decision, and the system gives output in the form of an action. This *sense-think-act cycle* emerging from the computer metaphor, resulted in a strong emphasis on the information-processing aspects of cognition. Within psychology, this view on cognition, also termed *cognitivism*, led to valuable theories about representation, memory and reasoning. In the field of artificial intelligence, the focus was on symbol processing, reasoning, and search, which led for instance to sophisticated chess programs and reasoning systems.

Although the cognitivistic paradigm has given rise to many interesting theories and applications, it became apparent that the paradigm was not appropriate for operation in the real world when it was first applied to the robot Shakey (Raphael, 1976). In the early 1970's, Shakey displayed some impressive abilities. However, its behavior was far from real-time, taking Shakey several minutes to perform the next action. In contrast, Walter (1950, 1951) developed robots a decade earlier that did operate real time using subsymbolic and analog mechanisms. It became evident that the focus on symbolic representations and reasoning alone did not lead to real-life performance. The problems of perception and action were severely underestimated. The reason behind this underestimation is the so-called Moravec's paradox (Moravec, 1988): conscious tasks that we humans think are very difficult, like playing chess and reasoning, are actually relatively simple for computers. However, the subconscious sensori-motor skills that all animals exhibit seemingly effortlessly, are actually very complex and require huge computational resources.

One of the problems with the sense-think-act approach to cognition is that perception is viewed as a passive process that only serves to process the raw information from the external world and to pass it on to the think and act modules. Perception in that view is completely independent from action. However, in reality, perception and action are strongly intertwined: **perception is an active process**. Perception leads to action, but through the interaction with the environment action leads to new percepts as well.

This view on cognition is termed *embodied cognition* (Pfeifer & Scheier, 1999; Brooks, 1999; Varela, Thompson, & Rosch, 1991; Lakoff & Johnson, 1999). An intelligent system is viewed as a system that is embodied and situated, a system that can perceive the

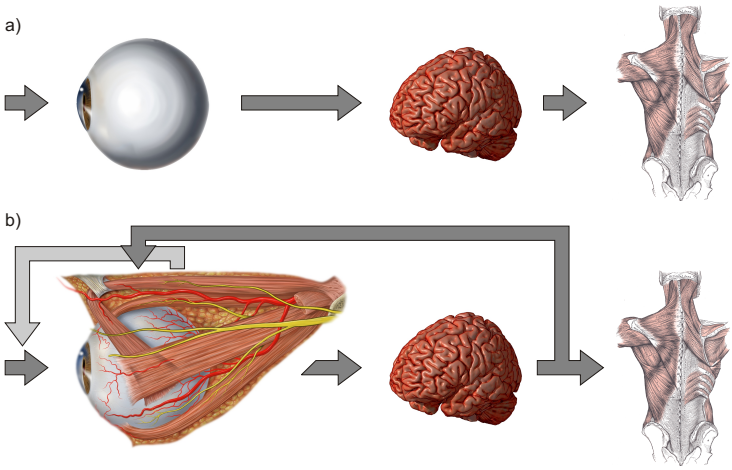

Figure 1.1: Active versus passive vision. a) *Sense-think-act*, the classical view on cognition: the world is perceived through the sensory system and the information is processed by the brain, resulting in an action. In this view, vision is the passive processing of the information that falls on the retina. b) However, in most natural systems, vision is highly active, as illustrated by the fact that the human eye has photoreceptors, which sense the world, but also extraocular muscles, which control the orientation of the eye. Based on the sensory information and top-down knowledge, the brain actively controls and structures the visual input. Sensing not only results in acting, but acting also results in sensing.

environment and act upon this environment. An intelligent system is thus a system that is in constant interaction with its environment. Perception not only serves to assist action, but action also serves to assist perception by structuring the input. The active capabilities of natural systems are constantly used to direct perceptual attention to specific parts of the external world. By actively shaping the content of perception, the computational resources are more efficiently used and many otherwise difficult perceptual tasks are greatly simplified.

This dissertation focuses on visual perception from an embodied-cognition perspective. In particular, it discusses visual attention and active object recognition in natural and artificial systems.

1.1 Visual Attention

Humans make on average 3 to 5 eye movements every second, which sums up to something in the order of 4.5 billion eye movements in a lifetime. The purpose of making an eye movement is to focus attention on relevant and interesting information in the visual field. In other words, it is an active process to explore and perceive the visual environment.

The necessity of making eye movement lies in the distribution of photo-receptor cells in the retina. These cells sense the light from the environment. There are two types of photo receptors, cones and rods. The *rods* are very light sensitive, but because there is only one type of rod cell, they are not sensitive to color. Moreover, the visual acuity (i.e., image resolution) of rod cells is relatively low. Rods are responsible for vision in poor light conditions and for sensing movement. The cells are mainly located in the periphery of the retina. The *cones*, on the other hand, are located in a small area at the center of the retina, in the fovea. The field of view of the fovea (or fovea centralis) is approximately 2°. In this small area, the cone cells are densely packed. The fovea therefore provides high-acuity vision. Moreover, there are three types of cones, which are sensitive to different wavelengths, making the cells sensitive to color.

The fovea is therefore used to identify the color and form of objects in high resolution with the drawback that only a small part of the total field of view is used. To be able to view all relevant parts of the visual environment with high acuity, humans, like many other vertebrates, constantly shift the focus of the fovea. This is done by making eye, head, and body movements. This process is called *overt visual attention*, as opposed to *covert visual attention*, which is the process of mentally focusing attention.

The reason that the human eye has not evolved to have high-acuity vision throughout the retina is that that would result in an enormous stream of visual information, which would require a gigantic brain to process all the information. Moreover, it is unnecessary to have such a wide field of high-acuity vision, because large parts of the field of view are uninteresting and contain little or no information. Having a small fovea, and making eye movements is simply the most efficient and effective way to view the world.

Overt visual attention is a process that is controlled both *bottom up*, that is, from the stimulus, and *top down*, that is, from the interests, knowledge, and memory of the agent and form the task at hand. Although the top-down aspects of visual attention are

undeniably important and interesting to study, this dissertation focuses on the bottom-up aspects only. We are interested in the properties of the visual stimulus that attract visual attention. In our view, bottom-up visual attention mainly serves as a filter to select potentially interesting points in the stimulus that are then further processed under supervision of top-down influences.

Visual attention can thus be seen as a mechanism to filter the enormous amount of available visual information and to focus on the relevant parts. From a perspective of natural intelligence, it is interesting to study how this mechanism works, to find out on what information the direction of attention is based. Why are certain parts attended and others not?

This dissertation focuses on the bottom-up aspects of visual attention. Studying overt visual attention gives insights in these mechanisms, providing more knowledge about human visual processing and allowing to build predictive models. From an artificial-intelligence perspective, it is interesting to see if the visual-attention mechanisms can be copied to improve robot vision. Just as human vision, robot vision suffers from an overload of visual information. Developing attentive systems allows the robot to effectively and efficiently focus the computational resources on only the interesting parts of the visual field. Since nature solved the problem so well, insights in human visual processing are of great interest for artificial intelligence.

1.1.1 The Role of Symmetry in Visual Attention

In the literature, different bottom-up models have been proposed to predict human eye movements (see Chapter 2 for a discussion). These models have in common that they are based on contrast. The models determine the saliency at a given point in the visual field by comparing the center to its local surroundings. The saliency is high when features like intensity, color, and orientation are different in the center than in the surround. Although these models have been shown to predict human eye fixations to some extent, there is room for improvement.

This dissertation proposes the use of symmetry to predict human visual attention. The daily environments of humans contain many symmetrical objects, from living things to man-made objects. Since objects are highly interesting for the interpretation of a visual scene, symmetry is hypothesized to play a role in human visual attention. The role of symmetry in visual attention is studied in this dissertation, not only to predict human

eye fixations, but also to guide the attention of a mobile robot. The results show that visual-attention models based on symmetry can outperform models based on contrast. Symmetrical forms are furthermore shown to be perceived efficiently. This indicates that symmetry plays a role in human visual attention. Moreover, the use of symmetry in guiding the attention of a mobile robot is shown to improve the selection of interest points in the environment.

1.2 Objectives

This dissertation discusses a multi-disciplinary study of visual attention and active object recognition. The topics will be approached from both a natural and an artificial perspective. The first topic deals with the prediction of overt visual attention, that is, visual attention expressed by making eye movements. Human eye movements are recorded in an eye-tracking experiment, and compared to computational saliency models that make predictions of the location of eye fixations based on the stimulus. These models are then applied in a robotic context, to guide the visual attention of a navigating robot. The second topic deals with the recognition of objects using active vision to change the point of view. Here, inspiration is taken from object recognition in infants, and a model for a mobile robot is developed to actively learn and recognize objects.

Topic 1: The Role of Symmetry in Visual Attention. The process of making eye movements is an outstanding example of active vision. By actively shifting the focus of attention, a sharp and detailed image of the visual surroundings is acquired. The eyes do not randomly move around, but are focused on relevant and interesting parts of the visual field. This process is controlled by a visual-attention mechanism. This mechanism determines the fixation locations based on bottom-up low-resolution input to the parafoveal and peripheral parts of the retina, as well as on top-down information, such as earlier fixations, memory, world knowledge, interests, and the current task.

A large part of this dissertation deals with the bottom-up aspects of this visual-attention system. In particular, the role of visual symmetry in attracting attention is studied. Part I discusses the role of symmetry in human visual attention. Models of bottom-up visual attention are proposed to predict eye fixations based on local symmetry in the image. Part II studies the use of symmetry to guide the attention of a mobile robot to learn

and recognize its visual environment. This multi-disciplinary study of visual attention provides a better understanding of natural vision systems and leads to improvements in artificial vision systems.

Topic 2: Active Object Recognition. Besides visual attention, Part II discuss the use of active vision for object recognition. This topic is also inspired by natural systems and the resulting model is applied to an artificial system to recognize three-dimensional objects. When learning a new object, humans usually explore the object, to acquire different viewpoints of the object (see Figure 6.6 for an example). By the change of perspective, more information about the object can be gathered, and a complete three-dimensional representation of the object can be formed. Moreover, by manipulating the object, it becomes apparent what belongs to the object and what belongs to the environment. Not only in learning, but also in recognition, active vision often plays a role. Especially when the observation is ambiguous, new viewpoints need to be gathered to confidently recognize objects. Such ambiguity can arise when an object is indistinguishable from another object from a certain perspective, like a hatchback car and a station wagon look the same from the front. The amount of ambiguity also depends on the quality of the sensory system. A very good sensory system, like the eyes of a grown-up human being, has fewer problems with ambiguous observation than a poor system like the eyes of young infants or, as will be discussed in the thesis, cameras of mobile robots. Especially with poor quality sensors, there is a necessity to actively gather more visual information about the object. This thesis shows that active vision simplifies perceptual tasks and improves recognition performance.

Research Questions The main research questions in the dissertation are:

1. What are the spatial features that attract human overt visual attention?
2. How can we construct saliency models that predict human eye fixations?
3. Can we apply these saliency models to improve robotic vision to select points and regions of interest?
4. Can we improve robotic vision by using an active approach to vision?

Main Conclusions Based on the studies, experiments, and results described in the current work, we can draw the following main conclusions:

- From natural vision:
  - Humans pay more attention to the symmetrical parts of the visual field (Chapter 3).
  - Saliency predictions based on local symmetry compare significantly better with human eye fixation than predictions based on contrast (Chapter 3).
  - Symmetry is perceived efficiently and preattentively, that is without the need of attention focused on the object (Chapter 4).
  - It is likely that attention is object oriented and symmetry is a cue for the presence of an object, thus a predictor of overt visual attention. (Chapter 4 and Chapter 5)
- From artificial vision:
  - An active approach to object recognition simplifies the perceptual tasks and significantly improves the recognition performance (Chapter 7).
  - Points and regions of interest selected using local symmetry are more stable and more robust to noise than when using contrast features (Chapter 8 and Chapter 9).
  - The use of symmetry to guide the attention of a mobile robot results in a better representation of the environment (Chapter 8 and Chapter 9).
  - Local symmetry is possibly usable for context-free object segmentation (Chapter 9).

1.3 A Multi-Disciplinary Study

The presented research is multi disciplinary. Active vision and visual attention are studied from both a natural and an artificial perspective. Psychophysical experiments are conducted to investigate the bottom-up aspects of human visual attention, computational models are developed to predict the location of human eye fixations, and these attentive models are implemented in a robotic context. The goals of such a multi-disciplinary approach are illustrated in Figure 1.2.

On the one hand natural intelligence is studied to learn about natural intelligence and to use this knowledge to improve artificial systems. Natural systems are evolved over billions of years to deal with all kinds of problems and challenges in the world. For

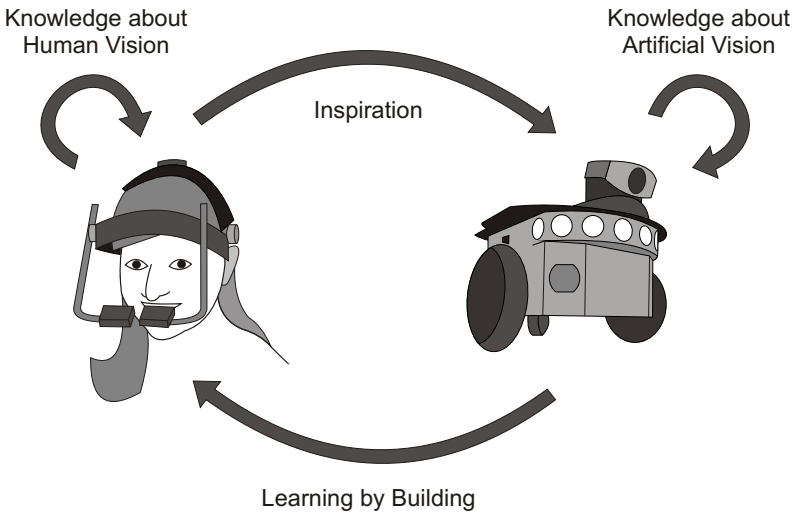

Figure 1.2: Understanding visual perception using an analytic and synthetic approach. By analyzing human behavior, we learn about visual attention in natural systems. This knowledge can be taken as an inspiration to develop artificial systems. By implementing and analyzing visual-attention models on robots, we learn about vision in artificial systems, which enables us to improve machine vision. Moreover, by building the systems, we learn about the general principles of vision, which also has consequences for the study human visual processing.

artificial systems, these problems and challenges are very similar. Robotic systems encounter the same world as natural systems. The solutions used by natural systems can be used to develop intelligent artificial systems. This thesis concentrates on the phenomenon of active vision. By studying topics such as visual attention and active visual exploration in humans, we learn about the relevant aspects of these behaviors. These aspects can be used to guide the visual attention of robotic systems, as well as to let robots actively explore objects and scenes in order to improve recognition.

On the other hand, by building these artificial systems we learn about intelligence in general. The insights that we gather from studying artificial systems also have relevance for natural systems. This process is termed *learning by building* by Pfeifer & Scheier (1999). When using an *analytic approach*, one studies intelligent systems by performing experiments on the system and analyzing the results. Although this

approach has resulted in many interesting insights, the problem remains that the explanation of the behavior is in abstract terms, containing many *black boxes*. However, if one applies a *synthetic approach* by actually building systems that reproduce the behavior, the explanation needs to be very concrete, simply because the artificial system will otherwise not work. In the process of building and analyzing the system, the positive and negative sides, and the possibilities and limitations of the explanation will become clear. By implementing and studying active vision methods in robotic systems, we learn about relevant aspects of visual attention and active visual exploration, witch also sheds a light on the same processes in natural systems. The synthetic study can thus lead to new analytic studies, thereby closing the loop in Figure 1.2.

Chapter 10 at the end of this dissertation discusses what has been learned from this multi-disciplinary study of visual attention and active vision. Figure 10.1 highlights the main findings.

1.4 Organization of the Thesis

The thesis consists of two main parts. Part I discusses visual attention in natural systems. This part begins with an introduction to visual attention in human vision in Chapter 2. Chapter 3 presents a model to predict human eye fixations using local symmetry. This model is compared to human data obtained in an eye-tracking experiment while the participants viewed complex photographic images. In Chapter 4, experiments are discussed to see if symmetry results in a pop out. The part ends with a discussion on the role of symmetry in visual attention in Chapter 5.

Part II deals with visual attention and active vision in artificial systems. An introduction is given in Chapter 6. Chapter 7 describes a method for robotic vision to recognize objects by exploration. In Chapter 8 and 9, the symmetry models discussed in the first part are applied to robotic navigation. In Chapter 8, symmetry is used to select points of interest as visual landmarks to construct a map of the environment, and Chapter 9 discusses a similar approach to select regions of interest, which might correspond to objects in the environment.

A general discussion on visual attention and active vision in natural and artificial systems is given in Chapter 10. This discussion concludes the dissertation.

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.

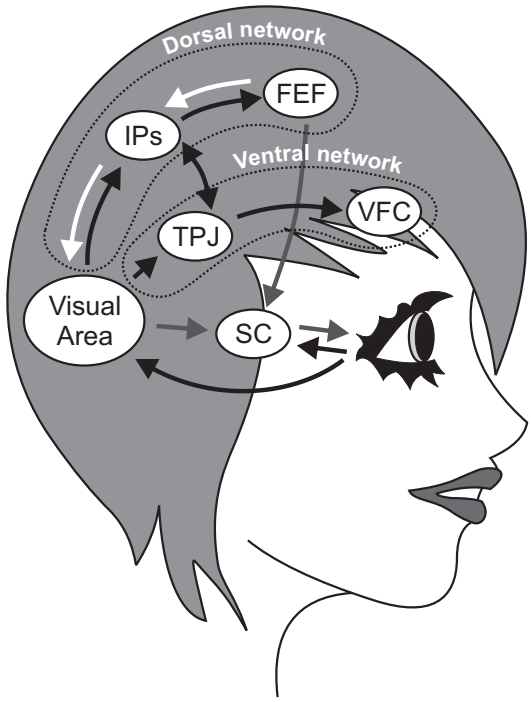

Figure 2.1: Top-down and bottom-up control of visual attention and eye movements in the human brain. The superior colliculus (SC), located in the midbrain, plays a central role. This brain area receives input directly from the retina, and from other brain areas, most notably the visual areas in the visual 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*

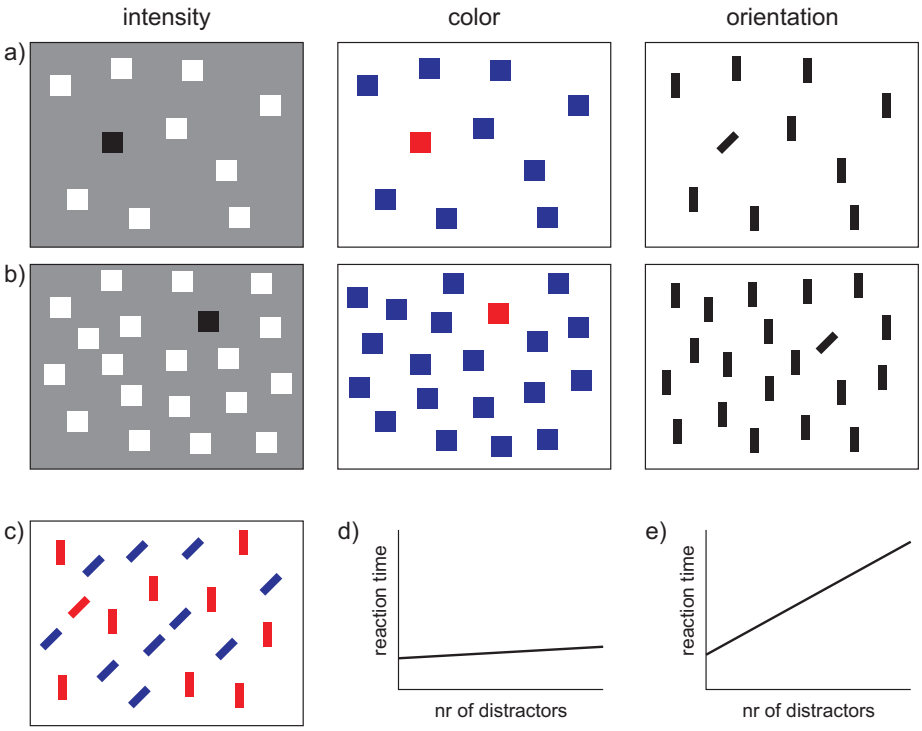

Figure 2.2: The pop-out effect. a) Three examples of feature-search display where the target is clearly visible, because it differs in one unique feature. As can be appreciated by looking at (b), human reaction times in finding the target are hardly influenced by the number of distractors, that is, the 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).

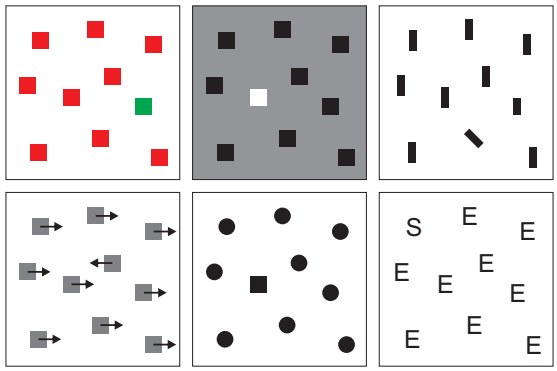

Figure 2.3: A list of basic features. From left to right, top to bottom: color, intensity, orientation, motion, and two examples of shape. The pop out of the target depends on the feature difference at the center and at its surround.

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

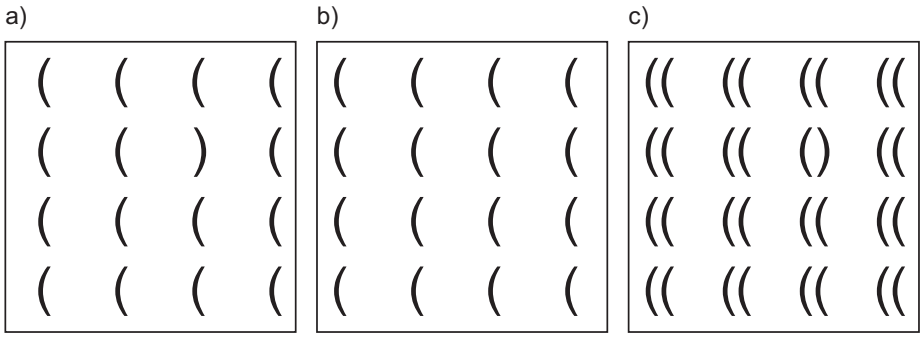

Figure 2.4: Configural superiority. Search for the rightward curved line segment among 15 distractors in (a) takes longer than among 31 distrators in (c). Although one would intuitively think that the addition of the 16 identical items in (b) would lead to less efficient search, it actually leads to 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

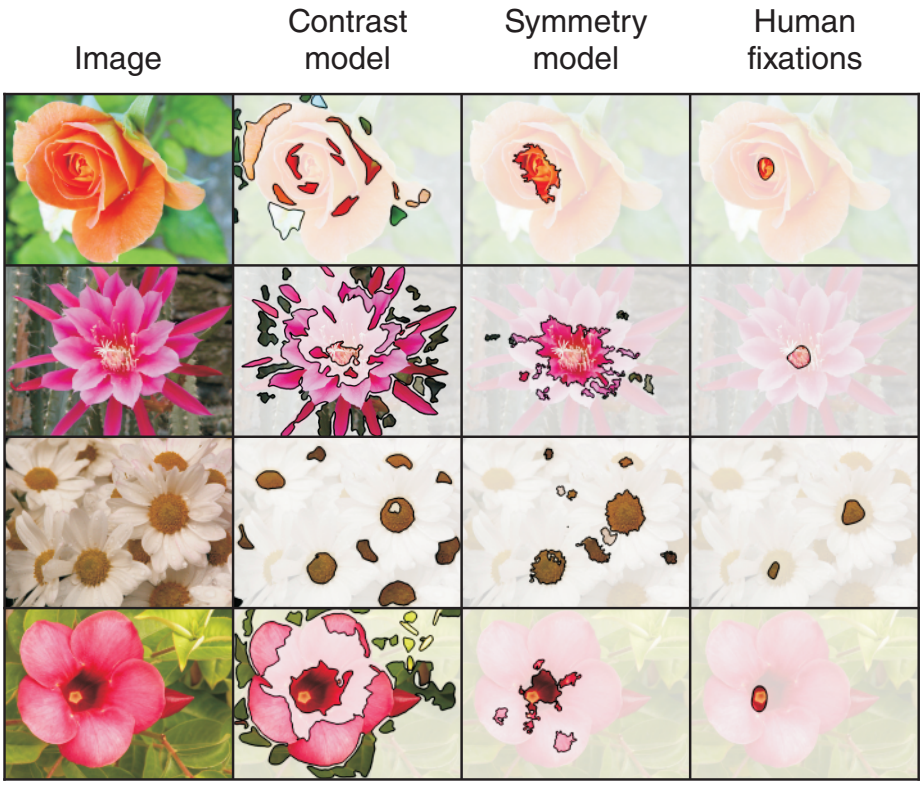

Figure 2.5: Examples of images containing symmetrical objects. The second column shows the contrast-saliency maps, the third column gives the symmetry-saliency maps, and the human-fixation density maps are shown in the last column. The preference of humans to fixate on the center-of-symmetry of the flowers 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

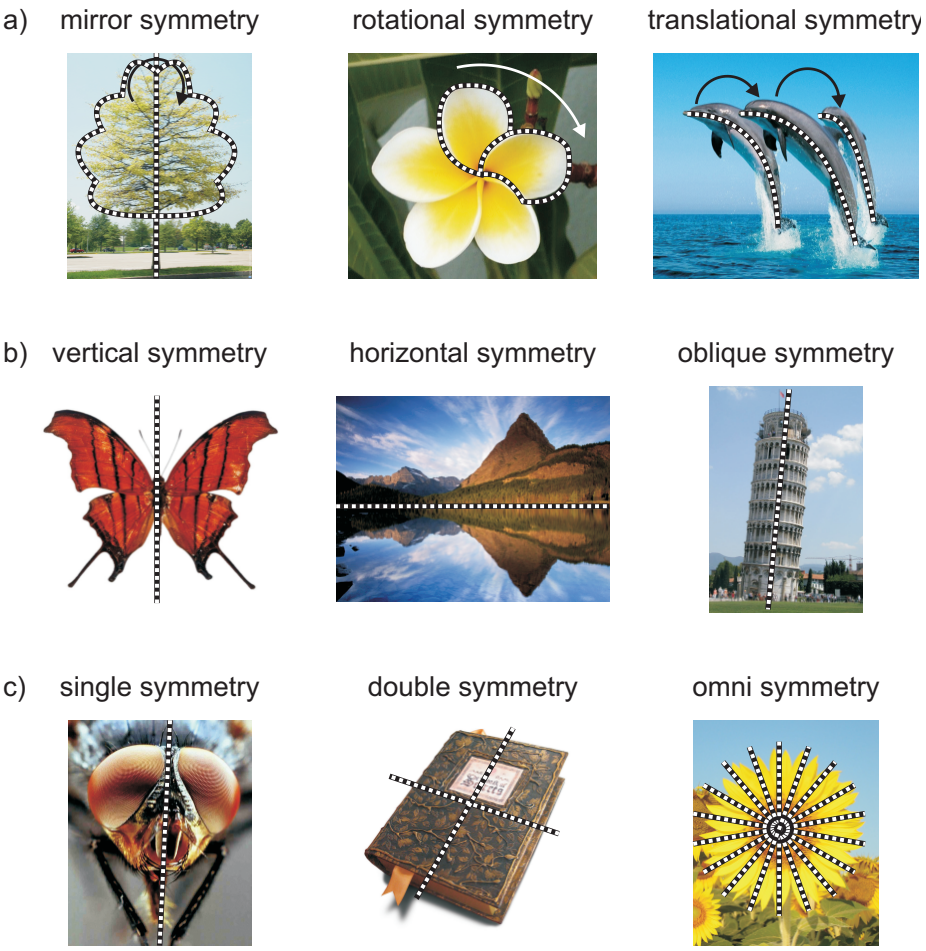

Figure 2.6: Different types of symmetry: a) shows the three types of symmetry, mirror symmetry, rotational symmetry, and translational symmetry. b) In mirror symmetry, the axis of symmetry can be oriented vertically, horizontally, or obliquely. c) Patterns with different numbers of symmetry axes. Omni 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.

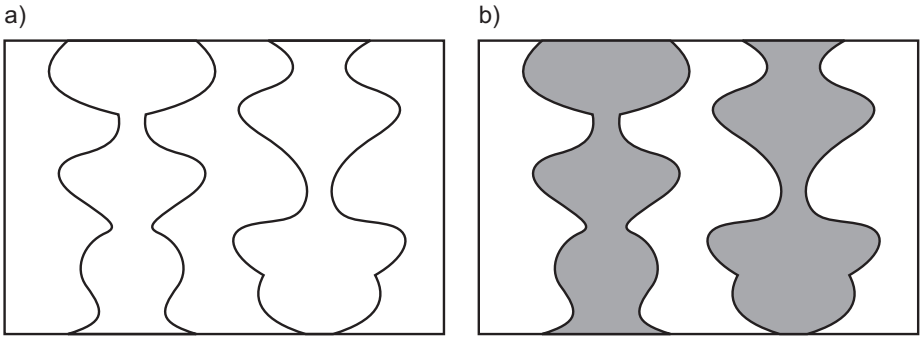

Figure 2.8: Symmetry as a cue for figure-ground segregation. If people are asked to determine what the foreground is and what the background in (a), they generally determine that the symmetrical areas, marked gray in (b), are the foreground. Symmetry is non-accidental. When two contours are symmetrical,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

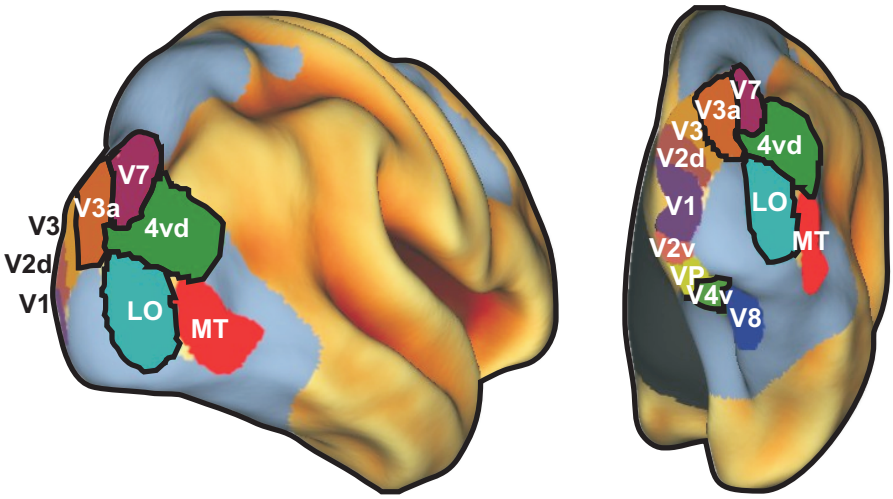

Figure 2.9: Visual brain areas involved in symmetry perception. The images show the right hemisphere of a human brain. The left image gives a view from right back, and the right from full back. V3a, V4d/v, V7, and the lateral occipital complex (LOC) are shown to play a role in symmetry perception (Sasaki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

3

Predicting Human Eye Fixations by Local Symmetry

Abstract

Most bottom-up models that predict human eye fixations are based on contrast features. The saliency model of Itti et al. (1998) is an example of such a contrast-saliency model. Although the model has been successfully compared with human eye fixations, we show that it lacks accuracy in the prediction of fixations on symmetrical forms. The contrast model gives high response at the borders of the forms. However, human observers consistently look at the symmetrical center of these forms. We propose a saliency model that predicts eye fixations using local symmetry. To test the model, we performed an eye-tracking experiment with participants viewing complex photographic images, and compared the data with our symmetry model and the contrast model. The results show that our symmetry model significantly better predicts eye fixations on a wide variety of images including many that are not selected for their symmetrical content. Moreover, our results show that especially early fixations are on highly symmetrical areas of the images. Our results show that symmetry is a strong predictor of human eye fixations, and that it can be used as a predictor of the order of fixation.

A modified version of this chapter is submitted as:

Kootstra, G., & Schomaker, L. R. B. (submitted). Prediction of eye fixations on complex visual stimuli using local symmetry. *Journal of Vision*.

Parts of the chapter have been published as:

Kootstra, G., & Schomaker, L. R. B. (2009a). Prediction of human eye fixations using symmetry. In *Cognitive Science Conference (CogSci)*. Amsterdam, The Netherlands.

Kootstra, G., Nederveen, A., & de Boer, B. (2008a). Paying attention to symmetry. In M. Everingham, C. Needham, & R. Fraile (Eds.) *British Machine Vision Conference (BMVC2008)*, (pp. 1115-1125). Leeds, UK.

3.1 Introduction

Humans continuously make eye movements to investigate the visual environment in an efficient manner. Interesting parts of the visual field are focused on, and inspected with high acuity. Eye movements are influenced both top-down, for instance based on the task at hand or past experiences, and bottom-up, based on properties of the stimulus. Although both influences play a role, we are only interested in the role of the stimulus in guiding eye fixations. The questions that are address in this chapter are: what are properties of the stimulus that attract overt visual attention, and can human eye fixations be predicted with bottom-up models?

More specifically, the role of local symmetry as an alternative to contrast for the prediction of eye fixations is investigated. We propose saliency models that calculate the conspicuousness in an image on the basis of symmetry, and discuss the results of comparing these models to human eye fixations recorded in an eye-tracking experiment. The main result shows that local symmetry is a better predictor of human gaze than contrast.

This chapter is organized as follows. The backgrounds of the presented research are discussed first. Then, the symmetry-saliency models are presented, along with the performed eye-tracking experiment and the methods to compare the models with the human data. Next, the experiments and results are presented, and the chapter ends with a discussion on these results.

3.2 Background

As discussed in Section 2.2, human eye movements are controlled both top down as well as bottom up. Although it is clear that both influences play a role, this chapter focuses on the bottom-up influences. We are interested in the role of the stimulus in the guidance of eye movements, specifically in the visual features that can be used to predict human eye fixations. This gives insights in the inherent properties of the stimulus that attract attention. To investigate this, a saliency model is proposed that determines the salient regions in an image, which is then compared to human eye fixations on the same images. Whereas most existing saliency models focus on contrast features to determine parts of the image that stand out from their local environment, the use of local symmetry to predict the eye movements is advocated in this dissertation.

3.2.1 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 control of eye movements. Not only does it give more insight in visual processing in natural systems, the model can also be used to predict overt visual attention and to improve computer and machine vision systems. Although dynamic features are undeniably of importance for the prediction of eye fixations, we focus on the static features in this paper.

Most existing bottom-up saliency models use contrast features to determine the saliency in an image. The influential saliency model of Itti and Koch, for instance, calculates the saliency of an image on the basis of contrast in three different feature channels: intensity, color, and orientation (Itti & Koch, 2001; Itti et al., 1998) (see Appendix A for details). The model is based on a biologically-plausible architecture for visual attention (Koch & Ullman, 1985), and is an implementation of the feature-integration theory of human visual search (Treisman & Gelade, 1980). It can correctly predict human behavior in visual pop-out experiments (Itti & Koch, 2000). The model has also been compared to human eye fixations on complex photographic images by Parkhurst et al. (2002). They showed that the saliency at the points of human fixation, as measured by the model, is significantly higher than expected by chance. Similarly, Ouerhani et al. (2004) found a positive correlation between the resulting saliency maps and human fixations.

Other saliency models, like the model of Le Meur et al. (2006) are also based on contrast calculations. They found a positive correlation between their model and human data that was slightly higher than the performance of Itti and Kochs model. Privitera & Stark (2000) investigated a set of simpler contrast-saliency operators. These operators were also found to predict human fixation points to some extent. Besides the contrast operators, Privitera and Stark also tested some other operators, including a simple symmetry operator, which also resembled the human data to some extent. The saliency model of Bruce & Tsotsos (2009) compares the distribution of features in the center to the surround, and defines the saliency based on the contrast between the two. The center-surround structure also emerged as the most representative receptive fields when fitting a non-parametric model to human eye-fixation data Kienzle, Franz, Schölkopf,

& Wichmann (2009). However, the model used in this experiment could not result in the concept of symmetry, as we propose in this chapter.

Although contrast has been the dominant feature for saliency models, a clear deficiency in the current visual attention models can be seen in Figure 2.5. The figure shows the results of the eye-tracking experiment presented in this chapter, as well as the results of two saliency models. For the images that are shown in the first column, the participants had a clear preference to fixate on the center of these symmetrical objects (last column). The response of the contrast-saliency model (Itti et al., 1998), shown in the second column, however, is much more spread out and not focused so much on the center of the object, but on the borders where the object contrasts with the backgrounds. The saliency model based on local symmetry that is proposed in this chapter, on the other hand, does more specifically predict the fixations on the center (third column). The results of this chapter show that this is true not only for photographic images that are selected explicitly to contain symmetrical objects as shown in the figure, but more generally for a wide variety of images containing natural and man-made content. Local symmetry calculations can thus be used to predict human gaze.

3.2.2 Symmetry in Vision

Symmetry is an abundant visual feature with esthetic properties. Humans very rapidly detect symmetrical patterns (Palmer & Hemenway, 1978) and recognition performances increase for symmetrical patterns (Royer, 1981). The perception of symmetry is suggested to be pre-attentive (Wagemans, 1999). Fixations on symmetrical forms are concentrated at the center of the form, or at the crossing points of the symmetry axes (Kaufman & Richards, 1969), or near the axis of symmetry (Locher & Nodine, 1987). This is similar to the tendency of eye saccades to land at the geometric center of a target object or target configuration (Findlay, 1982; He & Kowler, 1989; Ottes et al., 1984; Bindemann et al., 2009). The center-of-gravity of a pattern is approximately its center of symmetry. We propose to predict the tendency to fixate on the geometric center on the basis of symmetry. The advantage of using symmetry is that the symmetrical center can be determined without the need of prior segmentation of the objects in the scene. Since symmetry is a context-free cue for figure-ground segregation (Driver et al., 1992), it can be used for bottom-up object detection.

This chapter investigates whether symmetry can be used to predict the eye fixations of

humans watching complex photographic images with natural and man-made scenes.

A more detailed discussion on the role of symmetry in vision can be found in Section 2.6.

3.2.3 Fixation Sequence

When humans view an image for a couple of seconds, they make a sequence of saccades to investigate the interesting regions of the image. According to the scanpath theory of Noton & Stark (1971a,b), humans make the same sequence of fixations every time they view the same pattern. This suggests that the sequence is remembered as part of the pattern representation. This spatial memory is then used as top-down guidance for the eye movements. Since the focus is on the bottom-up components of eye movements, scanpaths are not considered in this chapter.

Parkhurst et al. (2002) compared human eye fixations in a free-viewing experiment with the contrast-saliency model (Itti et al., 1998). Investigating the amount of contrast near the point of fixation, they found that it drops over the fixation sequence. Earlier fixations are on parts of the image containing more contrast than the later fixations. Similarly, this chapter shows that the amount of local symmetry at the point of fixations also gradually drops over the fixation sequence. This effect is even larger for local symmetry than for contrast. The reason for the drop of contrast and symmetry at the points of fixation might be that the early fixations are more stimulus-driven than the later, since context then plays a larger role in the guidance of the eyes. However, it is also possible that all attended parts of the scene have above-average contrast and local symmetry, and the sequence is based on the strength of these features. Local symmetry, and to a lesser extent contrast, can then be used to predict the sequence of fixations. It must be noted, however, that this is only true in free-viewing conditions with no particular target. When participants are engaged in a search task, bottom-up saliency is not a good predictor of overt visual attention (Foulsham & Underwood, 2007).

3.3 Methods

In this section, we first present the symmetry-saliency model, and give a short overview of the contrast-saliency model of Itti et al. (1998) with which we compare the results

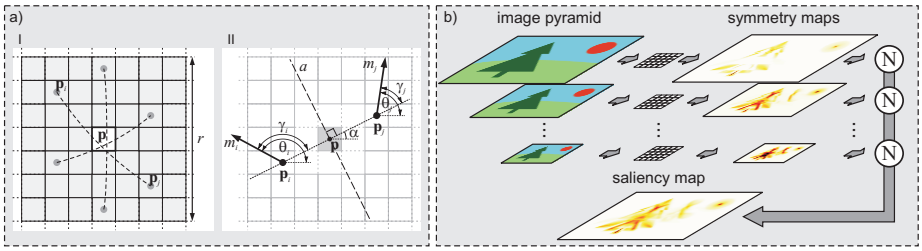

Figure 3.1: The multi-scale symmetry-saliency model. a) shows the basic symmetry operator. All pixel pairs in the symmetry kernel contribute to the local symmetry value of the central pixel (a1). The contribution of a pixel pair is calculated using the intensity gradients at the pixel locations (a2). b) gives the layout of the multi-scale symmetry model. A Gaussian image pyramid of five scales is constructed. The symmetry operator is applied to all images in the pyramid, resulting in symmetry maps at different scales. The maps are normalized and added to form the symmetry-saliency map.

as a point of reference. Subsequently, the eye-tracking experiment is explained, and the data presented. The section ends with a description of the two methods used to compare the human data with the saliency models.

3.3.1 Symmetry-Saliency Model

We developed three saliency models based on local symmetry calculations. The models are built upon the isotropic and radial symmetry operator of Reisfeld, Wolfson, & Yeshurun (1995), and the color symmetry model of Heidemann (2004). These three symmetry operators are all based on the same basic symmetry operator. We extended the operators to multi-scale symmetry-saliency models. We first describe the basic symmetry operator, followed by the multi-scale symmetry models.

3.3.1.1 Basic Symmetry Operator

The basic or *isotropic symmetry operator* calculates the amount of local symmetry at a given pixel, $\mathbf{p} = (x, y)$, in an image by applying a symmetry kernel to this pixel. Using a sliding window approach, the symmetry is calculated for all pixels. The amount of local

symmetry at p is calculated based on the intensity gradients of the surrounding pixels in the kernel. Pixels pairs in the symmetry kernel contribute to the local symmetry value. A pixel pair consists of two pixels, $\mathbf{p}_i$ and $\mathbf{p}_j$, so that $\mathbf{p} = (\mathbf{p}_i + \mathbf{p}_j)/2$ (see Figure 3.1a1). In other words, the two pixels forming a pair are point symmetric in the center of the kernel. The contribution of the pixel pair to the local symmetry of $\mathbf{p}$ is calculated by comparing the intensity gradient g_i at $\mathbf{p}_i$ and gradient g_j at $\mathbf{p}_j$. The intensity gradients are obtained by approximating the image derivatives in the horizontal, I_x, and vertical, I_y, direction using Sobel filters:

$$I_x = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} * I, \qquad I_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * I, \tag{3.1}$$

The gradient vector $g_i = (I_x(\mathbf{p}), I_y(\mathbf{p}))^T$, with the magnitude, m_i, and orientation, θ_i, determined as:

$$m_i = \sqrt{I_x(\mathbf{p})^2 + I_y(\mathbf{p})^2} \tag{3.2}$$

$$\theta_i = \text{atan2}(I_y(\mathbf{p}), I_x(\mathbf{p})) \tag{3.3}$$

Based on the orientation of the gradients at point i and j, the symmetry is measured by:

$$c(i,j) = (1 - \cos(\gamma_i + \gamma_j)) \cdot (1 - \cos(\gamma_i - \gamma_j)), \tag{3.4}$$

where $\gamma_i = \theta_i - \alpha$ is the angle between the orientation of the gradient, θ_i, and the angle, α, of the line between $\mathbf{p}_i$ and $\mathbf{p}_j$ (see Figure 3.1a2). The first term in Equation (3.4) has a maximum value when $\gamma_i + \gamma_j = \pi$, which is true for gradient orientations that are mirror symmetric with respect to $\mathbf{p}$. Using only this term would also respond to symmetry values for two pixels that have the same gradient orientation and thus lie on a straight edge. Since we are not interested in detecting edges, but in finding the centra of symmetrical patterns, the second term in the equation demotes pixels pairs with similar gradient orientations.

The symmetry measurement is weighed by a distance function and the magnitudes of

the gradients to get the local symmetry contribution of the pixel pair:

$$s(i,j) = d(i,j,\sigma) \cdot c(i,j) \cdot log(1+m_i) \cdot log(1+m_j), \tag{3.5}$$

where m_i is the magnitude of the gradient, and $d(i,j,\sigma)$ is a Gaussian weighting function on the distance between $\mathbf{p}_i$ and $\mathbf{p}_j$ with a standard deviation of σ. The multiplication with the gradient magnitudes assures that only strong edges contribute to the local symmetry value, since these are likely to belong to objects in the scene. The logarithm is used to attenuate the influence of large magnitude values.

The total symmetry value at point p is calculated by summing the contributions of all symmetrical pixel pairs in the kernel, $\Gamma(\mathbf{p})$ The symmetry kernel has a size of $r \times r$ (see Figure 3.1a2). We used $r = 24$ in our experiments. The amount of local symmetry calculated by the isotropic symmetry operator is then:

$$S_l^{\text{iso}}(\mathbf{p}) = \sum_{(i,j)\in\Gamma(\mathbf{p})} s(i,j), \tag{3.6}$$

where S_l^{iso} is the resulting isotropic symmetry map at scale l. The use of different scales to acquire a multi-scale symmetry-saliency model is discussed in the next section.

Based on this isotropic symmetry operator, Reisfeld et al. (1995) developed a *radial symmetry operator* that is extra sensitive to patterns containing multiple axes of symmetry. Due to the summation in Equation (3.6), the isotropic operator has already a higher activation for patterns with multiple axes of symmetry. However, the radial operator extra promotes these kinds of patterns. To achieve this, the orientation of the symmetry contribution of every pixel pair is calculated by:

$$\phi(i,j) = (\theta_i + \theta_j)/2. \tag{3.7}$$

Next, the pixel pair that contributed most to the symmetry value at point p is determined by:

$$(i',j') = \arg\max_{(i,j)\in\Gamma(\mathbf{p})} s(i,j) \tag{3.8}$$

and the symmetry orientation at point $\mathbf{p}$ is established:

$$\psi(\mathbf{p}) = \phi(i',j'). \tag{3.9}$$

This orientation is then used to promote the contributions of pixel pairs with dissimilar orientations:

$$S_l^{\mathrm{rad}}(\mathbf{p}) = \sum_{(i,j)\in\Gamma(\mathbf{p})} s(i,j)\cdot \sin^2(\phi(i,j)-\psi(\mathbf{p})). \tag{3.10}$$

Both the isotropic and the radial symmetry operator are based on the intensity of the pixels only. Heidemann (2004) extended the basic operator to a *color symmetry operator*. This operator compares pixels in three color channels, red, green, and blue, to determine the symmetry value:

$$S_l^{\mathrm{col}}(\mathbf{p}) = \sum_{(i,j)\in\Gamma(\mathbf{p})} \sum_{(k_i,k_j)\in K} c(i,j,k_i,k_j), \tag{3.11}$$

where K contains all combinations of two color channels, $K = \{(\mathrm{R},\mathrm{R}),(\mathrm{R},\mathrm{G}),\ldots,(\mathrm{B},\mathrm{B})\}$ $c(i,j,k_i,k_j)$ s the symmetry contribution calculated by comparing pixel $\mathbf{p}_i$ in color channel k_i with pixel $\mathbf{p}_j$ in color channel k_j. Besides the addition of color, Equation 3.4 is altered so that the function gives the same results for gradients that are rotated by 180° in order to account for patterns on gradually changing background:

$$c^{\mathrm{col}} = \cos^2(\gamma_i+\gamma_j)\cdot\left(\cos^2(\gamma_i)\cdot\cos^2(\gamma_j)\right). \tag{3.12}$$

The first term in the equation is a 180°-periodic symmetry term. The second term has a similar role as the second term in Equation 3.4, to discount for pixels that lie on an edge.

3.3.1.2 Multi-Scale Symmetry Model

The three basic symmetry operators discussed in the previous section calculate the symmetry response on one scale. Although a larger kernel size could in theory be able to detect larger symmetrical structures, there are two problems with that approach. Firstly, since two pixels at opposite sides of the kernels center are compared, the pattern needs to be perfectly symmetrical to have matching gradients at pixels far from the center. This will cause problems when using complex stimuli of real-world scenes like we do in our study. Secondly, larger symmetry kernels greatly increase the computational load of the algorithm.

To be able to detect larger symmetrical patterns and to allow for small deviations from

perfect symmetry and speed-up of calculation, we apply a multi-scale approach using Gaussian image pyramids (see Figure 3.1b), similarly to (Itti et al., 1998).

The image, I_0, at scale zero is at its original resolution (1024×768 pixels in our experiments). At subsequent scales, the image is first convolved with a Gaussian kernel, G, for low-pass filtering, and then downsampled to obtain an image that is half the width and height of the previous scale:

$$I'_{l-1} = I_{l-1} * G \tag{3.13}$$

$$I_l(x,y) = I'_{l-1}(2x,2y) \tag{3.14}$$

In our experiments, we used five different scales ($L = 5$). The resolution of the first scale, I_0, was 1024×768 pixels, and that of the highest scale, I_4, 64×48.

To determine the saliency map, the symmetry operator is applied to all Gaussian images in the pyramid. This results in L symmetry maps at different scales. These maps are combined by first normalizing the maps, then resizing them to the same scale ($l = 2$), and finally adding the different maps:

$$S = \bigoplus_{l=0}^{L-1} N(S_l), \tag{3.15}$$

where $\oplus$ is the summation operator that first resizes all elements to the same scale, and then sums the maps pixel wise.

The normalization function, N, is adopted from (Itti et al., 1998) and has the purpose to promote symmetry maps at scales with only a few outstanding points, as opposed to symmetry maps that contain many similarly symmetrical patterns. The normalization function first scales the values in the map to the range $[0,1]$, so that the global maximum has a value of 1.0, and then multiplies all values in the map with $(1-\bar{m})^2$, where $\bar{m}$ is the average value of all local maxima in the map that have a value greater or equal than 0.10. If there are many similarly symmetrical patterns, $\bar{m}$ will be large, and the map will thus be multiplied by a small value. If, on the other hand, there is one clear global maximum, $\bar{m}$ will be small, and the map will be weighed more strongly in calculating the total saliency map. Finally, the resulting saliency map will be normalized so that the total sum of all its elements is 1.0. Another normalization procedure based on lateral inhibition is discuss in (Itti & Koch, 2000). However, in our experience, that procedure results in too few salient locations. We try to predict eye fixations in a

free-view experiment with complex photographic stimuli where participants have many potentially interesting locations to focus on.

We designed our multi-scale symmetry-saliency model similarly to the multi-scale implementation of the contrast-saliency model (Itti et al., 1998) in order to provide a fair comparison of both methods.

3.3.2 Contrast-Saliency Model

We compare our symmetry-saliency model with the contrast-saliency model (Itti et al., 1998). In this section, a short overview of the contrast model is given to give the reader an idea of the mechanisms. For a full description, we refer to Appendix A.

The contrast-saliency model calculates saliency based on contrast in three different feature channels: intensity, color and orientation. Contrast is calculated by center-surround operations. The center is excited by the presence of a given feature, whereas the surround is inhibited, or vice versa. In the intensity channel, this corresponds to bright on dark or dark on bright. In the color channel, contrast is calculated using chromatic double-opponency channels, red on green, blue on yellow, or vice versa. Both color and intensity contrasts are implemented by using Gaussian image pyramids. The center-surround calculations are done by subtracting the image at different scales. The center is then taken as a pixel on a certain scale and the surround as the corresponding pixel on a coarser scale. For the calculation of orientation contrast, the Gaussian intensity images are convolved with Gabor filters in four different orientations. Again an image pyramid is constructed, and the center-surround orientation contrast is calculated by subtracting the Gabor-filtered images at different scales.

To obtain a multi-scale contrast-saliency model, contrast is calculated on three different scales, 2,3,4 (0 being the original resolution) and with a difference of both 3 and 4 scales between the center and the surround scales. The resulting *feature maps* on the different scales are normalized and combined similar to Equation 3.15, to form three conspicuity maps, for intensity, color, and orientation. To calculate the total contrast-saliency map, the conspicuity maps are first normalized using the earlier discussed normalization method, and then the average over the three maps is taken. Different from Itti, Koch, and Nieburs implementation, the resulting saliency map is at scale two, so that it is comparable with our symmetry-saliency map.

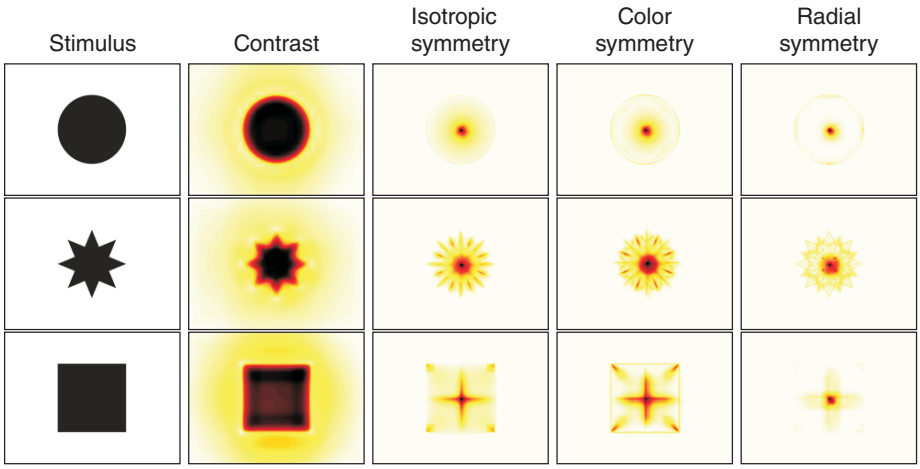

Figure 3.2: Examples of saliency maps produced by the three symmetry models and the contrast model as a response to the artificial stimuli. The color map goes from white (no response) to dark red (highest response). The contrast model has high response for the complete form. For the circle and square the highest points of activation are respectively near the edges and corners. The symmetry models, on the other hand, respond more specifically to the symmetrical center of the form, with the highest specificity for the radial symmetry model

Itti et al. (1998) ddiscuss a procedure to select a fixation location using winner-takes-all and inhibition-of-return operators. These operators are useful for modeling visual search, or to integrate bottom-up and top-down influences. However, since we are interested in the influences of saliency per se, we do not use this selection procedure, but rather compare the human fixations with the full saliency maps.

Some examples of saliency maps resulting from the symmetry models and the contrast model for artificial stimuli are given in Figure 3.2. There is a large difference between the symmetry and the contrast responses. Whereas the symmetry models specifically highlight the center of the objects, the contrast model gives a much more spread-out activation. For the circle and the square, the most salient points are even near the corners of the forms instead of at the center. The saliency map of the radial symmetry model

is a little more focused on the center than those of the other symmetry models. Apart from that, the differences among the three symmetry models are relatively modest.

3.3.3 Eye-Tracking Experiment

To test the performance of both the symmetry and the contrast saliency model, we conducted an eye-tracking experiment to record eye fixations while participants viewed complex photographic images. The experiment is discussed in this section.

3.3.3.1 Participants

31 students (15 female, 16 male) of the University of Groningen took part in the experiment for credit points. The age of participants ranged from 17 to 32. All had normal or corrected-to-normal vision.

3.3.3.2 Stimuli

A total of 99 photographic images in five different categories were presented to the participants. 19 images were in the natural-symmetry category. These images were selected explicitly for containing symmetrical natural objects. To test if our methods are not only valid for scenes containing explicit symmetrical forms, but more generally for a wide range of images, we included four other categories in the image set: 12 images of animals in a natural setting, 12 images of street scenes, 16 images of buildings, and 40 images of natural environments. Figure 3.3 gives examples of the different categories included in the dataset. The five categories span a wide variety of images, containing natural symmetries and natural and cultural scenes, with organic and rectilinear shapes. All these images were taken from the McGill calibrated colour image database (Olmos & Kingdom, 2004).

The images were displayed full-screen with a resolution of 1024×768 pixels on an $18''$ CRT monitor of 36 by 27 at a distance of 70 cm from the participants. The visual angle was approximately 29° horizontally by 22° vertically.

Figure 3.3: Image examples for all five categories used in the experiment. In total, 99 images were used: 19 images of natural symmetries, 12 of animals, 12 of street scenes, 16 of buildings, and 40 of natural scenes.

3.3.3.3 Experimental Setup

Since we are interested in the bottom-up components of visual attention, the participants were asked to freely view the images. We did not give them a task, since that would give a strong bias on the eye movements. Still, the eye movements are likely to be also controlled top-down, by interests and experiences of the participants. We will discuss our method to capture the consensus among participants in the next subsection.

The images were presented in random order to the participants. Each image was displayed for five seconds. After each presented image, the participant could decide when to continue. The experiment was split up in sessions of approximately five minutes. Between the sessions, the participants had a short break, in which the experimenter had a relaxing conversation to keep the participants motivated and focused.

Figure 3.4: The distribution of human eye fixations for the different categories. The contour plots show the normalized kernel-density estimation of the fixations of all participants for all images in the category. The distributions of the natural-symmetry and animal category are biased to the center, whereas the fixations on the street scenes and buildings are more uniformly distributed.

3.3.3.4 Eye Tracker and Data Acquisition

We used the Eyelink I head-mounted eye-tracking system (SR research) to record the gaze of the participants. Fixations were extracted using the accompanying software. At the beginning of the experiment, the eye tracker was calibrated using the SR-research software. Before every session, the calibration was verified and the experiment continued when the system was correctly calibrated. If not, the eye tracker was recalibrated. Before every trial, i.e., before every presentation of an image, drift was measured by letting the participant focus on a cross displayed in the center of the screen, and the estimation corrected if necessary. Because of the drift correction method, the first fixation was strongly biased. We therefore eliminated this fixation from the data. Using the eye tracker, we acquired 99 trials of five seconds for all 31 participants. A few trials were not used in the data analysis due to interruptions or other incidents.

3.3.3.5 Eye-Tracking Data

On average, the participants made 15.6 ($\pm$3.6) fixations while viewing an image for five seconds. The normalized distributions of fixations for the five categories are given in Figure 3.4. The figure shows that the fixations are not uniformly distributed of the image, but biased towards the center. The figure shows that the fixations are not uniformly distributed over the image, but biased towards the center. The standard deviation of the angular distance from the fixations to the center is respectively 8.0°, 8.2°, 9.0°, 9.1° and 8.6° for the images of natural symmetries, animals, street scenes, buildings, and natural scenes. This center bias is expected to occur in free-viewing experiments

(Tatler, 2007), and might be a result of both the tendency of photographers to place the important objects near the center, and the tendency of humans to center the eyes. In our data, the center bias is stronger for the natural symmetry and animal images, and weaker for the images of street scenes and buildings. In the center-bias experiment, discussed in the results section, we incorporate different strengths of center bias in the saliency maps to investigate the influence of a center bias in predicting human eye fixations.

3.3.3.6 Center Bias

As can be seen in Figure 3.4, the human eye fixations are biased towards the center of the image. To investigate the role of a center bias on the comparison between the saliency models and the human data, we include a center bias in the models similar to (Parkhurst et al., 2002). To do so, the values in the saliency map, S, are weighted with a two-dimensional Gaussian distribution with its mean at the center of the image, and a standard deviation, σ_b, that determines the strength of the center bias, with small values corresponding with strong center bias:

$$S'(\mathbf{p}) = S(\mathbf{p}) e^{-\|\mathbf{p}-\mu\|^2/2\sigma_b^2}, \tag{3.16}$$

where $\mathbf{p}$ is the location of a pixel in the map and $\mu = (512.5, 384.5)$ is the center of the image. The resulting central-biased saliency map, S', is normalized so that the total sum is 1.0.

3.3.4 Comparison Methods

We used two methods to compare the human eye-fixation patterns with the predictions from the saliency models. A *correlation method* similar to that used in (Le Meur et al., 2006; Ouerhani et al., 2004) and a *fixation-saliency method*, similar to that used in (Parkhurst et al., 2002). Both methods are discussed in this section.

3.3.4.1 Correlation Method

To correlate the human data with the output of the saliency models, we transform the eye-fixation data to *fixation-distance maps* (see Figure 3.5). These fixation-distance

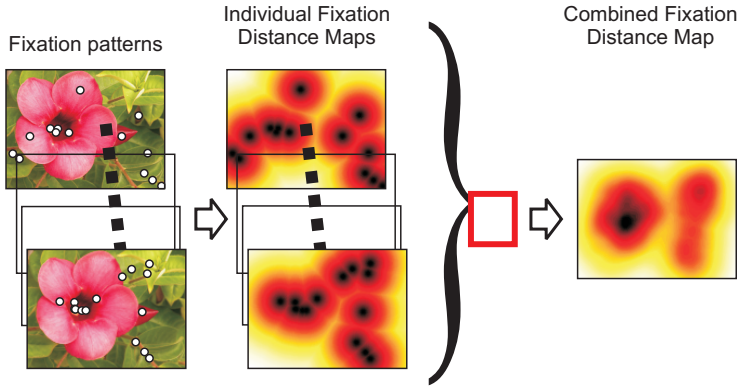

Figure 3.5: The fixation patterns of individual participants viewing an image are transformed to individual fixation-distance maps using the inverse distance transform. The summation of the individual maps results in the combined fixation-distance map. It can be appreciated that there is substantial variation in the individual fixation patterns. However, some fixations are shared among the participants. This consensus becomes clear in the combined fixation-distance map.

maps give the probability that a fixation lands on a certain location based on the human data. Similarly, the saliency maps can be seen as giving the probability of a fixation on that location based on the saliency models. To construct a fixation-distance map from an eye-fixation pattern, the inverse distance transform of the fixation data is calculated. The distance transform, F', gives the distance to the nearest fixation for all pixels in the image. This results in values of zero at the points of fixation with a linear increase at pixels further away from the fixations:

$$F'(\mathbf{p}) = \|\mathbf{p} - \mathbf{f}_n\|, \tag{3.17}$$

where $\mathbf{p} = (x, y)$ is the pixel location, $\mathbf{f}_n = (x_n, y_n)$ is the location of the nearest human-fixation point, and $\|\|$ is the Euclidian distance between the two. Next, the fixation-distance map, F, is obtained by subtracting all values from the maximum value in the distance transform:

$$F(\mathbf{p}) = \max(F') - F'(\mathbf{p}) \tag{3.18}$$

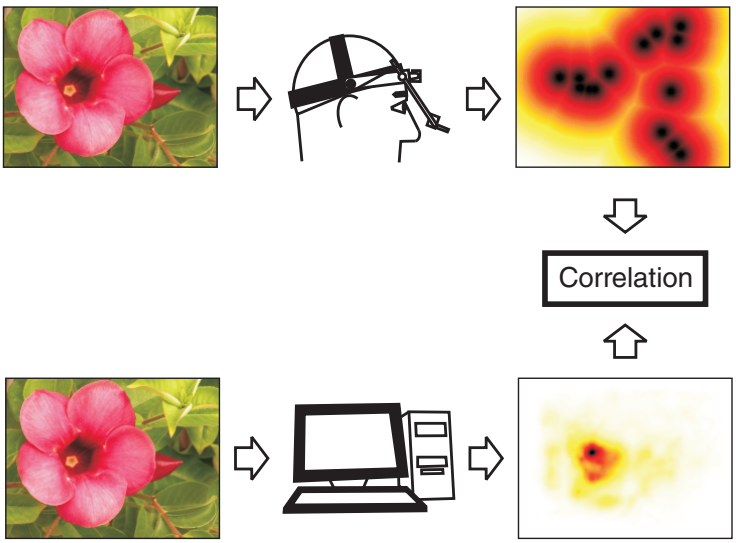

Figure 3.6: The correlation method. The fixation-distance map obtained from the human eye fixations is correlated with the saliency map calculated from the same image. The correlation results in a correlation coefficient that shows how well the saliency model predicts the human data.

F is normalized so that the sum of its elements is 1.0. This results in a map with high values at the points of fixations, and lower values further from these points. This is slightly different from the approach in (Kootstra et al., 2008a; Le Meur et al., 2006; Ouerhani et al., 2004), where a fixation density map is calculated using Gaussian kernels. Our method puts emphasis on the location of fixations, rather than on their density. Moreover, this correlation method is parameter free, i.e., there is no width of the kernel to be set.

In Figure 3.6, the correlation method to compare the saliency maps with the fixation-distance maps is depicted. The two maps are correlated with each other to get the correlation coefficient, ρ:

$$\rho = \frac{\sum_{\mathbf{p} \in P} ((F(\mathbf{p}) - \mu_F)(S(\mathbf{p}) - \mu_S))}{\sqrt{\sigma_F^2 \sigma_S^2}}, \tag{3.19}$$

where P is the set of all pixel coordinates and μ and σ^2 are respectively the mean and the variance of the values in the maps. The correlation coefficient has a value between 1 and 1. A ρ of 0 means that there is no correlation between the two maps, which is true when correlating with random fixation-distance maps. Values for ρ close to zero indicate that a model is a poor predictor of human fixation locations. Positive correlations show that there is similar structure in the saliency map and the human fixation map.

In the above described correlation method, the predictions of the saliency models are compared to the fixation-distance maps of individual participants. However, the photographic images viewed by the participants are highly complex stimuli that generate many fixations, with substantial variation among the participants. Because of this variation, the correlations of individual fixation-distance maps with the saliency maps will be low. However, some of the fixations are shared by all participants, and are more likely to be caused by bottom-up factors. Because we are interested in general models and not in models that predict visual attention of specific persons, we want to test how well the saliency models predict the consensus among participants as well. To test this, we calculate the correlation coefficient for the *combined fixation-distance maps* (Figure 3.5). These combined maps are calculated by summing the individual fixation-distance maps:

$$F_c = \sum_{i=1}^{N} F_i, \tag{3.20}$$

where F_i is the individual fixation-distance map for participant i, F_c the combined fixation-distance map showing the consensus, and $N = 31$. F_c is normalized so that the elements sum up to 1.0. The saliency maps are compared tot the combined fixation-distance maps using Equation (3.19).

3.3.4.2 Fixations-Saliency Method

The fixation-saliency method tests how the saliency at the points of human fixation according to the saliency methods compares to the average saliency for the image (see Figure 3.7). Whereas the correlation methods looks at both the presence and the absence of fixations and saliency, the fixation-saliency method focuses on the locations where there actually are eye fixations. With the method, we can investigate whether the local symmetry and contrast at the fixation points are above average. Moreover, it gives

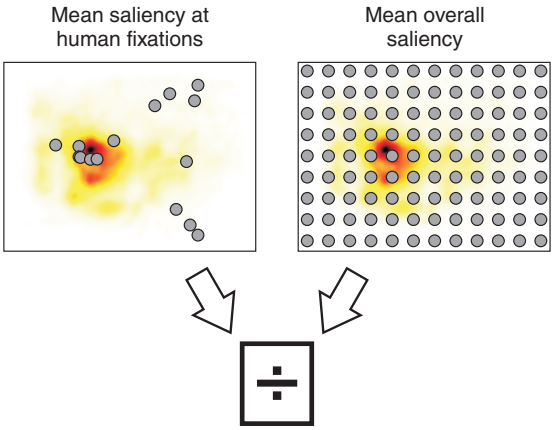

Figure 3.7: The fixation-saliency method. The saliency, as calculated by the saliency models, is measured in a patch around the fixation points. The mean saliency for the human fixations is divided by the average saliency in the image, resulting in the fixation-saliency score

the possibility to investigate the progression of saliency over the fixation sequence.

The fixation-saliency score, λ, is calculated by calculating the average saliency according to the saliency model in a patch around the points of fixation divided by the average saliency over a large number of random points, $K = 1000$. For a given participant and image, that is:

$$\lambda = \frac{K}{M} \frac{\sum_{i=1}^{M} s(\mathbf{f}_i)}{\sum_{j=1}^{K} s(\mathbf{r}_j)}, \tag{3.21}$$

where $\mathbf{f}_i$ is the ith fixation of a total of M fixations, $\mathbf{r}_j$ is a randomly generated point in the image, and $s()$ gives the average saliency in a patch around the point:

$$s(x,y) = \frac{1}{(2R+1)^2} \sum_{j=-R}^{R} \sum_{i=-R}^{R} S(x+i, y+j), \tag{3.22}$$

where $R = 28$ pixels. When $\lambda > 1$, the saliency is higher at the fixation points than

average, showing that the given saliency model can predict the fixations to some extent.

The above method calculates the average fixation saliency for all fixations in the sequence. To investigate the progression in saliency over the sequence, we calculate the fixation saliency for individual fixations, λ_i:

$$\lambda_i = K \cdot s(\mathbf{f}_i) / \sum_{j=1}^{K} s(\mathbf{r}_j). \tag{3.23}$$

3.4 Results

In this section we discuss the results of the comparison of the symmetry and contrast saliency models with human eye fixations. We firstly show the results of the correlation and fixation-saliency methods on the fixation patterns of individual participants viewing an image. Secondly, we discuss the results of the correlation comparison with the fixations of all participants combined. Next, the saliency over the fixation sequence is shown. Finally, an analysis of the center bias is discussed.

3.4.1 Individual Fixation Patterns

3.4.1.1 Correlation

In Figure 3.8 the results of the correlation between the individual fixation-distance maps and the saliency maps are given. The five groups of bars contain the results for the different image categories. The bars show the mean correlation coefficients, ρ, over all participants and images in the category for the different saliency models. The error bars give the 95% confidence intervals on the mean. The scores of the saliency methods are plotted along with the inter-participant correlation, and the correlation of the human data with random fixations. The first, which indicates how well one persons fixations correlate with those of the others, is depicted by the horizontal gray bar with a solid mid-line, giving the mean and 95% confidence interval. The correlation with random fixations is depicted by the horizontal dashed line, which is, as expected, virtually zero for all categories. All means and confidence intervals in this paper are calculated using multi-level bootstrapping. Significant differences can be appreciated by looking at the 95% confidence intervals.

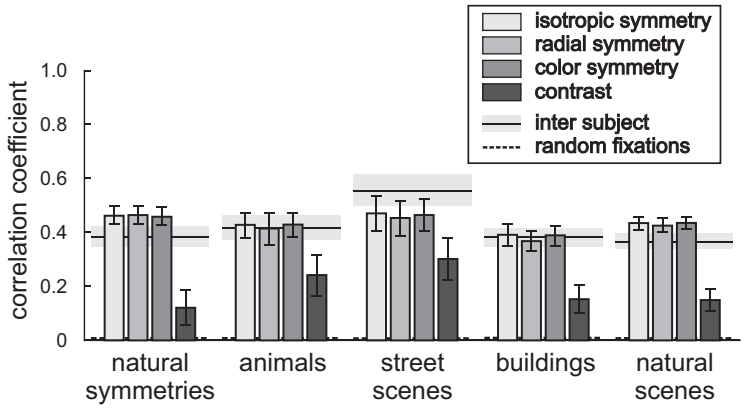

Figure 3.8: Correlation between the saliency maps and the individual fixation-distance maps. The groups of bars relate to the different image categories. The bars give the mean correlation coefficients. The error bars are the 95% confidence intervals. The horizontal gray bars with the solid line show the mean and 95% confidence interval of the inter-participant correlation. The correlation of the human data with random fixations are given by the dashed lines, which are close to zero. It can be appreciated that the symmetry models significantly outperform the contrast model, not only on the natural-symmetry category, also on the other categories.

The inter-participant correlation is calculated for every image by correlating the fixation-distance maps of every participant with those of all other participants, resulting in a similarity measure among participants. The plot shows that there is variability among the participants. The saliency methods are also faced with this variability, which pulls down the correlation values. The inter-participant correlation can therefore be used to put the scores of the saliency methods into perspective. It must be noted that the correlation scores of the models can be higher than the inter-participants scores when the variation among participants is high. The models can then predict the consensus among the participants better than the participants themselves can.

Figure 3.8 clearly shows that the symmetry models compare significantly better with the human data than the contrast models for the images containing natural symmetries. This is as expected, since the images were selected on the basis of symmetry. Moreover, also for the other categories the correlation scores are significantly higher for the symmetry models than for the contrast model. This suggests that the symmetry models

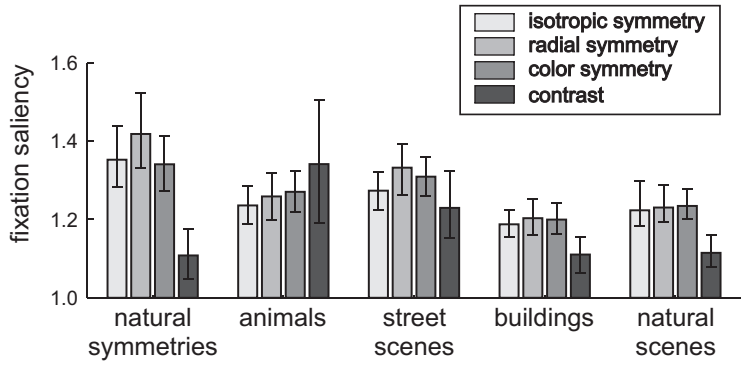

Figure 3.9: The fixation-saliency results. The bars give the mean saliency near the eye fixations relative to the average saliency in the complete image. The 95% confidence intervals are given by the error bars. The local symmetry at human points of fixation is significantly higher than the contrast at these points for most of the categories, except for the animal images.

have general validity. The performance of the symmetry models is in the same range as the inter-participant correlations. The performance of the contrast model correlates with the inter-participant score. High inter-participant scores reflect that the individual fixation patterns are more similar, presumably because there are fewer interesting locations for the participants to focus on. The contrast model scores better in these cases than it does when there is more variability among the participants. The performance of the symmetry models, on the other hand, is significantly better for all image categories, and they seem to predict the consensus among participants better even when there is more variability. Among the three symmetry models, isotropic, radial, and color, we do not see significant differences in performance.

3.4.1.2 Fixation Saliency

If we look at the fixation-saliency value, λ, in Figure 3.9, we see that both contrast and symmetry are higher at the points of fixation than in the rest of the images ($\lambda > 1$), showing that both can predict eye fixations to some extent. Especially for the natural-symmetry category, the symmetry models score significantly better than the contrast model. Also for the other categories, except for the animal category, symmetry scores

significantly better, with an exception for the isotropic model in the street-scene category, which scores better, but not significantly ($\alpha = 0.05$). The animal category gives a different result. There the contrast model scores better than the symmetry model, although not significantly. This result might be explained by the fact that, in contrast with the images in the other categories, many images in this category contain objects animals that are highly distinguishably and sharply depicted on an out-of-focus background. The fore- and backgrounds in the other images are less distinct and more cluttered. In the animal images, there are fewer interesting locations and the background also has less contrast. Among the different symmetry-saliency models, there are no clear differences with the exception of the radial symmetry model on the natural-symmetry category, which scores somewhat better than the others.

The results of both the correlation method and the fixation-saliency model show that, in general, the symmetry-saliency model predicts the eye fixations significantly better than the contrast-saliency model.

3.4.2 Combined Fixation Patterns

In Figure 3.8, the saliency maps are correlated with the individual fixation-distance maps. Because there is much variety in the fixation patterns among the participants, the correlation scores are relatively low. Some of the locations in the images, however, are attended by most participants. To investigate how well this consensus is predicted by the saliency models, we combined the fixation-distance maps of the individual participants. The correlation coefficients, ρ, of this analysis are given in Figure 3.10. The bar plots show a similar structure as that in 3.8: the symmetry models significantly outperform the contrast model. However, the correlation coefficients went up from around 0.4 to around 0.7 for the symmetry models. This shows that the symmetry models do a good job in predicting the fixation consensus among the participants. Again, this is not only true for the images containing explicit symmetrical forms, but for all categories. This shows that the common fixations of the participants are well captured by the symmetry-saliency models.

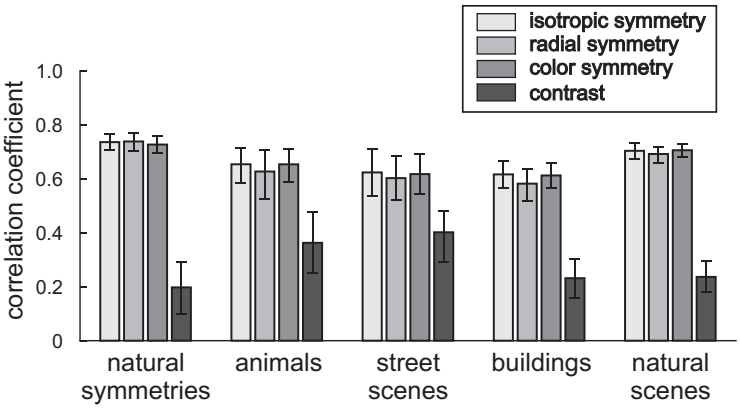

Figure 3.10: Correlation between the saliency maps and the combined fixation-distance maps, representing the consensus among the participants. The bars and error bars give the mean and 95% confidence intervals on the mean of the correlation coefficients. The results show the same pattern as for individual fixation-distance maps, with significantly higher scores for the symmetry models. However, the correlation coefficients are much higher, showing a better fit of the models with the participants consensus.

3.4.3 Fixation sequence

In the above, we compared the complete five seconds fixation sequence with the saliency models. In Figure 3.11, the progression of fixation saliency, λ_i, as a function of the fixation number is shown. It can be appreciated that the symmetry is especially high for the first fixation, and gradually drops for later fixations. This shows that the participants first attend highly symmetrical parts of the image. The contrast at the points of fixation, however, is lower, and runs much more stable over the sequence, except for the animal condition. The difference between the symmetry models and the contrast model is significant for the first fixations for all categories except for the animal images. For later fixations, the difference is less apparent, but still in favor of the symmetry models, and significant for the nature category. The results for the animal condition are different. The fixation saliency of the models is not significantly different and higher for contrast. Contrast is here also high for the first fixations, and lower for later. This might again be explained by the different style of the photos compared to the other categories.

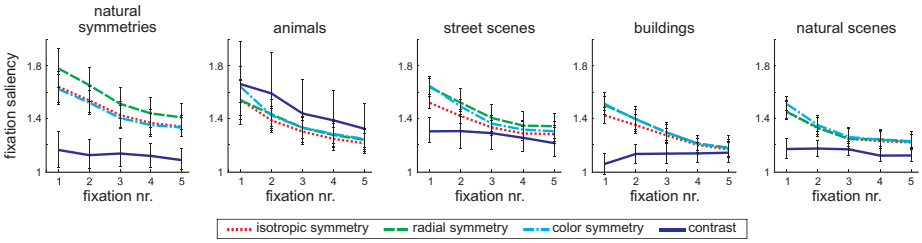

Figure 3.11: Fixation saliency, i.e., the saliency as measured by the models near the human fixation points, over the fixation sequence. The fixation saliency is plotted as a function of time measured by the fixation number. The lines give the mean fixation-saliency scores, and the error bars the 95% confidence intervals on the mean. The symmetry at fixation points is especially high for early fixations, and drops for later, showing that the fixations can be order on the basis of symmetry. The contrast at fixation points is lower than symmetry and is more constant over the sequence, except for the animal category, where the contrast model shows a similar result as the symmetry models.

3.4.4 Central Bias

In order to test whether the performance of the models is influenced by the center bias that the participants displayed Figure 3.4, we added a center bias to the saliency maps as explained in the methods section. Figure 3.12 shows the correlation coefficients as a function of the center-bias strength, σ_b, where the combined fixation-distance maps are compared with the center-biased saliency maps. The curves of the contrast-saliency model show a maximum correlation value for σb between 6° and 9°. The maxima are at 6°, 7°, 9°, 8°, and 7° for respectively the natural-symmetry, animal, street-scene, building, and natural-scene category. This is similar to what is reported in (Parkhurst et al., 2002). The curves of the symmetry-saliency models, on the other hand, do not show a maximal value. They gradually grow when the center-bias is weakened and reach an asymptote between 12° and 15°. The maximally performing central-bias standard deviation for the contrast model is somewhat lower than the standard deviations on the distance to the center as observed in the human data (see subsection about eye-tracking data). The results show that the contrast model can be improved using a center bias, whereas the symmetry models give better results without such a bias. The performances of the symmetry models are significantly better than that of the contrast

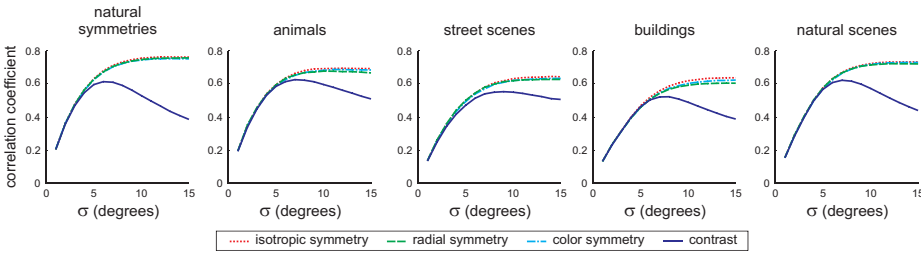

Figure 3.12: The influence of a center bias added to the saliency maps on the correlation coefficients. The plots give the coefficients for the comparison of the human data with the center-biased saliency maps. The curves give the mean correlation coefficients. The curves for the contrast model show a clear peak for a center bias with σb between 6° and 9°. The symmetry models, on the other hand, show no peak and even increase in correlation with the human fixation distance maps when the center bias is relaxed.

model, even when the center bias is applied. There is virtually no difference among the different symmetry models. The fact that the performance drops for the contrast model when the center bias is weakened, suggests that the model incorrectly predicts eye fixations at the periphery of the images. The symmetry models, however, seem to predict valuable fixations in the periphery, since the performance increases, even for standard deviations higher than those observed in the human data. It is therefore better not to apply a center bias to the predictions of the symmetry models.

3.5 Discussion

In this chapter, three saliency models for the prediction of human eye fixations based on local symmetry were presented. The models were compared to a saliency model that is based on contrast features (Itti et al., 1998). To test the models, we conducted an eye-tracking experiment using a wide variety of different images. The results show that the symmetry-saliency model compares substantially better with the human data than the contrast-saliency model.

The analysis of the correlation between the models predictions and human fixations shows significantly better performance for the symmetry models, not only for the im-

ages containing explicit symmetries, but for all image categories. The comparison with the combined fixation-distance maps shows that the models capture the fixation consensus among the participants particularly well. This suggests that local symmetry can be used as a general model for the prediction of human eye fixations.

The analysis of the fixation saliency gives similar results. The amount of local symmetry at the points of human eye fixation is well above average and exceeds the contrast at fixation points for most image categories except for the animal images. Moreover, the amount of symmetry at the points of fixation is especially high for the first fixations with a gradual drop for later fixations. The contrast saliency shows a flat curve over the fixation sequence. This suggests that humans attend to locally symmetrical regions in an image, and moreover that symmetry can be used to order the fixation sequence.

The distribution of the human fixation data is a little biased towards the center. The addition of a center bias results in a maximum performance for the contrast model at a slightly stronger bias than in the human data. The performance of the symmetry models, on the other hand, does not have a maximum, but grows for weaker center biases. This suggests that the symmetry models find valuable salient points in the periphery, which are attended to by the human observers. The contrast model, on the other hand, suggests salient points in the periphery that do not correspond to human fixations.

The fixation saliency of the contrast model is different for the images in the animal category than for the other categories. The main difference between the categories is that the images in the animal category depict the animals clearly and sharply on blurred and out-of-focus backgrounds. This results in high contrast between foreground and background. The other categories contain images with sharper and more cluttered backgrounds.

The experiments reveal no significant difference among the three symmetry models, whereas the radial symmetry model was expected to perform better since humans are also more sensitive to patterns with multiple axes of symmetry. However, the isometric symmetry model already results in higher activation for these kinds of patterns. The extra promotion of multiple symmetry axes apparently only slightly changes the symmetry saliency maps. Similarly, the addition of color does not result in substantial changes in performance as well.

Although the performance of the contrast models in the presented experiments is less than that of the symmetry models, contrast obviously also plays a role in visual atten-

tion. Both the correlation and the fixation saliency of the contrast model are well above chance levels, conform the findings of for instance (Le Meur et al., 2006; Parkhurst et al., 2002; Parkhurst & Niebur, 2003). Moreover, the symmetry models also exploit contrasts in the image gradients to determine symmetry. The main difference between the symmetry and contrast model is the specificity, as can be seen in Figure 2.5 and Figure 3.2. The contrast model gives a more spread-out activation less focused on the centra of objects. This reduces the similarity to the human data. It makes sense to combine the symmetry and contrast model to further improve the prediction of eye fixations.

Both analysis methods show a correlation between local symmetry and human eye fixations. However, that does not prove that there is a causal relation between symmetry and overt visual attention. However, we think that a causal relation is likely, especially considering that symmetry can be used for figure-ground segregation. We discuss this further in Chapter 5.

In (Findlay, 1982; He & Kowler, 1989; Kaufman & Richards, 1969; Ottes et al., 1984) eye fixations are reported to land at the center of gravity of objects. A center of gravity is strongly correlated to the center of symmetry of an object. Our research therefore suggests that the center-of-gravity effect is not only true for simple artificial stimuli like the ones used in the above-mentioned studies, but also for complex photographic images of natural and man-made scenes.

In reality, fully symmetrical forms are almost never observed. When an object is viewed from a nonorthogonal angle, its appearance in the two-dimensional projection is not perfectly symmetrical. This is termed *skewed symmetry*. Although this is true, still the amount of local symmetry as calculated by our model is higher for the skewed symmetry than for other random configurations in the image. On a side note: Humans have more difficulties perceiving skewed symmetry, but are capable of doing so with sufficient depth cues Wagemans (1993).

As pointed out in (Land & Hayhoe, 2001; Schumann, Einhäuser-Treyer, Vockeroth, Bartl, Schneider, & König, 2008), natural human behavior might be different from the behavior observed in lab experiments. We therefore think it is interesting to study the role of symmetry in overt visual attention during natural behavior such as playing cricket (Land & McLeod, 2000) or making tea (Land & Hayhoe, 2001). In a dynamic setting, symmetry in motion can also be used.

To conclude, our results suggest that symmetry plays a role in the guidance of eye movements, either directly or indirectly by being a cue for the presence of objects. We advocate the study of the role of symmetry in human vision.

Does Symmetry Result in a Pop-Out?

Abstract

The previous chapter suggests that symmetry attracts attention. The analyses revealed a correlation between human eye fixation and local symmetry. However, it does not mean that there is a causal relation between the two. To shed more light on this matter, the role of symmetry in attracting visual attention is investigated in more controlled experiments in this chapter. In the first experiment, it is tested to what extent the perception of symmetry is efficient and preattentive in a pop-out experiment. The second experiment examines whether symmetrical figures attract more attention than non-symmetrical figures in a scene-memory task. The results of the first experiment show that symmetry causes a pop out, especially for polygonal figures. For the stacked bar and dot patterns, search for the target is less efficient. However, some elements of a pop-out effect are found. This suggests that the detection of symmetry is preattentive. The second experiment shows no effect of symmetry. This shows that visual attention is mainly object oriented.

4.1 Introduction

The previous chapter has shown that symmetry is a good predictor of human eye fixations. The local symmetry in the image correlates well with the fixation locations. Moreover, the amount of local symmetry is higher at the fixation points, especially for the first fixations. However, this shows a correlation, but not a causal relation between symmetry and visual attention. Does symmetry directly attract attention, or is symmetry a cue for the presence of an object and is visual attention basically object oriented? This is one of the questions that this chapter tries to answer. The other question is whether symmetry is a feature that is processed efficiently and preattentively. The results in the previous chapter suggest that this is the case, since especially the first fixations are on locally symmetrical parts of the image. In this chapter, these questions are addressed in two experiments, a pop-out experiment and a scene-memory task.

4.1.1 Effortless Symmetry Perception

Julesz (1971, 1981) defined perception as efficient and preattentive when important characteristics of a stimulus can be detected when the stimulus is exposed for a very brief period ($<$ 160ms). In this definition, symmetry detection of single presented figures has been found to be preattentive. Humans are able to detect symmetry in simple shapes that are very briefly presented (Carmody, Nodine, & Locher, 1977), as well as in dot patterns (Wagemans, Van Gool, & d'Ydewalle, 1991), line segments (Locher & Wagemans, 1993), and abstract art displays (Locher & Nodine, 1989). Especially symmetry at a coarse scale can be detected very quickly (Barlow & Reeves, 1979; Royer, 1981; Palmer & Hemenway, 1978). It is proposed that symmetry detection takes place in two phases, a quick but coarse first phase, followed by a slower second phase, in which a more detailed investigation of the pattern takes place (Palmer & Hemenway, 1978). Other evidence for preattentive detection of symmetry comes from Baylis & Driver (1994), who demonstrated that the reaction times to detect symmetry are hardly influenced by the complexity of the pattern. This suggests parallel and preattentive perception of symmetry Wagemans (1999).

Another method to investigate preattentive processes is by using visual-search experiments such as discussed in Section 2.3. If the search for an object that differs from the other objects in a search display on a specific feature is efficient, the perception of

that feature is likely to be preattentive. The search for a feature is efficient when the slope of the reaction time $\times$ set size curve is near zero, in other words when the target pops out despite the number of distractors. A further indication of preattentive feature detection is the existence of a search asymmetry (Wolfe, 1998; Treisman & Gormican, 1988). According to Treisman & Gormican (1988), it is easier to find a deviation among canonical stimuli than the reverse. It is, for instance, easier to find a tilted bar among vertical distractors than it is to find a vertical bar among tilted bars, and it is easier to find an up-side-down elephant among normally oriented elephants than vice versa (Wolfe, 2001).

In Section 2.3.1, we discussed a number of basic features that cause a pop out. These features include color, intensity, orientation, motion, and shape. Search asymmetries have been found for these features as well (Wolfe, 1998). Only a few studies focused on symmetry as a basic feature. Wolfe & Friedman-Hill (1992) showed that symmetry relations among items in a search display influence the reaction times. If the distractors are symmetrical to each other, search is easier. If, on the other hand, some of the distractors are symmetrical with the target, search becomes more difficult. This suggests that symmetry relations are processed in parallel and preattentively.

Olivers & van der Helm (1998) used more complex symmetrical and non-symmetrical stimuli in a pop-out experiment. They used dot patterns, polygons, and block-contour shapes and presented search displays with a symmetrical target among non-symmetrical distractors and vice versa. Their results showed that symmetry does not result in a pop-out effect. The reaction times increase when the number of items in the search display increase. However, in their experiments, the set of distractors was highly heterogeneous. All distractors had a completely different shape, with the only common property that they were either symmetrical among the vertical axis or not. However, Bauer et al. (1996) showed that search becomes inefficient when the distractors are sufficiently heterogeneous. This might explain the results of Olivers & van der Helm (1998).

In our first experiment, we investigate the preattentive perception of symmetry using the visual search paradigm. Our experimental setup differs from that of Olivers & van der Helm (1998) in that we use a homogeneous set of distractors. We furthermore use set sizes of 4, 8, and 12 items per search display, instead of the 1–4 items used by Olivers & van der Helm. This is more conform the set sizes used in other visual-search experiments (see Section 2.3).

4.1.2 Attention for Symmetry

The first experiment presented here studies whether the perception of symmetry is efficient and preattentive. If so, this does not mean that symmetry is also a visual attractor if the task is not to spot the odd-one-out. To investigate whether attention is paid to symmetry outside the visual-search paradigm, we ask the participants to remember a scene consisting of a number of symmetrical and non-symmetrical figures in the second experiment. The question is whether people pay more attention to the symmetrical forms than to the non-symmetrical ones to memorize the scene. If the results confirm this question, symmetry is a stronger visual attractor than the mere presence of an object. If not, it indicates that human visual attention is essentially object oriented.

4.2 Symmetry Pop-Out Experiment

In the symmetry pop-out experiment, the participants have to search for the odd-one-out, a symmetrical figure among non-symmetrical figures or vice versa.

4.2.1 Methods

4.2.1.1 Participants

Eleven undergraduate students participated in the experiment for course credits. The age of the participants ranged from 18 to 25 years. All participants had normal or corrected to normal vision. The data of one participant is discarded, since the mean reaction times exceeded the means of the other participants by three standard deviations.

4.2.1.2 Experimental Design

The participants were shown visual-search displays. The task was to decide as quickly as possible if the search display contained a figure that differed from the rest or not by pushing buttons on a keyboard. The reaction time was recorded.

Three different stimulus types were used, stacked blocks, polygons, and dot patterns (see Figure 4.1a). Figure 4.1b shows the experimental layout. In fifty percent of the

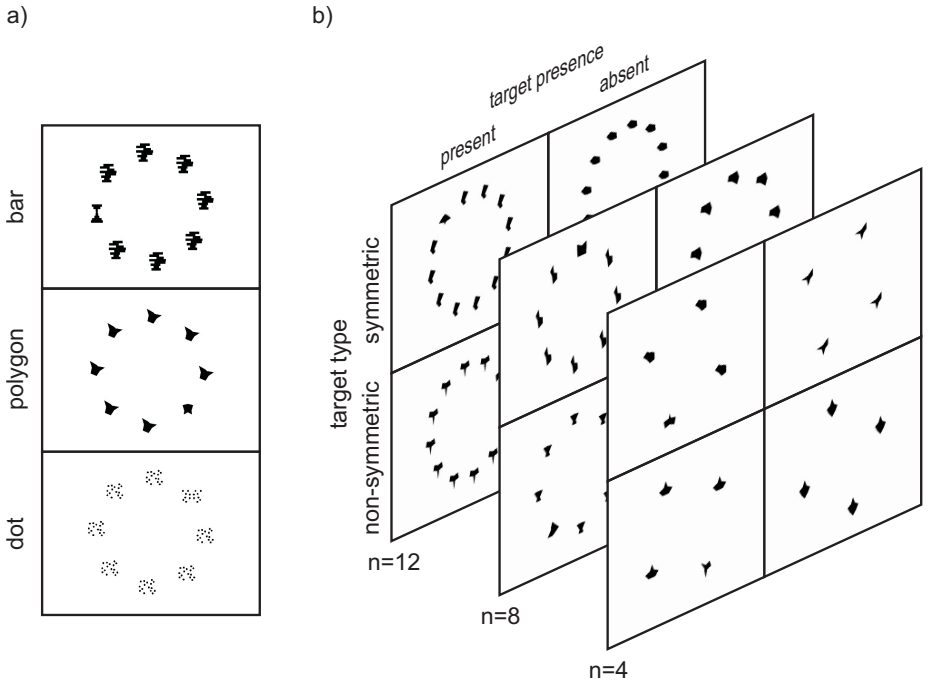

Figure 4.1: The experimental setup of the pop-out experiments. a) The three different types of stimuli used: blocks, polygons, and dot patterns. b) For each stimulus type, a $3 \times 2 \times 2$ layout was used: three different set sizes, two different conditions (symmetric and non-symmetric), and target presence.

trials, a target was present in the display. In the other cases, the target was absent, and all figures in the display were identical. We used two conditions to investigate search asymmetries. In the symmetric condition, the target was symmetric and the distractors were non-symmetric, and in the non-symmetric condition, the target was non-symmetric and the distractors were symmetric. Three different set sizes were used, 4, 8, and 12. Each unique setting of stimulus type, target presence, symmetry condition, and set size was repeated 20 times, giving a total of $3 \times 3 \times 2 \times 2 \times 20 = 720$ trials. The trials were randomly shuffled. Prior to each trial, the participants were asked to fixate on a cross in the center of the screen.

The experiment started with 20 dummy trials to let the participants get used to the task.

Subsequently, the 720 trials were presented in eight blocks of 90 trials. After each block, the participants got one minute of rest. Halfway the experiment, a five minute break was taken. The number of correct trials was fed back to the participants after every 30 trials, in order to keep them motivated to do the task. In total, the experiment lasted a little less than an hour.

4.2.1.3 Stimuli

Three different stimulus types were used, stacked blocks, polygons, and dot patterns, similar to those used in (Olivers & van der Helm, 1998). The figures were presented in a circle with a diameter of two third of the height of the screen (approximately 17°), with the figures uniformly spread out over the circle. The circle was rotated over a random angle, to have different positions for the figures each trial. Each figure had a maximum width and height of 2.9°. All symmetrical figures were mirror symmetric with a vertical symmetry axis. Care was taken that the non-symmetrical figures did not accidentally show symmetry.

The stimuli were presented to the participants at a resolution of 1024×768 on an $18''$ CRT monitor of 36 by 27 cm at a distance of approximately 60 cm from the participants. The visual angle was approximately 33° horizontally by 25° vertically.

The three stimulus types are defined as follows:

Stacked blocks A stacked-blocks figure consists of eight filled bars stacked on top of each other. For a non-symmetrical figure, the left and right sides of the bars have randomly appointed lengths, with a minimum of 5% and a maximum of 50% of the height of the figure. For a symmetrical figure, the left and right sides are mirrored copies.

Polygons A polygon consists of 16 points. Each point is defined as a vector from the center of the figure, where the angles of the vectors are evenly spread out like a wagon wheel, that is, with intervals of 22.5°. The lengths of the vectors are randomly assigned between 5% and 50% of the maximum width of a figure. For symmetrical figures, the left and right side are mirrored copies. Different from (Olivers & van der Helm, 1998), the polygons are filled to increase their visibility.

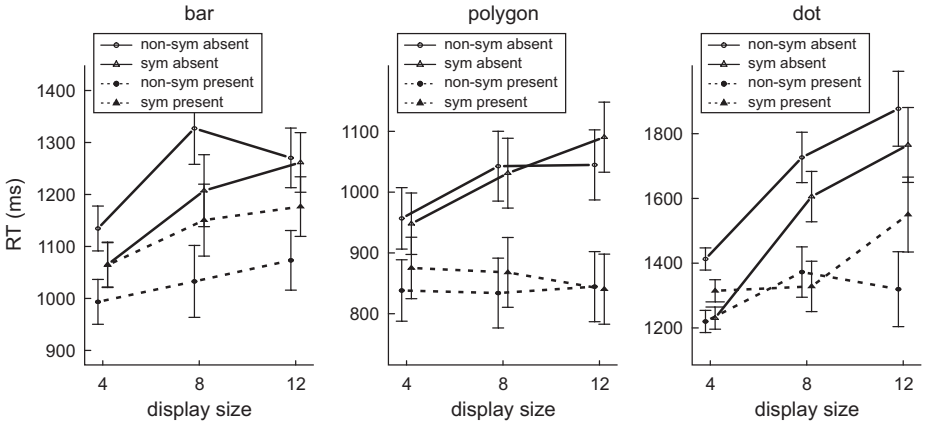

Figure 4.2: Reaction times of the symmetry pop-out experiment for the stacked bars, polygons, and dot patterns. The curves give the mean correct reaction times. The error bars depict the 95% confidence intervals. The dashed lines refer to the target-present trials, whereas the solid lines show the target-absent trials. The circles show results for the non-symmetric condition and the triangles for the symmetric condition. A pop-out effect of symmetry is especially clear for the polygon figures.

Dot patterns Each dot pattern consists of 16 dots randomly placed in a circle with a diameter of 2.9° and a minimum distance between two dots of 0.4°. The dots have a diameter of approximately 0.3°. In the symmetrical case, eight dots are placed and reflected about the vertical symmetry axis.

4.2.2 Results and Discussion

Data for which the reaction times lie outside three standard deviations of the participant's mean response to that specific set size × condition × target combination is disregarded, since these trials are likely to be false. Figure 4.2 shows the reaction times of the visual search experiment. The lines are the mean correct reaction times, and the error bars depict the 95% confidence intervals on the mean. It can be appreciated from the figure that the slopes of the target-absent curves are steeper than that of the target-present curves. This indicates that there is a symmetry pop-out effect. The pop-

out effect is especially clear for the polygon figures, which show an actual flat curve for the the target-present condition. The plots furthermore reveal search asymmetries for all stimulus types. The search for a non-symmetrical target among symmetrical distractors is faster than the search for a symmetrical target. The reaction times greatly differ for the different stimulus types. Search is fastest for the polygon figure, followed by the stacked bars. The search for a dot-pattern target is much slower.

Below, the results of within-subjects repeated measures analysis of variance (ANOVA) on the mean correct reaction times for the different stimulus types are given, followed by an analysis on the slopes of the reaction time $\times$ set size curves. This is similar to the analysis done by Olivers & van der Helm (1998).

4.2.2.1 ANOVAs

We performed a three-way within-subjects ANOVA on the mean correct reaction times with the factors set size (4, 8, 12), condition (symmetry / non-symmetry), and target (present / absent) for all stimulus types. There are two effects that we are interested in especially: 1) The ANOVA indicates a pop-out effect when there is a significant set size $\times$ target interaction, which indicates that the search efficiency differs between target presence and absence. 2) A significant condition $\times$ target interaction hints towards a search asymmetry. We furthermore performed a post-hoc two-way within-subjects ANOVA on the mean correct reaction times with set size and condition as factors. This analysis adds two more effects of interest: 3) This post-hoc ANOVA indicates as pop-out effect when there is a significant main effect of set size in the target-absent case and there is not, or a less strong main effect of set size in the target-present case. 4) A search asymmetry is indicated by a significant main effect of condition in the target-present case. The presence or absence of the four effects of interest are summarized in Table 4.1.

The three-way ANOVA for the stacked blocks shows a significant main effect for target ($F(1,9) = 19.0$, $p = .002$), and for set size ($F(1,9) = 33.8$, $p < .001$) and a significant condition $\times$ target interaction ($F(1,9) = 24.5$, $p < .001$). The set size $\times$ target interaction is not significant ($F(1,9) = 1.41$, $p = .27$). The post-hoc two-way ANOVA shows a significant main effect for set size ($F(1,9) = 15.5$, $p < .01$) and condition ($F(1,9) = 10.8$, $p < .01$) in the case of target absence. In the target presence case, set size is also a main effect, although less significant ($F(1,9) = 9.53$, $p = .01$), as is

condition ($F(1,9) = 10.4$, $p = .01$). In both the target presence and absence case, there is no significant set size $\times$ condition interaction.

For the polygons, the three-way ANOVA shows a significant main effect for target ($F(1,9) = 49.9$, $p < .001$) and a significant set size $\times$ target interaction ($F(1,9) = 7.46$, $p = 0.02$). No significant main effect for set size is found, due to the flat slope in the target-present case. The post-hoc two-way ANOVA shows a main effect for set size for target absence ($F(1,9) = 6.46$, $p = .03$), but not for the target-present case ($F(1,9) = 0.94$, $p = .36$). In both cases there are neither significant main effects for condition nor any interaction effects.

For the dot patterns, the three-way ANOVA reveals significant main effects for set size ($F(1,9) = 67.0$, $p < .001$) and target ($F(1,9) = 34.0$, $p < .001$), a significant set size $\times$ target interaction ($F(1,9) = 13.0$, $p < .01$), and a significant condition $\times$ target interaction ($F(1,9) = 29.0$, $p < .001$). The post-hoc two-way ANOVA shows significant main effects for set size ($F(1,9) = 41.8$, $p < .001$) and condition ($F(1,9) = 12.8$, $p < .01$) in the target-absent case. In the target-presence case, their is also a main effect of set size, though less significant ($F(1,9) = 17.9$, $p < .01$), and for condition ($F(1,9) = 13.7$, $p < .01$). No interaction effects are found.

Table 4.1: Summary of ANOVAs on the mean correct reaction times. The table shows whether the four effects of interest are strongly present (+), mildly present ($\pm$), or absent (-). Two effect hint towards a pop-out effect and two indicate a search asymmetry. The reader is referred to the text for the definition of the effects.

stimulus type	effect 1 pop out	effect 2 search asym	effect 3 pop out	effect 4 search asym
stacked blocks	-	+	$\pm$	+
polygons	+	-	+	-
dot patterns	+	+	$\pm$	+

Table 4.1 summarizes the results of the ANOVAs and post-hoc ANOVAs. In general we can conclude that there is a pop-out effect of symmetry. This suggests that the detection of symmetry is efficient and preattentive. The pop-out effect can also be seen in Figure 4.2. The reaction times are slower for the target-absent cases and the slopes of the curves are steeper. Furthermore, the data shows a search asymmetry. It can be seen in

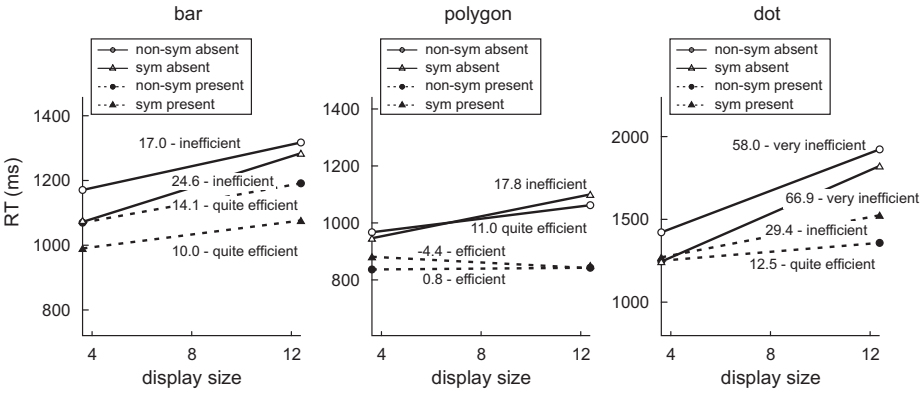

Figure 4.3: The reaction time $\times$ set size slopes. The lines give the fitted linear regression models. The dashed lines refer to the target-present trials, whereas the solid lines show the target-absent trials. The circles show result for the non-symmetric condition and the triangles for the symmetric condition. For every line, the slope (ms/item) and a classification of the search efficiency is given.

the plots that the search for a non-symmetrical target among symmetrical distractors is more efficient than the search for a symmetrical target among non-symmetrical distractors. This is in accordance with the search-asymmetry theory of Treisman & Gormican (1988) stating that it is easier to find a deviation among canonical stimuli than the reverse, where canonical here is symmetrical. According to Treisman & Gormican, this indicates preattentive detection of symmetry as well.

4.2.2.2 Slope Analysis

Figure 4.3 shows the regression lines calculated by fitting a linear model to the data. The slopes (ms/item) as well as the classification of slopes are given as well. The classifications are according to Wolfe (1998), who divided the slopes into four different categories: efficient ($\pm$0 msec/item), quite efficient ($\pm$5-10 msec/item), inefficient ($\pm$20-30 msec/item), and very inefficient ($>$ 30 msec/item).

For all stimulus types, the slopes for target absence are steeper than for target presence. This hints towards preattentive detection of the target. There is however a great difference of the efficiency of search for the different stimulus types. Search for the polygon

targets is efficient, whereas search for the stacked blocks is quite efficient, and the search for a symmetrical dot pattern among non-symmetrical ones is quite inefficient. Also the slopes for the target-absent cases differ. From quite efficient for polygons to very inefficient for dot patterns.

In general it can be concluded that the slopes show pop-out effects for all stimulus types. The effect is strongest for the polygons, followed by the stacked blocks, and weakest for the dot patterns. This suggests that symmetry can be preattentively detected.

4.3 Scene-Memory Experiment

In the scene-memory experiment, the participants are briefly presented with a scene containing a number of symmetrical and non-symmetrical figures. When a second scene is presented, the participants are asked to detect if the two scenes are identical or if a change occurred. By recording the eye movements and the response times, it is investigated whether more attention is given to symmetrical figures to remember the scene. If this turns out to be the case, symmetry is a salient feature and there is a causal relation between symmetry and eye fixations. If however, the symmetrical and the non-symmetrical figures receive as much attention, the results of Chapter 3 indicate that attention is not paid to symmetry per se, but that symmetry is a cue for the presence of an object, and visual attention is primarily object oriented.

4.3.1 Methods

4.3.1.1 Participants

Twenty undergraduate students from the University of Groningen participated in the experiment for course credits. The age of the participants ranged from 19 to 30 years. All participants had normal or corrected to normal vision. Two participants were taken out of the data set. One because the calibration of the eye tracker failed, the other because there was a disturbance half way through the experiment.

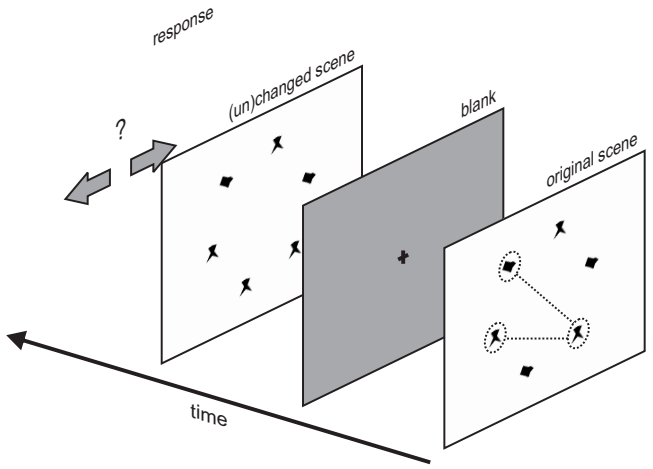

Figure 4.4: Trial setup of the scene-memory experiment. First, the original scene is presented for 1500ms with 6, 8, or 12 items. Next, a 1000ms blank screen is presented with a fixation cross placed in the center. When the second scene is presented, the participants are asked to respond as quickly as possible whether the two scenes were identical or that a change occurred.

4.3.1.2 Experimental Design

Prior to each trial, a fixation cross was presented in the center of the screen on which the participants had to focus. The setup of a trial is shown in Figure 4.4. The participants are asked to remember the first scene presented. This original scene consists of 6, 8, or 12 polygon figures placed in a circle, identical to the symmetry pop-out experiment. A scene contained symmetrical and non-symmetrical figures. The number of symmetrical figures was either $\frac{1}{2}N$, $\frac{1}{2}N-1$, or $\frac{1}{2}N+1$, where N is the total number of items in the scene, with the three cases uniformly distributed. This prevented that the participants could simply count the number of symmetrical figures in the second scene to solve the task. All symmetrical figures per trial were identical as were all non-symmetrical figures. The polygons did change over the trials. The original scene was presented for 1500ms, during which the participants made 5-6 fixations on average. The scenes with 8 and 12 items can therefore not completely be viewed by the participants. If there is a preference for symmetry, it should be visible for these scenes.

Next, a blank screen is presented to prevent an afterimage. A fixation cross was drawn in the center of the blank screen. After 1000ms, a second screen was presented. In 50% of the trials this scene was identical to the original scene. The other changed trials were either in the symmetric or the non-symmetric condition, both in half of the cases. In the symmetric condition, a symmetrical figure was replaced with a non-symmetrical figure. In the non-symmetric condition, a non-symmetrical item was changed into a symmetrical one. The participants were asked to decide as quickly as possible whether the two scenes were identical or not. They responded by a key press.

4.3.1.3 Data Acquisition

The reaction times and number of correct trials were recorded. If more attention is paid to symmetrical figures when remembering the scene, we expect to see faster response times and a higher percentage of correct trials in the symmetric condition than in the non-symmetric condition, since it should be more apparent in that a symmetrical figure changed when the participants pay more attention to symmetry.

We also recorded the eye movements of the participants when remembering the first scene. If participants pay more attention to symmetrical forms, the eye-tracking data should reveal this. We used the Eye Link I head-mounted eye tracker (SR Research). Fixations were extracted using the accompanied software. At the beginning of the experiment, the eye tracker was calibrated using the SR-research software. Prior to each trial drift was measured by letting the participants fixated on a cross in the center of the screen. The drift was corrected if necessary.

Based on the eye-fixation data, the *symmetry-fixation score*, s is determined. This is calculated as follows:

$$v = F_s/F_n \tag{4.1}$$
$$a = N_s/N \tag{4.2}$$
$$b = \log(0.5)/\log(a) \tag{4.3}$$
$$s = v^b, \tag{4.4}$$

where F_s is the number of symmetrical figures on which a fixation fell, F_n is the total number of figures on which a fixation fell, N_s is the number of symmetrical figures in the scene, and N is the total number of figures in the scene. In the cases that the

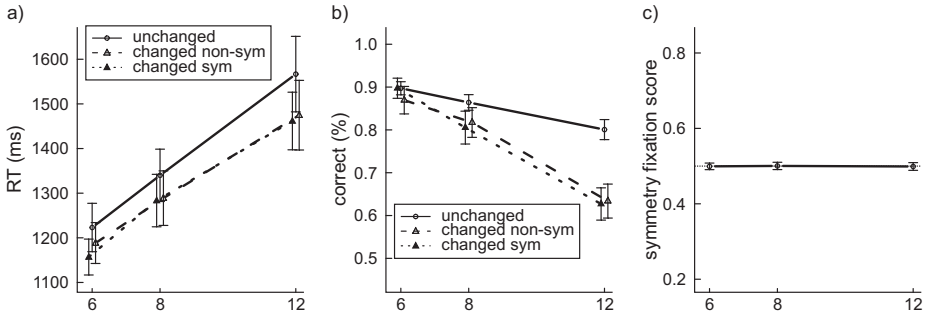

Figure 4.5: Scene-memory results. a) The reaction times (ms) as a function of the set size, b) the percentage of correct trials as a function of the set size , and c) the symmetry-fixation score. The lines depict the means over all participants. The error bars are the 95% confidence intervals on the mean.

number of symmetrical and non-symmetrical items in the scene is the same, s is simply v. In the other cases, the proportion of fixated symmetrical figures is corrected with respect to the proportion of symmetrical items in the display. If for instance 40% of the items are symmetrical and 40% of the fixations are on symmetrical items, then the symmetry-fixation score is 0.5.

4.3.2 Results and Discussion

The results of the scene-memory experiment are displayed in Figure 4.5. As expected, the reaction times increase with the number of items in the display, since it takes more time to inspect if any of the figures changed (see Figure 4.5a). The reaction times are faster when there is a change than when the scene is unchanged. This is also expected, since on average, fewer items need to be inspected to notice a change. However, we see no difference between the symmetric and the non-symmetric condition. Figure 4.5b shows the proportion of correct trials. The performance drops as a function of the set size. This correlates with the oral response that the participants gave after the experiment. They reported having difficulties noticing change when there were many items in the display. This is reflected by the scores close to chance for 12 items. The fact that the unchanged scores are higher than the changed scores for larger set sizes shows that the participants have a bias towards responding that there is no change. Also here

there is no difference between the symmetric and the non-symmetric condition. Finally, Figure 4.5c shows the symmetry-fixation scores. The score is 0.5 for all display sizes. This indicates that the participants fixated on symmetrical figures just as much as on non-symmetrical figures.

The fact that there is no difference for reaction times nor for correctness scores between the symmetric and non-symmetric condition suggests that the participants do not pay more attention to symmetry. This is supported by the eye-movement data, which shows that symmetrical and non-symmetrical items are evenly fixated on. The results thus show that participants have no preference for symmetrical objects over non-symmetrical objects when remembering a scene, but instead pay attention to any object that is present in the scene.

When asked after the experiment, some participant reported that they systematically viewed the figures in a clockwise fashion. Via manual inspection of the data, this was confirmed. The used setup and layout of the displays might therefore not be completely appropriate for the study. However, had there been a preference for symmetry, we would expect to have found it, especial for the larger set sizes, since the participants did not have enough time to inspect all figures.

4.4 Discussion

In this chapter, two experiments have been presented that investigate the role of symmetry in attention. The first experiment indicates that there is a pop-out effect for symmetry. The reaction time $\times$ set size slopes are less steep for target presence than for target absence for all stimulus types. Especially for the polygonal figures, the search for the target is efficient. The performed statistical tests also indicate pop-out effects. Moreover, they reveal a search asymmetry. Non-symmetrical targets among symmetrical distractors are found faster than vice versa. The presence of a pop-out effect and a search asymmetry suggest that detection of symmetry is preattentive.

The results of the first experiment contradict with the findings of Olivers & van der Helm (1998). They found no pop-out effect of symmetry. However, this might be due to the fact that the distractors used in their experiments were highly heterogeneous. Each distractor had a completely different shape, with only the presence or absence of symmetry in common with the rest. It is known that search becomes inefficient when

the distractors are sufficiently heterogeneous (Bauer et al., 1996). In the experiment presented in this chapter, a homogeneous set of distractors is used. This causes the problem that difference in symmetry coincides with difference in shape. However, an asymmetry in search for the target is observed, which is in accordance with the search-asymmetry of Treisman & Gormican (1988) with symmetry being the canonical case. This is not expected to be found if shape would have been the feature used by the participants. The effects found can therefore be attributed to symmetry.

In the second experiment, the participants are asked to memorize a scene with multiple symmetrical and non-symmetrical figures. The results show that equally much attention is paid to the symmetrical and the non-symmetrical items. The participants paid attention to the objects in the display irrespective of their symmetry. Although this might seem to contradict the results found in Chapter 3 at first sight, this is not the case. The results show that human visual attention is primarily object oriented. If we take this as a basic assumption, the fact that symmetry has been found to be a good predictor of the location of human eye fixations indicates that symmetry is used as a cue for the presence of an object. However, there are more cues for figure-ground segregation. For the stimuli used in the second experiment, closure, for instance, is a strong cue. In fact, the black objects are very easily detectable on the uniform white background. Since this is the case, symmetry has no added value to the segregation of figure and background. The stimuli used in the previous chapter are more complex, with different types of objects on highly cluttered backgrounds. For these kinds of stimuli, symmetry is a valuable cue. We will elaborate on object-oriented visual attention and the role of symmetry as a cue for figure-ground segregation in the next chapter.

5

Object-Oriented Visual Attention

In the previous chapters, visual attention was discussed in terms of basic features, such as local symmetry features and contrast features. However, to interpret a visual scene, humans seem to pay attention to objects and not so much to individual features (Scholl, 2001; Yeshurun et al., 2009). Objects are more than sets of features, since the spatial arrangement of the features is also important. This was illustrated by Figure 2.4. Due to the configuration of the parentheses, the features symmetry and closure emerge that are not properties of the simple elements. These configural features are demonstrated by the pop-out effect (as shown by the figure). The configural features are stronger visual attractors than the basic features.

In this chapter, the role of objects in visual attention is discussed. Section 5.1 discussed the preattenive perception of objects and the role of bottom-up mechanisms for the segregation of object and background. The Gestalt principles for grouping and figure-ground segregation play an important role in this process. They are discussed in Section 5.2. Finally, Section 5.3 interprets the findings in Chapter 3 and 4 and discusses the role of symmetry in visual attention.

5.1 Paying Attention to Objects, Not to Features

5.1.1 Object-Oriented Attention

If objects play a role in bottom-up visual attention, they must be perceived preattentively, that is, without the focus of visual attention. Objects have indeed been found to have a preattentive existence. Enns & Rensink (1993), for instance, used search displays with figures that could be interpreted as three-dimensional objects (see Figure 5.1). The search for a target with a different lighting direction resulted in efficient search (a). However, when a display contained patterns with matched complexity but with features that could not be interpreted as three-dimensional objects, search was inefficient (b). Rensink & Enns (1995) demonstrated that occluded objects are preattentively observed as complete objects, and not as dissected objects due to the occlusion. Similarly, Wolfe (1996) showed that visual search is not disturbed when there is a lattice displayed in front of the search field. Although the lattice creates spurious basic features at the conjunctions with the search items, the items are observed by humans as being an object behind another object. These examples show that objects are perceived preattentively and that the object representation plays a role in the allocation of visual

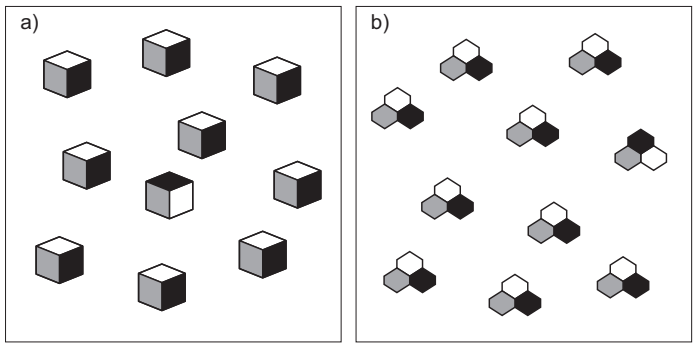

Figure 5.1: Efficient search for 3D object properties. a) results in a pop out of the target, which can be interpreted as a three dimensional object with a different lighting pattern. Although of the same complexity and with similar features, b) results in inefficient search. This shows that objects are preattentively observed, and object representations play a role in visual attention. (Adapted from Enns & Rensink, 1993)

attention. This means that bottom-up and preattentive visual attention is not only controlled by low-level basic features, but also by higher-level features that identify objects in the visual field.

Another interesting example of the preattentive aspects of figures was discussed earlier in this thesis and demonstrated in Figure 2.4. Pomerantz (2006) and Pomerantz et al. (1977) showed that there can be a configural superiority effect: the search for the parenthesis with the odd curvature is more efficient when the parenthesis forms a symmetrical object configuration with another parenthesis. The symmetrical versus the non-symmetrical configurations are more salient than the odd curvature.

Baylis & Driver (1993) found that humans are better in judging the relative position of two contours when both contours belong to the same object. When the same contours belong to two distinct objects, on the other hand, reaction times and error rates increase. Similarly, it is relatively easy to judge the mirror symmetry of two contours when the contours belong to the same object. However, when the contour belong to two distinct object, symmetry recognition of the two contours is hard (Driver & Baylis, 1995).

The above-discussed studies show that there is a preattentive and stimulus-driven notion of objects that influences attentional processes, which cannot be explained by the

basic features only. Objects consists of basic features, but possess properties that are not possessed by the basic features. An object is a combination or conjunction of these features. Figure 2.4 gives a nice example of features of an object that are not part of the basic features of which the object consists. With the addition of the extra curved line, the configural features symmetry and closure emerge. The next section deals with the question of how the basic features are perceptually grouped to be perceived as objects.

5.1.2 Segregation of Objects From Their Background

There are mechanisms in the visual system that determine which features belong together and form objects. While it is likely that there are top-down processes involved as well, Gestalt psychology gives a number of bottom-up principles that play a role in the grouping of elements and in figure-ground segregation (Koffka, 1935; Köhler, 1947). One of these principles is symmetry. Symmetry is a cue to segregate figures (objects) from there background (Driver et al., 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 belong together and are considered object, and the non-symmetrical parts are considered background. This shows that symmetry can be used by the human visual system to detect objects in a *context-free* manner, that is, bottom-up without any knowledge about the objects. The success of local-symmetry detectors in predicting human eye movements as presented in Chapter 3 can therefore be explained by the fact that symmetry is a cue for figure-ground segregation. Symmetry predict the location of objects in the image and these objects attract visual attention. Symmetry therefore correlates well with human eye fixations.

Driver et al. (1992) tested the symmetry perception of a patient with visual hemineglect. The patient pays no attention to information on the left side, but has no loss in the left visual field. Due to this attentional deficit, he performed at change levels when asked to judge if patterns contained vertical mirror symmetry. He could not compare the left and right sides, because that required attention. However, when asked to judge what is foreground and what background in images similar to 2.8, he judges the symmetrical regions as foreground, just as people with normal vision do. Apparently symmetry can be perceived without the need of attention. This shows that symmetry perception is preattentive and a strong context-free cue for figure-ground segregation.

In the Gestalt psychology, not only symmetry, but more principles for perceptual grouping and figure-ground segregation are studied and proposed. In the next section, we discuss the Gestalt psychology and the main principles for perceptual grouping and figure-ground segregation.

5.2 Gestalt Laws of Figure-Ground Segregation

"Gestalt" is the German word for *form* or *configuration*. The Gestalt approach is based on the notion that the whole is more than the sum of its parts, that is, that the figure is more than the sum of its basic features. Kurt Koffka (1886-1941) (Koffka, 1935), Wolfgang Köhler (1887-1968) (Köhler, 1947), and Max Wertheimer (1880-1943) (Wertheimer, 1945) are some of the early Gestaltists. Although some of the work of the Gestaltists is criticized, for instance the phenomenological approach of introspection, the main principles are largely supported today (Palmer, 1992).

Although the early Gestaltists were not interested in mental processes, but rather emphasized perceptual phenomena, the Gestalt principles are well founded and have relevance for psychology, physiology, neurology, and computer science. The main contribution of Gestalt psychology to the study of perception are the principles for *perceptual grouping* and *figure-ground segregation* that explain the grouping of individual component to objects, and the segmentation of objects from their backgrounds.

5.2.1 Principles for Perceptual Grouping

The main idea behind the Gestalt theory of perceptual grouping is the *law of Prägnanz*, or the *law of good figure*. The law states that humans tend to perceive a given visual array in the most simple organization. Humans organize the individual basic elements in the visual field to form the most simple, stable, and coherent interpretation. The Gestalt principles are thought to be utilized by the brain to process the sensory input into objects in the world (Wertheimer, 1945). Based on the principles, the individual parts in the stimulus are grouped to form wholes. Several principles have been proposed through the years (see Wertheimer, 1923, for an early publication on the topic). The most established principles are given in Figure 5.2 and listed below:

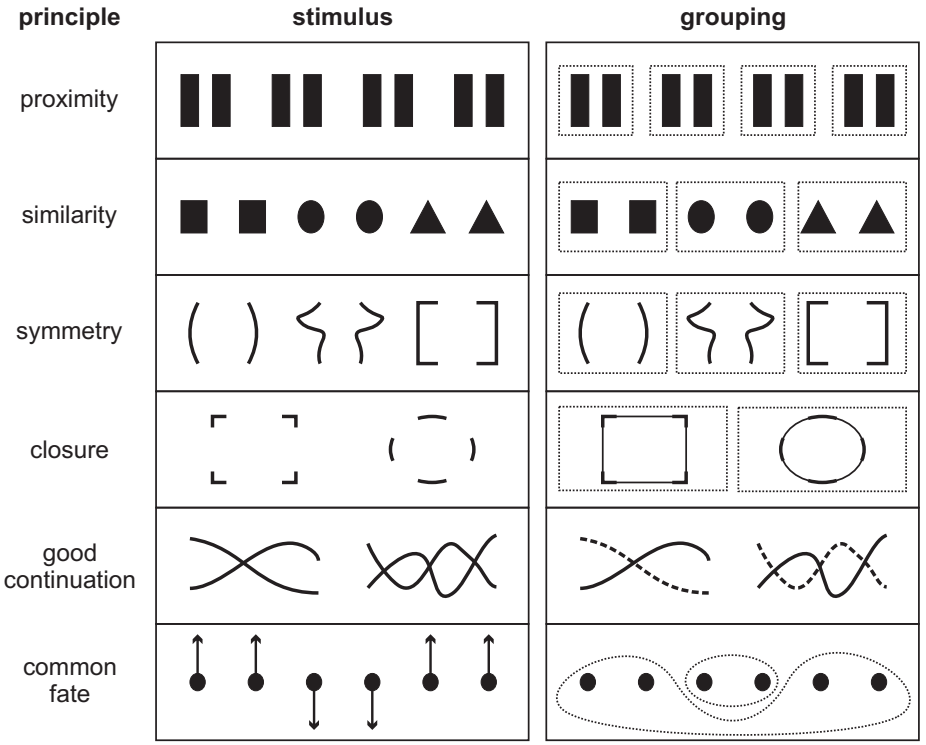

Figure 5.2: The Gestalt principles for perceptual grouping. The stimulus is given in the second column, and the perceived grouping in the last column. The groups are depicted by the dashed regions for all stimuli except for the principle of good continuation, where the lines that are perceived as one are depicted by the solid and the dashed line. The principles are further discussed in the text.

Principle of proximity Things that are located close together are likely to be grouped together.

Principle of similarity Items that appear similar, that is, item that have similar features, are more likely to belong together.

Principle of symmetry Things that are symmetrical with respect to each other tend to be grouped together. Symmetry is a non-accidental property. It is very unlikely that two unrelated parts exhibit symmetry. Therefore, if symmetry is present, it is likely to come from the same source in the external world and therefore belongs to the same object.

Principle of closure The human mind tries to fill in missing information to complete a regular figure. This means that items that are positioned so that they are perceived as belonging to an incomplete structure, or a space that is not completely enclosed, are perceived as a whole.

Principle of good continuation Humans tend to perceive continuous figures. Components that have a continuous and smooth transition are more likely to belong together. Abrupt transitions are less likely to occur within one object.

Principle of common fate Things that move together, belong together, that is, items that move in the same direction with the same speed are likely to form one object.

5.2.2 Principles for Figure-Ground Segregation

The perceptual-grouping principles are closely related to the principles for figure-ground segregation. The parts of the visual display that comply with one of the grouping principles are most likely the figure, whereas parts of the image that violate the principles are most likely background. A symmetrical configuration, for instance, is likely to be an object in front of a non-symmetrical background. Apart from the grouping principles, a number of additional figure-ground segregation principles have resulted from Gestalt psychology. The most important and classical principles ones are shown in Figure 5.3 and listed below:

Principle of symmetry Regions with symmetrical edges are likely to be objects (Driver et al., 1992; Bahnsen, 1928). The reason is that symmetry is a non-accidental

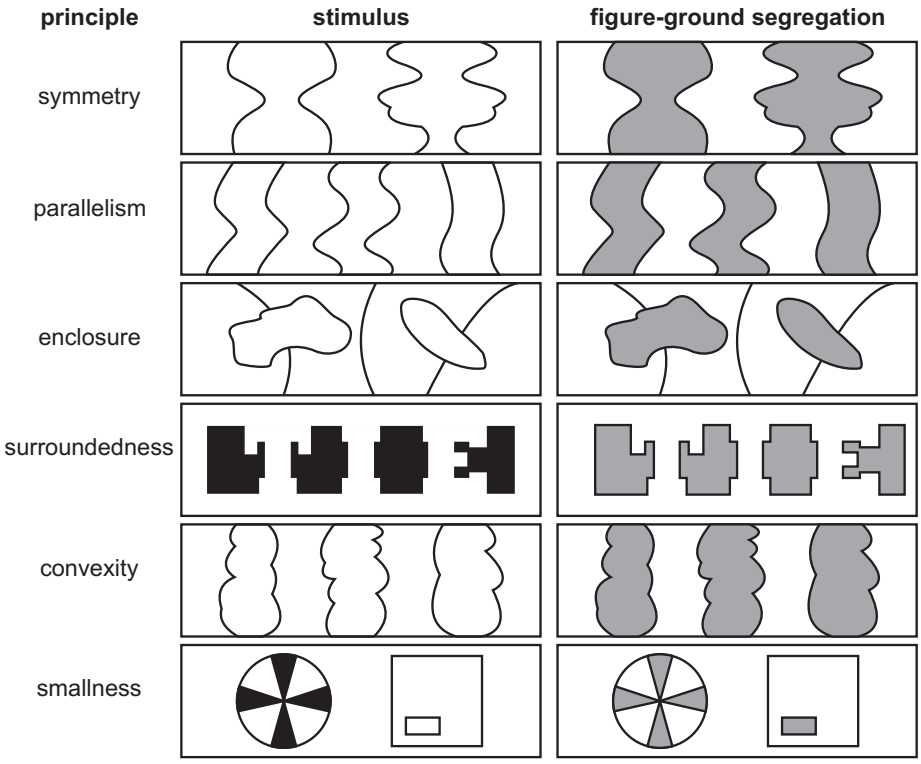

Figure 5.3: The Gestalt principles for figure-ground segregation. The stimulus is shown in the second column. The third column gives the figures perceived by humans by the gray regions.

stimulus and the pattern is thus likely to belong to one object. The non-symmetrical parts are considered to be background.

Principle of parallelism Regions that have parallel edges are generally perceived as objects (Metzger, 1953).

Principle of enclosure Parts that form a closed structure are likely to be objects.

Principle of surroundedness Areas surrounded by others tend to be seen as figure. The example in Figure 5.3 demonstrates that this can be misleading; the text 'TIE' is hard to perceive at first.

Principle of convexity Regions that have convex edges are more likely to be objects (Kanizsa & Gerbino, 1976; Driver & Baylis, 1995). It is Possibly because the convex parts protrude into other regions (Hoffman & Singh, 1997)

Principle of smallness Smaller areas are tend to be perceived as figure against a larger background

Kimchi & Peterson (2008) showed that the figure-ground segregation based on such principles can be performed by humans without the need of focal attention. Moreover, Kimchi, Yeshurun, & Cohen-Savransky (2007) demonstrated that the organization of elements in the visual field into objects automatically attracts attention. The Gestalt principles therefore seem to have a preattentive existence.

5.2.3 Figural Goodness and Symmetry

The Gestalt principles contribute to the *goodness* of a figure. The Gestalt notion of figural goodness refers to the simplicity and orderliness of a figure. Symmetry is considered one of the important determinants of figural goodness (Palmer, 1991; Hochberg & McAlister, 1953; Attneave, 1954; Koffka, 1935). Due to the intrinsic redundancy and organization, symmetrical forms are simple and ordered. Humans judge symmetrical patterns as less complex than non-symmetrical patterns ((Palmer, 1991). Patterns with mirror symmetry are judged simpler than patterns with 180°rotational symmetry. Moreover, the perceived complexity is strongly correlated with the number of symmetry axes. More symmetry axes result in less complexity. Fisher et al. (1981) concluded that infants already respond to the goodness of organization of forms on the basis of symmetry.

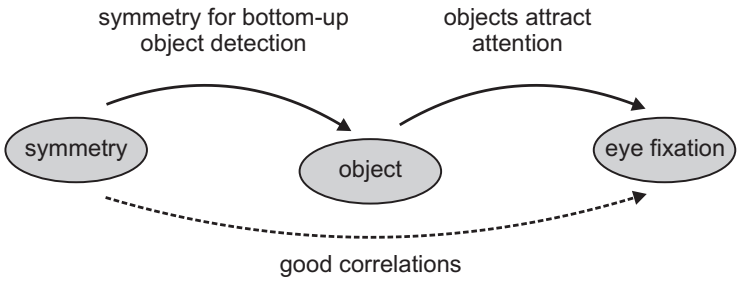

Figure 5.4: The main conclusion of Part I: The good correlations between the predictions of the symmetry-saliency model with the human eye fixations can be explained by the object-oriented nature of human visual attention and the role of symmetry in figure-ground segregation.

5.3 The Role of Symmetry in Attention

A number of conclusions can be drawn from this chapter that put the results of Chapter 3 and Chapter 4 in a new perspective. Firstly, objects are perceived preattentively. Secondly, the preattentive perception of objects plays a role in bottom-up human visual attention. Visual attention is predominantly object oriented. Finally, there are a number of bottom-up cues for perceptual grouping and figure-ground segregation. Symmetry is one of these cues.

These conclusions can explain the good correlations between the predictions of the symmetry-saliency model and the human eye fixations observed in Chapter 3. Human visual attention is object oriented and symmetry plays a role in figure-ground segregation. The presence of symmetry is thus a cue for the presence of objects, causing humans to focus attention on the symmetrical parts of the visual field (see Figure 5.4). Hence, symmetry is a good predictor for human visual attention.

The results of the symmetry pop-out experiments in Chapter 4 show that symmetry is perceived preattentively. This is in concurrence with the observations that objects can be perceived preattentively and that symmetry is a bottom-up cue for object segmentation.

The scene-memory experiment presented in Chapter 4 showed no effect of symmetry in the memory of items in a display. The participants rather paid attention to any object, symmetrical or non-symmetrical. The fact that symmetry had no effect can

be explained by the object-oriented nature of visual attention and the many different Gestalt principles for figure-ground segregation. Although symmetry is one of the principles, there are others that apply to the figures used in the experiment as well. The figures could also be segregated from the background by enclosure, surroundedness, and smallness. The non-symmetrical figures could therefore be easily identified. That is why the participants showed no increased interest in the symmetrical figures for remembering the scenes.

It can be concluded from this part of the thesis that humans pay attention to the symmetrical parts of a visual scene. Predictions based on local symmetry in the image compare well with human eye fixations. Furthermore, it can be concluded that the perception of symmetry is efficient and preattentive. It is suggested that human visual attention is object oriented and that symmetry is a cue for bottom-up object detection. This all suggests that symmetrical parts of a visual scene are interesting to pay attention to. The applicability of symmetry to improve artificial vision systems is therefore investigated in the next part of the thesis. Models for the detection of symmetrical interest points and symmetrical regions-of-interest are proposed to select landmarks to describe the visual environment of a mobile robot. Furthermore, the use of active vision to simplify perceptual tasks is discussed.

Part II

Artificial Systems

6

Visual Attention and Active Vision in Artificial Systems

Just like natural systems, artificial systems have to deal with an enormous amount of visual information. Even more so than the human brain, current computer systems are not apt to process all this information. Although the processing power of modern computers constantly improves, there are two important factors that limit visual processing. First, the fact that computers are serial, and second, we still do not know how to use all the processing power as efficiently as evolved in natural systems. Therefore, computers are much slower, and less efficient in their calculations than natural brains. To deal with the insufficient processing power, visual attention is highly important. To have real-time artificial systems that interact with the world, mechanisms are needed to focus computation on relevant information instead of waisting time and resources on irrelevant information. We believe that the study of visual attention in natural systems can help to develop adequate attention mechanisms for computer science and robotics. Natural systems furthermore take advantage of actively perceiving the world. By actively changing viewpoint, additional information becomes available, which improves perceptual tasks like object segmentation and recognition.

Strategies of visual attention and active vision in natural systems are adopted in this part of the dissertation to improve machine vision. This chapter introduces a number of concepts and techniques used in machine vision. The following chapters discuss a number of robotic studies that have been carried out. In Chapter 7, active vision is used to improve object recognition in the real world. In Chapters 8 and 9, the results of the first part of the dissertation are used to focus the attention of a robot to select visual landmarks in the environment based on symmetry. These landmarks are used to build a map of the environment.

In Section 6.1, the use of visual attention for object representation and recognition is introduced. Section 6.2 presents interest-points models for image representation and the scale-invariant feature transform (SIFT) and its problems are discussed in Section 6.3, since it is one of the standard models to detect and describe interest points. My approach to use symmetry to detect interest points is introduced in Section 6.4 Next, simultaneous localization and mapping and the use of visual attention to select landmarks in the environment are discussed in Section 6.5, followed by an introduction to the concept of active vision in Section 6.6.

6.1 Visual Attention for Object Representation

In order to recognize objects and scenes, they need to be represented in an appropriate form. Some different representations are discussed in this section, including interest-point representations. The link between interest points and visual attention is discussed as well.

Most current models for object recognition use an appearance-based object representation. That means that the appearance of the object (or of parts of the object) is stored as a set of features. The most straightforward representation is a feature vector that consists of the set of all pixels that make up the object. This representation, however, is not very useful, since it has no invariance to changes. If the viewpoint changes only slightly, for instance, all pixels will have different color values and the feature vector consisting will therefore change drastically. The same happens when there is a change of illumination or when the object gets partly occluded. Other representations are more suitable to represent objects. In what follows, we give a short overview of methods for full-object representation and discuss their problems. Although many other representations exist, we solely focus on interest points as a method to represent objects in the remainder of the section. A full overview of methods for object representation is outside the scope of this thesis.

6.1.1 Full-Object Representation

In the case of *full-object* representation, the features complete objects are used to represent the object. The use of *color distributions* is an example of such a representation. The distribution of colors that are found on the object are represented using color histograms (Shapiro & Stockman, 2003) or Gaussian-mixture models (Alata & Quintard, 2009). This representation is invariant to shifts in position, to rotation, and to scale. The disadvantage of the representation is that it is not very distinctive, because all information of spatial structure is lost. This can be improved by splitting the object into a number of subregions and computing a histogram for each subregion. Another disadvantage is the fact that colors appear very differently in different lighting conditions.

Another example of full-object representation is the use of *principal-component analysis* (PCA), which has been successfully applied to human-face recognition (Turk & Pentland, 1991). PCA can be used to find the most important *eigenvectors* (or *eigen-*

faces in the case of face recognition). The eigenfaces each have an eigenvalue, which indicates the portion of the variance that is captured by the eigenface. By using the first number of most principal eigenfaces, new face images can be transformed into a low-dimensional feature vector, capturing the most important linear combinations of the input for recognition and classification. The advantage of using PCA is that structural information is captured well and that it results in a low-dimensional feature vector. The disadvantage, however, is that additional pre-processing steps are needed to make the representation invariant to position, rotation, scale, and illumination. Because all pixels are treated independently, the objects need to be correctly aligned and the pixel values normalized.

6.1.2 *Interest-Point Representation*

Two common problems with the full-object representation are that it cannot handle occlusions of the object, and that it changes drastically when the object is observed from a different viewpoint. To solve these problems, objects are nowadays often represented by a set of local *interest points* (Harris & Stephens, 1988; Lowe, 1999; Schmid & Mohr, 1997) (see Figure 6.1). This has the advantage that when the object is partially occluded, some interest points will disappear, but others will remain visible, making recognition still possible. Also, when the object is rotated around its vertical axis, its full appearance can drastically change, because parts of the object become visible, whereas other parts disappear. Many of the interest points, however, remain visible, and the local changes due to the rotation at these points is minimal. Interest-point representation is furthermore more robust to noise and to clutter in the background, because the representation is spread out over a number of interest points. Some points might get affected, but others will maintain their descriptive power, making it more robust than a full-image representation.

6.1.3 *Visual Attention and Interest Points*

As we have seen in Part I, humans make eye movements to focus attention on interesting parts of the visual field. Similarly, computer-vision systems use interest-point detection to efficiently focus computational power. Instead of having to process the complete image with computationally expensive image-processing techniques, a rela-

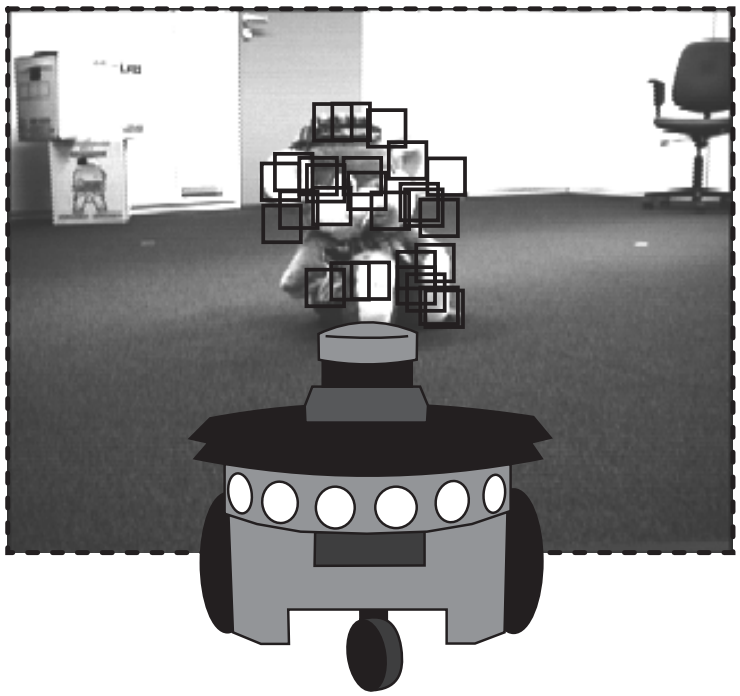

Figure 6.1: An illustration of the use of interest-points to represent an object. Instead of representing the full object, interest points are detected that are stable under various changes in viewpoint and illumination. The object is then represented by a set of interest-point detectors.

tively simple method can be used to detect the interest points, followed by sophisticated methods to describe and match the local neighborhoods of these points for recognition.

The eye fixations reflect the interest of humans. This is for instance used in (Van Maanen, 2009, Chapter 6) to personalize the presentation of paintings. Computational models for visual attention that predict human eye fixations can therefore also be used to select interest points for image representation and recognition. This is illustrated by the similarity between the contrast-saliency model of Itti et al. (1998) to model human visual attention and the scale-invariant feature transform of Lowe (2004) for the detection of interest points. Both models use difference-of-Gaussians calculations to

detect saliency in the image. This close relation between both fields motivated us to use our symmetry-saliency models to develop interest-point detectors. Since the symmetry models outperform the contrast-saliency model in predicting the points of human eye fixation, we are interested to see if symmetry can also be used to improve interest-point detection. We further elaborate on this in Section 6.4.

6.2 Interest Points

Interest-point methods consist of two parts, the *detector*, which selects the interest points, and the *descriptor*, that represents the local patch surrounding the points. A good detector robustly selects the same points independently of changes in viewpoint, illumination and noise. The descriptor should distinctively describe the points, while also being invariant to changes in viewpoint, illumination and noise.

6.2.1 Interest-Point Detectors

The *Harris corner* detector (Harris & Stephens, 1988) is one of the most widely known interest-point detectors. It detects interest points lying on corners in the image based on the second-moment matrix, which is a matrix representations of the partial derivatives in the image. The method, however, is not scale invariant. Lindeberg (1998) developed a method to automatically select the scale of an interest point, That method has been refined by Mikolajczyk & Schmid (2001) to the *Harris-Laplace* detector. This detector finds interest points using a scale-adapted Harris corner detector, and subsequently determines the associated scale by finding the scale that optimizes the Laplacians in the Gaussian-blurred images. The *Hessian-Laplace* detector is a variation that selects points for which the trace and determinant of the Hessian matrix simultaneously assume a local extremum. This way, blobs in the image are selected as interest points. Lowe (2004, 1999) proposed the *scale-invariant feature transform* (SIFT), which calculates difference-of-Gaussians at different scales to approximate the Laplacian of Gaussians. Points are selected that are local extrema both spatially and in scale. This results in interest points that lie on blobs in the image. The speeded-up robust features (SURF) is a fast interest-point detector, which approximates the Gaussian second-order partial derivatives needed to compute the Hessian matrix by utilizes integral images (Bay, Tuytelaars, & Van Gool, 2006).

The Harris-Laplace, Hessian-Laplace, SIFT, and SURF detectors are rotation- and scale-invariant interest-point detectors. Mikolajczyk & Schmid (2002, 2004) proposed an interest-point detector that is also affine-invariant. Initially, the Harris-Laplace detector is used to approximate the location of an interest point. In an iterative process, the location, scale, and shape of the point is changed to become affine invariant. The affine-invariant shape is determined using the second-moment matrix, and can be used to transform the image so that an affine-normalized descriptor can be computed. A similar method is proposed by Tuytelaars & Van Gool (2004), in which parallelograms are determined at corners in the image and transformed to squares to obtain affine-invariance.

The Harris-Laplace detector finds interest points at corners of objects in the image. The other detectors all find interest points at blobs in the image. This also results in interest points near corners, since these have a blob-like structure.

The above mentioned interest-point detectors are all based on the first or second partial derivatives in the image. This makes these detectors vulnerable to noise in the image, because that results in local maxima in the derivatives, as is discussed in Section 6.3. The use of symmetry to improve noise robustness is proposed in Section 6.4 and Chapter 8.

For more information on interest-point detectors, we refer to a survey by Tuytelaars & Mikolajczyk (2008).

6.2.2 Interest-Point Descriptors

The most popular interest-point descriptors use histograms to describe the texture in a local patch surrounding the interest point. The SIFT descriptor (Lowe, 2004) represents the patch by histograms of oriented gradients. The patch is divided in 16 subregions and the descriptor consists of histograms of the gradient orientations for all subregions. The representation is made rotationally invariant by normalizing the orientations with respect to the dominant orientation, scale invariance is achieved by adapting the size of the patch to the scale of the interest point, and normalization of the feature vector makes it robust to changes in illumination. SIFT is discussed in more detail in the next section and in Appendix B. PCA-SIFT (Ke & Sukthankar, 2004) uses principle-component analysis to reduce the dimensionality of the SIFT feature vector. The GLOH descriptor (Mikolajczyk & Schmid, 2004) is also based on SIFT, but uses a circular instead of

a square patch around the interest point to make the feature vector more rotationally invariant. The *shape context* feature (Belongie, Malik, & Puzicha, 2002) uses a three-dimensional histogram for the representation of the location and orientation of edges in the patch.

We use SIFT as interest-point detector and descriptor in this thesis. In comparative studies, SIFT is among the best performing interest-point detectors for three-dimensional object detection (Moreels & Perona, 2007), and for landmark selection in simultaneous localization and mapping (Mozos, Gil, Ballesta, & Reinoso, 2008). Although the affine-invariant detector scores better with objects with planar surfaces, SIFT performs better when the objects have non-planar surfaces (Moreels & Perona, 2007), because the planar assumptions used to correct for affine transformations are violated. Also the performance of the SIFT descriptor is currently state-of-the-art (Mikolajczyk & Schmid, 2005), although in competition with GLOH.

For more information on interest-point descriptors, we refer to (Mikolajczyk & Schmid, 2005).

6.2.3 Storing the Interest Points

The detected and described interest points need to be stored for recognition. The most straightforward method is to store all feature vectors in a database along with the associated object. For recognition, the newly observed interest points can be compared with the database by using a distance measure on the feature vectors, such as the Euclidean distance. The similarity between the observation and objects stored in the database is then determined by the number of matching interest points. A better matching can be achieved when the relative position of interest points on the object are stored as well (Lowe, 1999, 2004).

A method that has good results in image matching is the *bag-of-features* approach (Csurka, Dance, Fan, Willamowski, & Bray, 2004), which is a analogy of the *bag-of-words* approach in text classification. In that approach, a set of N prototypical interest points are learned from an initial set of interest points using clustering methods. Instead of storing all individual feature vectors in an image, the interest points are first clustered into these prototypes and the representation of the image is a vector of N elements indicating if a prototype is present or not. The main advantage of this model is that instead of storing $M \cdot K$ elements in the database, where M is the number of observed interest

points and K is the length of the interest-point descriptor, only N bits need to be stored to represent the image. Despite the loss of information about the position of interest points, the method has been shown to work well for image classification (Csurka et al., 2004; Winn, Criminisi, & Minka, 2005). The disadvantage of this method is that the interest points presented during execution should be similar to the ones used to learn the prototypical interest points, otherwise recognition will be impaired.

In (Kootstra, Ypma, & de Boer, 2007, 2008b), we present a method based on a *growing-when-required* (GWR) network (Marsland, Shapiro, & Nehmzow, 2002) that shares many similarities with the bag-of-features approach, but learns the features on-line instead of in an initial learning phase. The advantage of this approach is that new features can be added if needed, allowing for completely different interest points during execution. The GWR method for interest-point storage will be further explained in Chapter 7.

6.3 The Scale-Invariant Feature Transform

As discussed, the Scale-Invariant Feature Transform (SIFT) (Lowe, 2004) is one of the most popular methods for interest-points detection and description. The SIFT detector uses difference of Gaussians to select interest points. A point is selected when it has a local optimum difference-of-Gaussian value both spatially and in scale, that is, when its value is locally optimal with respect to neighboring points in scale and space. The points are described using histograms of oriented gradients. A neighborhood patch around the interest point is divided into 4×4 subregions. Histograms of the orientations of intensity gradients are constructed for all subregions. Each histogram has 8 bins, resulting in a 128-valued feature vector. The method is described in detail in Appendix B.

SIFT is scale and rotational invariant. Scale invariance is achieved by calculating the difference of Gaussians at different scales. The size of the neighborhood patch is determined based on the scale. Rotational invariance is created by rotating the patch using the dominant gradient orientation in the patch. Moreover, the descriptor is robust to changes in illumination and affine changes in the image. There are, however, two problems with SIFT that deserve some attention: the time performance, and its noise robustness.

6.3.1 Problem 1: Time Performance

There are three main reasons that make SIFT a rather slow interest-point detector and descriptor. (a) The detailed scale space utilized by the SIFT detector, together with the Gaussian filtering and the subtraction of images to obtain the differece-of-Gaussians images, results in a high computational load. (b) The 128-dimensional feature vector results in high computational costs for the matching of interest points. Especially because (c) SIFT results a high number of interest points per image.

The speeded-up robust features (SURF) of Bay et al. (2006) is a much faster interest point detector, while having a repeatability comparable to SIFT (Bay, Ess, Tuytelaars, & Van Gool, 2008). The SURF detector choses points that maximize the determinant of the Hessian matrix. Points that have a locally maximal Hessian's determinant both spatially and across scale, are selected as interest points. This results in points that have large second derivatives in two directions, that is in points on blob-like structures. To calculate the Hessian, SURF utilizes integral images, or summed-area tables, for fast approximation of the Gaussian second-order partial derivatives.

The standard SURF descriptor (SURF-64) also has a shorter feature vector, containing 64 elements. The method to describe the neighborhood of an interest point is very similar to SIFT, but instead of building a histogram of gradient orientations for every subregion, the gradient vectors and their absolute values are summed, resulting in four values per subregion. This speeds up the computations, while maintaining a performance similar to SIFT (Bay et al., 2008). Another method to reduce the dimensionality of the SIFT descriptor is PCA-SIFT, which results in a 36-dimensional descriptor after applying principal-component analysis to the original feature vector (Ke & Sukthankar, 2004). Although PCA-SIFT results in faster interest-point matching, its descriptor is less distinctive than the original SIFT descriptor (Mikolajczyk & Schmid, 2005).

As mentioned, SIFT results in a high number of interest points per image. Moreover, a large proportion of these points are not re-detected in subsequent observations (Kootstra & Schomaker, 2009b; Kootstra et al., 2008b). Chapter 7 describes an active method to let a robot explore object to enable it to reject unstable interest points. A similar method to reduce the number of interest points and increase their repeatability is presented in Chapter 8 to find stable interest points to serve as landmarks in a map of the robot's environment.

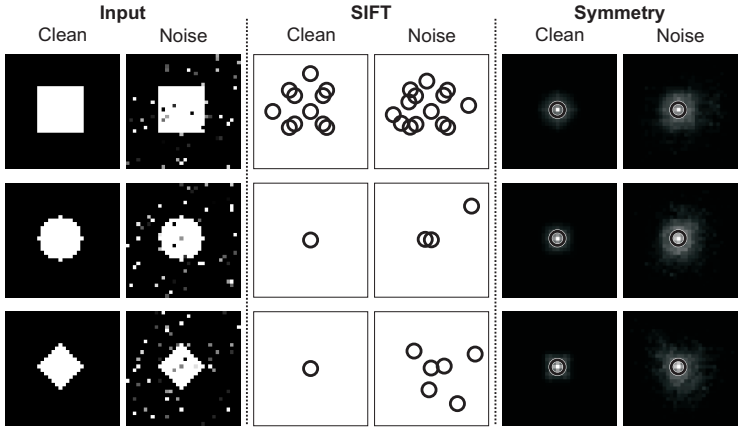

Figure 6.2: Response to original and noisy stimuli. Ten percent of the pixels in the noisy images are affected by Gaussian noise. The first two columns show the input images. The middle two columns show the detected SIFT interest points. It can be observed that the noisy stimuli result in many spurious interest points. The last two columns show the saliency maps for these stimuli calculated with the symmetry-saliency model presented in Chapter 3. The maps show a clear peak at the center of the figures for both the clean and the noisy stimuli. It can be seen that the local maxima, illustrated by the circles, are stably detected at the center of the figures, despite the presence of noise. This illustrates that symmetry is less susceptible to noise.

6.3.2 Problem 2: Noise Robustness

The problem of the SIFT detector with noise robustness is demonstrated in Figure 6.2. The first two columns show a number of clean and noisy stimuli along with the SIFT interest points. The noisy images are constructed by adding Gaussian pixel noise to ten percent of randomly chosen pixels. It is apparent that the addition of noise results in many spurious interest points detected by SIFT. Especially on the lower scales, the difference of Gaussians are strongly affected by the noise. These spurious points pollute the interest-point databases, making the matching slower and more prone to mismatches. The matching of SIFT interest points has indeed been shown to deteriorate in noise conditions (Kootstra, de Jong, & Schomaker, 2009; Kootstra & Schomaker, 2009b). In the next section, we demonstrate that detection of interest points with sym-

metry is more robust to noise. This is discussed in more detail in Chapters 8 and 9. The SIFT descriptor is also affected by noise (Bay et al., 2008). This is due to the fact that the noise distorts the local gradients that are used in the descriptor.

6.4 *Using Symmetry to Detect Interest Points*

In Chapter 3, we proposed a saliency model based on local symmetry in the image. We showed that the saliency maps calculated by the model correlate well with human eye fixations. This strongly suggest that humans pay attention to symmetrical parts of the visual field. If humans pay attention to symmetry, it might also be advantageous for machine vision to pay attention to symmetry.

It can be appreciated from the last two columns in Figure 6.2 that symmetry detection is much less affected by the addition of noise. These columns show the saliency maps as calculated with the symmetry-saliency model. The saliency maps show a clear peak at the centers of the forms, even with the addition of noise. This can be explained by the fact that symmetry is non-accidental. It is highly unlikely that the addition of noise results in a symmetrical configuration. The symmetrical shape of the stimulus stays visible even though parts of the form are distorted. It can be seen from the figure that the noise slightly affects the crispness of the peak. This is mainly the result of the distortion of the form's edges. However, the center clearly peaks out, and can be reliably detected under noise. This is also true for other symmetrical stimuli of different sizes and at different positions in the image. Even interest points on shapes that are not perfectly symmetrical can still be reliably detected in noisy conditions. Also Heidemann (2004) showed that interest points detected on the basis of symmetry are robust to noise and 3D object rotation. Moreover, he showed that symmetry detection results in points that are more robust to changing light conditions than Harris corners and that these points are more unique compared to other locations in the image. This motivates our choice to use local symmetry for interest-point detection.

Although contrast features have received most attention in computer-vision research (e.g., Lowe, 2004; Itti et al., 1998), symmetry is successfully used in a number of studies. In earlier work, for instance, Marola (Marola, 1989) used symmetry for detection and localization of objects in planar images. Symmetry has furthermore been used to control the gaze of artificial vision systems (Backer, Mertsching, & Bollmann, 2001; Sela & Levine, 1997). A number of symmetry operators are porposed in the

literature. The mirror-symmetry operator of Reisfeld et al. (1995) compares gradients of neighboring pixels to determine the amount of local symmetry at a give location in the image. His work has been extended by Heidemann (2004) to the color domain. Reisfeld et al. also proposed a radial-symmetry operator that is more sensitive to symmetrical patterns containing multiple symmetry axes. Loy & Zelinsky (2003) developed a radial-symmetry operator that is optimized for speed.

The symmetry operators discussed above detect symmetry on a single scale. This has the disadvantage that the methods are not scale invariant. In this part of the dissertation, we propose two symmetry detectors that detect symmetrical patterns on multiple scales, allowing recognition over a wide range of different scales. In Chapter 8, we propose the *MUlti-scale Symmetry Transform* (MUST) to detect symmetrical interest points and in 9 we propose the *Symmetrical Region-of-Interest Detector* (SymRoID) to detect symmetrical regions of interest. Both methods are used to select landmarks in the robot's environment.

6.5 *Visual Attention for Simultaneous Localization and Mapping*

The goal of Simultaneous Localization and Mapping (SLAM) is to have a robot build a map of an unknown environment, which it can use to localize itself on. The two elements, localization and mapping, are mutually dependent: On the one hand, the robot needs to know where it is to be able to build an accurate map of the environment. On the other hand, the robot needs a map to be able to localize itself. In visual SLAM, visual landmarks are usually selected in the environment to build a map representation (see Figure 6.3). Most approaches use interest-point detectors to select these landmarks. In (Kootstra et al., 2009; Kootstra & Schomaker, 2009b), we proposes interest-point detectors based on symmetry to select landmarks for visual SLAM. We elaborate on this in Chapters 8 and 9. In this section, we introduce the SLAM problem and methods to solve it.

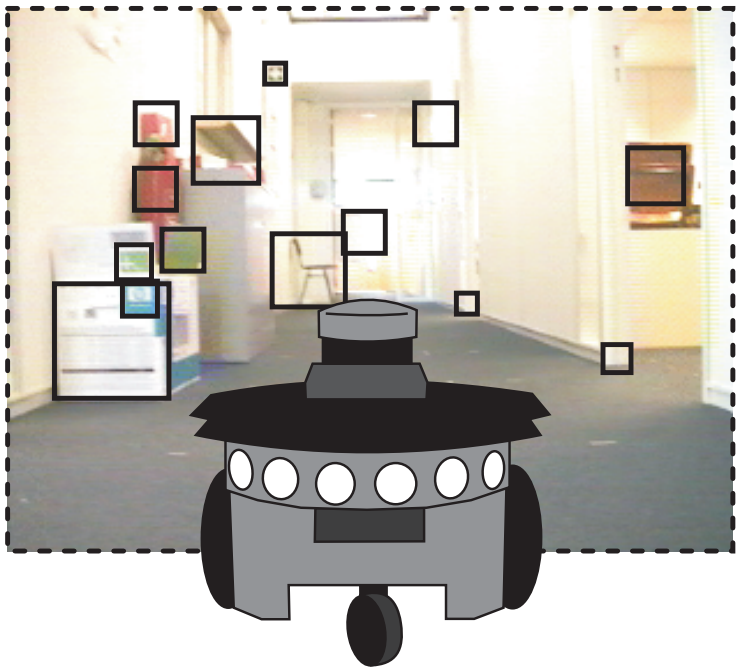

Figure 6.3: An illustration of the use of interest-point detectors to select landmarks in the environment for visual simultaneous localization and mapping (visual SLAM). The positions and descriptors of the selected landmarks are stored in the map.

6.5.1 The SLAM Problem

If the robot had perfect proprioceptive sensors to determine the change in pose and position and perfect exteroceptive sensors to map the environment, SLAM would not be a problem. The robot could simply make an observation of the environment and represent it in a map. Subsequently, it could travel in the environment and update its own position in the map accordingly. By repeating this sequence of two steps, an *observation step* and a *motion step*, the robot would build an accurate map of its environment. However, in reality, sensors are not accurate. Odometric sensors to determine the traveled path of the robot provide noisy information, and so do (exteroceptive) sensors to perceive the outside world, like sonars, laser-range finders, and video cameras. Con-

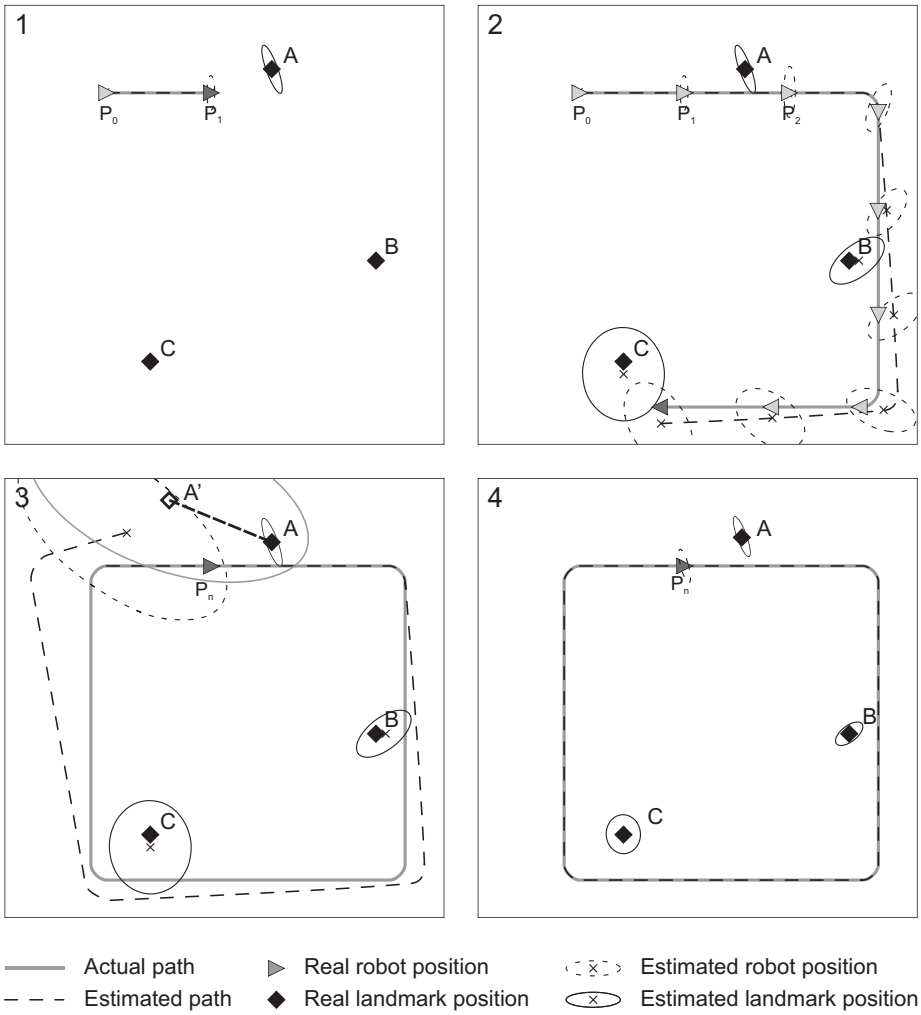

Figure 6.4: Mapping of the environment and loop closing. In the beginning, the robot's location can be estimated with high certainty, making the estimations of landmark positions certain as well (1). However, the uncertainty grows as a function of the traveled distance (2). At a certain moment, the robot re-observes a landmark that has been previously added to the map (3). At that moment, the position of the robot can be updated based on this observation. Since the re-observed landmark was estimated with high certainty, the robot's location can now also be estimated more reliably (4). Not only the current position of the robot, also all previous landmark observations can now be estimated with higher certainty.

sequently, the uncertainty in the map grows over time. In the beginning, the robot is relatively certain about its position, and the observations can therefore be placed in the map quite accurately. But with every displacement that the robot makes, it gets more uncertain about its own position, and therefore about the position of the observations.

An important element in SLAM algorithms to solve this problem of growing uncertainty is *loop closing*. The idea behind loop closing is depicted in Figure 6.4. The early observations of the robot are very certain (1), but uncertainty grows as a function of the traveled distance (2). As can be seen, the estimated path also deviates more and more from the true path of the robot. At a certain moment, however, the robot encounters a previous observation (3). At that moment, loop closing can take place. Since the location of the landmark was earlier estimated with high certainty, the position of the robot can be updated according to the re-observation (4). Importantly, this does not only result in a less uncertain estimation of the robot location, but the higher certainty can also be propagated back to the earlier observations, decreasing the uncertainty of all landmark estimations. By frequently closing loops, environments can be mapped and the robot's location can be estimated with sufficient certainty.

6.5.2 *Sensors to Perceive the Environment*

Many successful approaches to SLAM use *laser range finders* to map the environment (Thrun, Burgard, & Fox, 2005). A laser range finder is a device that accurately estimates the distance towards obstacles by sending a laser pulse and measuring the time taken by the pulse to be reflected off an object and return to the range finder. By sending modulated pulses in different directions, a laser range finder can measure the distance to obstacles in a 180 ° field of view. Due to its high accuracy, the laser range finder is a popular sensor for SLAM. It results in an accurate map of the environment, and therefore, the position of the robot can be updated accurately during loop closing. Moreover, the scans of the environment can be used to correct errors in the odometry by matching the current scan with a previous scan. The use of laser range finders has resulted in systems that successfully map large environments (Thrun et al., 2005).

Using vision for SLAM, however, remains a challenging topic (Frintrop & Jensfelt, 2008; Davison, Reid, Molton, & Stasse, 2007). A camera has the advantage over a laser range finder that it is a passive sensor that is low cost, low powered, and lightweight. A camera furthermore provides a rich source of information, which enables the use

of sophisticated detection and recognition methods. The challenge to use cameras for SLAM, however, is to extract relevant information from the high-dimensional visual data in real time. Moreover, the data provided by cameras is noisier and less accurate than that of laser range finders.

6.5.3 *Selecting Landmarks in the Environment*

Most visual SLAM systems use interest-point detectors to select distinctive visual landmarks in the environment. The positions of the landmarks in the environment are stored in the map, along with their descriptors. Most systems detect interest points on the basis of contrast, notably, the Scale-Invariant Feature Transform (SIFT) (Lowe, 2004; Se, Lowe, & Little, 2002), Speeded-Up Robust Features (SURF) (Bay et al., 2006; Murillo, Guerrero, & Sagues, 2007), and Harris corners (Davison & Murray, 2002).

In this dissertation, we propose the use of local symmetry to select landmarks in the environment (Kootstra et al., 2009; Kootstra & Schomaker, 2009b). This choice is motivated by the fact that man-made enviroments, and especially indoor environments, contain a lot of symmetrical structures and objects. This abundance of symmetrical forms can be exploited using local-symmetry detectors. Moreover, we already demonstrated that symmetry is less susceptible to noise, making it a good feature candidate to robustly detect and match visual landmarks.

The position of the landmark is often represented in the map by its estimated position in Euclidean space (Frintrop & Jensfelt, 2008; Durrant-Whyte & Bailey, 2006; Kootstra et al., 2009; Kootstra & Schomaker, 2009b), although other representations exist, such as the inverse-depth parametrization (Civera, Davison, & Montiel, 2008). The visual appearance of the landmark is usually represented using the SIFT descriptor (Se et al., 2002; Frintrop & Jensfelt, 2008; Kootstra et al., 2009; Kootstra & Schomaker, 2009b). Comparisons between landmarks in the current observation with landmarks in the map are made on the basis of these descriptors.

6.5.4 *Probabilistic Solutions to Visual SLAM*

The standard way to solve the SLAM problem is by using probabilistic Bayesian methods (Durrant-Whyte & Bailey, 2006; Bailey & Durrant-Whyte, 2006). In such methods, the uncertain knowledge about the state of the robot is combined with uncertain

observations in a statistically sound way. In the description of probabilistic SLAM, we will follow the definitions and notations in the SLAM tutorials of Durrant-Whyte and Bailey (Durrant-Whyte & Bailey, 2006; Bailey & Durrant-Whyte, 2006).

The goal of SLAM is to compute the probability distribution of the position of the robot and that of the landmarks in the environment, given the robot's initial position, its actions, and its observations. Mathematically, this is written as:

$$P(\mathbf{x}_k, \mathbf{m}_i | \mathbf{Z}_{0:k}, \mathbf{U}_{0:k}, \mathbf{x}_0), \tag{6.1}$$

where

- $\mathbf{x}_k = [x_k, y_k, \theta_k]^T$ is the state vector describing the location and orientation of the robot at time k.
- $\mathbf{m}_i = [m_{ix}, m_{iy}]^T$ is the location of landmark i.
- $\mathbf{X}_{0:k} = \{\mathbf{x}_0, \ldots, \mathbf{x}_k\}$, $\mathbf{U}_{0:k} = \{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$, $\mathbf{m} = \{\mathbf{m}_1^T, \ldots, \mathbf{m}_n^T\}^T$, $\mathbf{Z}_{0:k} = \{\mathbf{z}_1, \ldots, \mathbf{z}_k\}$.
- $\mathbf{u}_k$ is the motion vector, or the action, at time $k-1$, usually estimated by the robot's odometry.
- $\mathbf{z}_k$ is the observation of the location of landmarks taken by the robot at time k. This can contain multiple landmark observations.

The probability distribution is computed using a motion model and an observation model. Using the Markov assumption, the *motion model* gives the probability that the robot makes a transition to $\mathbf{x}_k$, given its previous location and the action taken:

$$P(\mathbf{x}_k | \mathbf{x}_{k-1}, \mathbf{u}_k). \tag{6.2}$$

The *observation model* describes the probability of making an observation $\mathbf{z}_k$ given the robot and all landmark locations:

$$P(\mathbf{z}_k | \mathbf{x}_k, \mathbf{m}). \tag{6.3}$$

The SLAM algorithm can be implemented, by iteratively applying the motion model in a prediction step, followed by applying the observation model in an update step.

The *prediction step* is:

$$P(\mathbf{x}_k, \mathbf{m} | \mathbf{Z}_{0:k-1}, \mathbf{U}_{0:k}, \mathbf{x}_0) = \int P(\mathbf{x}_k | \mathbf{x}_{k-1}, \mathbf{u}_k) \cdot P(\mathbf{x}_{k-1}, \mathbf{m} | \mathbf{Z}_{0:k-1}, \mathbf{U}_{0:k}, \mathbf{x}_0) d\mathbf{x}_{k-1}, \quad (6.4)$$

and the *update step*:

$$P(\mathbf{x}_k, \mathbf{m} | \mathbf{Z}_{0:k}, \mathbf{U}_{0:k}, \mathbf{x}_0) = \frac{P(\mathbf{z}_k | \mathbf{x}_k, \mathbf{m}) \cdot P(\mathbf{x}_k, \mathbf{m} | \mathbf{Z}_{0:k-1}, \mathbf{U}_{0:k}, \mathbf{x}_0)}{P(\mathbf{z}_k | \mathbf{Z}_{0:k-1}, \mathbf{U}_{0:k})}. \quad (6.5)$$

There are two main methods to the probabilistic SLAM problem that implement the motion model, (6.2), and the observation model, (6.3). The *extended Kalman filter* (*EKF-SLAM*) (Dissanayake, Newman, Durrant-Whyte, Clark, & Csobra, 2001) represents the joint distribution $P(\mathbf{x}_k, \mathbf{m} | \mathbf{Z}_{0:k}, \mathbf{U}_{0:k}, \mathbf{x}_0)$ as a Gaussian distribution. Also the motion and observation models are Gaussian distributed. The *Rao-Blackwellized particle filter* (Sim, Elinas, & Little, 2007), or *FastSLAM* algorithm (Montemerlo, Thrun, Koller, & Wegbreit, 2003, 2002), applies Monte-Carlo techniques for a non-Gaussian representation. The method is a combination between a particle filter for the representation of the robot's state and individual EKFs for the representation of the position of the landmarks. Since the extended Kalman filter is used in this dissertation, it will be discussed in more detail in the next section. We furthermore describe the Rao-Blackwellized particle filter, followed by a discussion on ambiguous observations.

6.5.4.1 The Extended Kalman filter (EKF-SLAM)

The extended Kalman filter represents the joint distribution $P(\mathbf{x}_k, \mathbf{m} | \mathbf{Z}_{0:k}, \mathbf{U}_{0:k}, \mathbf{x}_0)$ as a Gaussian distribution. The state vector $[\mathbf{x}^T, \mathbf{m}^{\mathbf{T}^T}]$ and its full covariance matrix is estimated in three steps, the prediction step, the update step, and the *augmentation* step. The covariance gives the uncertainty about the position of the robot and the landmarks in the map and the relations among them. In the prediction step, the position of the robot is updated based on odometric information. This does not only result in an adjustment of the state vector, but also in the covariance matrix, since the motion of the robot increases the uncertainty about its position. In the update step, the position of the robot and the landmarks are updated as well as their uncertainties based on the observation of known landmarks. The augmentation step is executed if new landmarks are observed. In that step, the state vector and covariance matrix are expanded to

include the new landmarks.

EKF-SLAM is used in Chapter 8 and 9 of the thesis. It is therefore discussed in detail in Appendix C. For more information on EKF-SLAM, the interested reader is referred to (Thrun et al., 2005; Durrant-Whyte & Bailey, 2006; Bailey & Durrant-Whyte, 2006).

6.5.4.2 The Rao-Blackwellized Particle Filter

The joint distribution $P(\mathbf{x}_k, \mathbf{m}|\mathbf{Z}_{0:k}, \mathbf{U}_{0:k}, \mathbf{x}_0)$ is represented in the Rao-Blackwellized particle filter (RBPF) by a set of particles $\mathbf{p}^{(i)} = \{\mathbf{w}_k^{(i)}, \mathbf{X}_{0:k}^{(i)}, P(\mathbf{m}|\mathbf{X}_{0:k}^{(i)}, \mathbf{Z}_{0:k})\}_{i=1}^{N}$. In other words, each particle holds an importance weight, an estimate of the robot trajectory, and a map of the environment based on the trajectory and the observations. The map is represented by independent Gaussian distributions for every landmark. The combination of a particle filter with Gaussian representations for the landmarks is for efficiency reasons. The dimensionality of the state space used by the particle filter is reduced, resulting in less particles needed to estimate the state vector.

Similar to the EKF, the RBPF iteratively applies a prediction, an update step, and an augmentation step. The RBPF furthermore resamples the particle population, in order to keep good and remove bad particles. In the FastSLAM 1.0 implementation of the RBPF (Montemerlo et al., 2002), the prediction step is done for all particles by sampling the next robot-pose estimation from the motion model, based on the previous pose estimation and the robot's action. Due to the sampling, the motion model is not limited to a Gaussian distribution, but can be any distribution. In the update step, the weights of all particles are updated based on the likelihood of the observations made by the robot given the pose estimation and the map. Furthermore, for all particles, an independent EKF update is performed for all observed landmarks with the pose estimation taken as the known robot pose. The augmentation step is straightforward since all landmarks distributions are independent. Finally, the particle population is resampled by importance sampling. A new particle is drawn from the population with replacement, where the probability that a particle is selected is proportional to its importance weight. The resampling does not take place every iteration, but after a fixed number of steps or at loop closure. FastSLAM 2.0 (Montemerlo et al., 2003) holds some improvements in efficiency over its predecessor.

For more information on RBPF and FastSLAM, we refer to (Thrun et al., 2005; Sim et al., 2007; Montemerlo et al., 2002, 2003).

6.5.4.3 *EKF-SLAM vs FastSLAM*

The EKF method offers a good mathematical solution to the SLAM problem, which is optimal under the assumption of Gaussian noise. Also FastSLAM offers a good solution, given sufficient particles to represent the posterior distribution.

One of the problems with EKF-SLAM is that the computational complexity grows quadratically with the number of landmarks, due to the update of the full state-covariance matrix in the update step. However, efficient implementations have been demonstrated (E.g., Guivant & Nebot, 2001). The extended Kalman filter can also be efficiently approximated using a *sparse extended information filter* (Thrun, Liu, Koller, Ng, Ghahramani, & Durrant-Whyte, 2004). This method uses the information matrix, which is the inverse of the state-covariance matrix. It takes advantage of the fact that this matrix has many off-diagonal values that are near zero, which can be exploited for sparsification. Another problem with EKF-SLAM is the Gaussian noise assumption and the unimodal Gaussian pose representation, which conflict with noise and uncertainty in real-world applications.

The advantage of FastSLAM over EKF-SLAM is that it uses a nonlinear and non-Gaussian motion model and a multimodal and non-Gaussian pose distribution. The landmarks are furthermore independently represented in the map, which results in a computational complexity linear in the number of landmarks. However, these advantages come with the costs of needing a large number of particles for a good approximation of the posterior distribution, especially in large environments, which increases the computational complexity.

6.5.4.4 *Ambiguous Observations*

EKF-SLAM cannot deal with ambiguous situations, situations when an observation is associated with more than one location. This is due to the fact that the position of the robot is represented by a Gaussian distribution, which does not allow multiple pose hypotheses. In reality, however, ambiguous situations do occur. Consider for instance a hallway in an office building with many identicaly looking doors. This will result in many identical landmarks in the map. The re-observation of such a landmark will cause multiple hypotheses of the robot position. This can, however, not be represented by a unimodal Gaussian distribution.

The advantage of the particle filter used by FastSLAM is that it represents the robot state by a number of particles, each representing a hypotheses of the robot position. The particle filter can therefore represent multimodal distributions. However, the standard particle filter suffers from *premature convergence*. The randomness in the particle-resampling process results in *random genetic drift*, which soon lets the particle population converge to a single solution. The standard particle filter is therefore also unable to represent multiple position hypotheses for an extended period of time. By applying *niching methods* known in the field of *genetic algorithms*, the diversity in the particle population can be maintained (Kootstra & de Boer, 2009).

The extended Kalman filter is udef in the visual SLAM systems presented in Chapters 8 and 9. Although improvements in performance can be seen with the use of FastSLAM in recent experiments (Wedema, 2009), the main interest is in the selection of landmarks and not so much in the SLAM techniques. When the interest points used can be detected more robustly and stably, this will give an increase in SLAM performance no matter what SLAM technique is used. The choice of uncerlying SLAM mechanism does therefore not influence the study of interest point detection.

6.6 *Vision as an Active Process*

The importance of active vision has been advocated by Ballard (1991), who used the term *animate vision*. According to Ballard, perceptual tasks can be greatly simplified using active perception. Many natural systems use their active capabilities to control the input of their sensors. In this way, the system can perceive things that it could not perceive in a passive way due to limitations of the sensory system or to computational limitations.

This dissertation started with a quote of Gibson stating that perception is an active process. The models for visual attention and interest-point detection proposed in this thesis are in a sense active-vision models. The models determine where to focus attention. In other words, the models determine what the visual input to the next processing stages will be. However, the attention is covert: no physical action is taken, but the focus of resources is purely mental. In this section, we give a number of examples of physical active-vision systems. Three artificial systems are presented, inspired by natural systems, that use active vision to solve perceptual tasks.

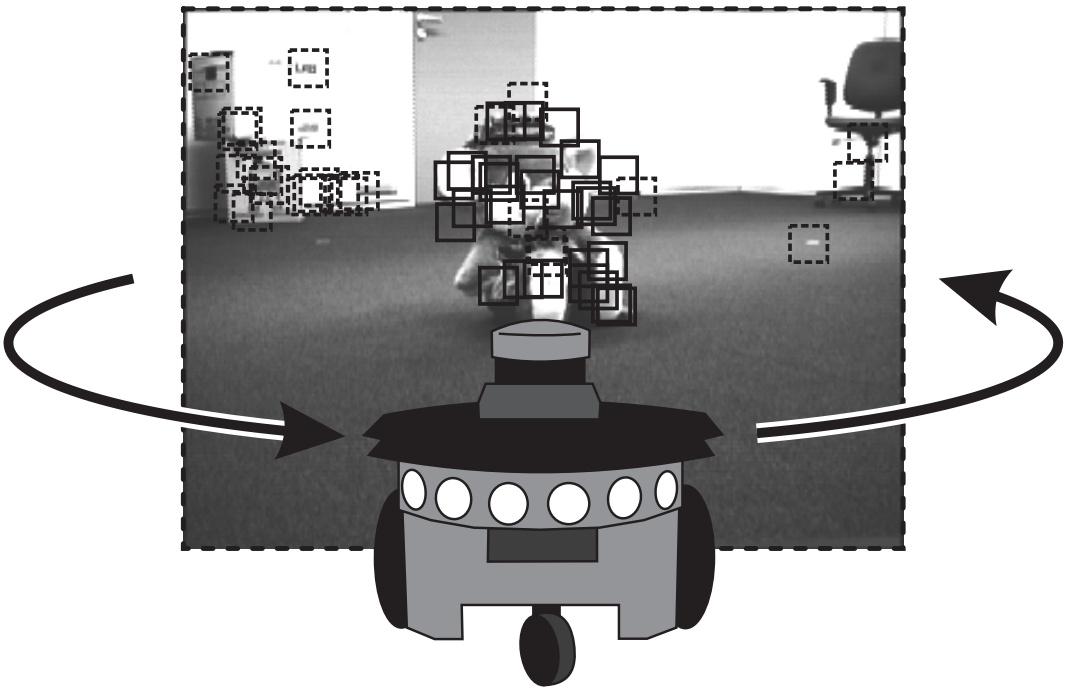

Figure 6.5: Using active vision to select stable interest points belonging to the object. The solid squares show the stable object points, whereas the dashed squares are the interest points that are filtered out.

6.6.1 Active-Vision Systems

A very clear example of the use of active vision is the peering behavior of a locust (Sobel, 1990). The eyes of a locust are close together and have little overlap in visual field. It can therefore not estimate distances in a passive way using stereo vision. To be able to determine the force needed to jump to a distant landmark, the locust performs a peering behavior, that is, it moves its head from left to right in a translational movement. The depth information can thus be obtained by the induced motion parallax. When the target was moved in the direction opposite to the locust's head movement, more parallax was artificially created, simulating a target closer to the locust. This resulted in a decrease of the jump velocity Sobel (1990). Based on these results, Lewis

Figure 6.6: An 18-month-old child exploring an unknown object. This enables the child to observe the object from different viewpoints. In the meanwhile, it makes it possible to discriminate the object from the background.

& Nelson (1998) developed a computational model for distance estimation on a mobile robot using a monocular camera. The robot performed peering behavior and the motion parallax was estimated using the biologically inspired Reichardt motion detector (Borst & Egelhaaf, 1993).

Another example of the use of active vision in natural systems is the *turn-back-and-look behavior* displayed by honeybees (Lehrer, 1993). When a bee leaves a food source that it visited for the first time, it does not directly fly back to the hive, but turns around to store a visual representation of the location. The bee performs a side to side movement, while gradually moving backwards, away from the food source. Doing so, the image motion induced by the behavior provides the bee with information about the three-dimensional structure of the source's surroundings. This enable the bee to select reliable landmarks to represent the location. These findings are used by Lehrer & Bianco (2000) to test the stability of visual landmarks by letting a robot perform the turn-back-and-look behavior. The motion parallax induced by the behavior is used by Kootstra (2002) to selected nearby landmarks, since nearby landmarks more precisely determine the target location than distant landmarks. The stability of interest-points for landmark selection in visual SLAM is tested in a similar way in (Kootstra et al., 2009; Kootstra & Schomaker, 2009b). This is explained in more detail in Chapter 8.

Figure 6.6 shows an example of an 18-month-old boy exploring a new toy. Instead of passively observing the toy, the child actively explores the object. Doing so, he gathers information about the object from many different viewpoints, thereby solving the

problem that an object can look completely different from different sides. Moreover, by exploring the object, it becomes visually clear what belongs to the object, and what to the background. The manipulation of objects is used by (Metta & Fitzpatrick, 2003) to let a robot learn to segment an object from its background. The exploration behavior was also an inspiration to develop an active-vision method for three-dimensional object recognition in the real world proposed in Chapter 7. By letting a robot circle around an object (see Figure 6.5), it achieves three things (Kootstra et al., 2008b, 2007). Firstly, the object can easily be segmented from the background using a *what-moves-together-belongs-together* approach. Secondly, the stability and robustness of the interest points can be tested to only store points that are observable from different viewpoints. Finally, more evidence for recognition is gathered from different viewpoints. This is explained in detail in Chapter 7.

6.7 Conclusion

In this chapter we introduced and discussed some computational aspects of visual attention and active vision. Visual attention can be used to focus computational resources on interesting parts of the visual field. This is used in interest-point models for image representation, such as the scale-invariant feature transform (SIFT), which is one of the state-of-the-art interest-point detectors and descriptors. However, SIFT has some problems, most notably the susceptibility of the detector to noise. The use of local symmetry has been argued to results in interest points that are more robust to noise.

In simultaneous localization and mapping (SLAM), the position of the robot and that of landmarks in its environment needs to be estimated. Two main probabilistic methods have been proposed in the literature, the extended Kalman filter (EKF-SLAM), and the Rao-Blackwellized particle filter (FastSLAM). The challenge of visual SLAM is to select stable landmarks to represent the environment. In Chapter 8 we propose an interest-point detector based on local symmetry in the image to select landmarks for visual SLAM and compare it to SIFT. The use of symmetry turns out to be less susceptible to noise and result in better SLAM performance. In Chapter 9, we build upon that result and develop a region-of-interest detector based on symmetry. Using regions instead of points is shown to be even more robust to noise. Moreover, it results in a higher stability, and most importantly in a better SLAM performance.

Active vision has the ability yo simplify visual tasks. In the next chapter, Chapter 7, this is exploited to represent and recognize three-dimensional objects in the real world. The proposed active-vision model is shown to improve recognition of objects in cluttered environments.

7

Active Object Recognition by Exploration

Abstract

Object recognition is a challenging problem for artificial systems. This is especially true for objects that are placed in cluttered and uncontrolled environments. To solve this problem, we discuss an active approach to object recognition in this chapter. Instead of passively observing objects, we use a robot to actively explore the objects. This enables the system to learn objects from different viewpoints and to select viewpoints for optimal recognition. Active vision furthermore simplifies the segmentation of the object from its background. As the basis for object recognition we use the Scale-Invariant Feature Transform (SIFT). SIFT has been a successful method for image representation. However, a known drawback of SIFT is that the computational complexity of the algorithm increases with the number of interest points. We discuss a growing-when-required (GWR) network for efficient clustering of interest points to reduce the size of the database. The results show successful learning of three-dimensional objects in real-world environments. The active approach is successful in separating the object from its cluttered background, and the active selection of viewpoint further increases the performance. Moreover, the GWR network strongly reduces the number of interest points.

This chapter is based on:

Kootstra, G., Ypma, J., & de Boer, B. (2008b). Active exploration and keypoint clustering for object recognition. In *International Conference on Robotics and Automation (ICRA)*. Pasadena, CA.

Kootstra, G., Ypma, J., & de Boer, B. (2007).Exploring objects for recognition in the real world. In *IEEE International Conference on Robotics and Biomimetics (ROBIO '07)*. Sanya, China.

7.1 Introduction

The real world poses many challenging problems for artificial systems. Consider for instance the problem of recognizing objects in the real world. Many object recognition systems that are successful in controlled laboratory environments have problems with the uncontrolled and unpredictable properties of the real world. Whereas, for instance, illumination and background can be controlled in an artificial setting, this is not true for real-world environments. Visual perception, therefore, becomes more challenging in the real world. Natural systems deal with these challenges by using active perception (Gibson, 1979). Instead of passively observing an object, many animals, including humans, explore the object to control the visual input (see figure 6.6). The use of active perception is also very important for artificial systems (Ballard, 1991; Pfeifer & Scheier, 1999). This chapter, discusses an active approach to three-dimensional (3D) object recognition in the real world by an autonomous robot. By actively changing its viewpoint, the robot observes an object from different angles, making it possible to learn to recognize the object from any given viewpoint. Moreover, the system selects the viewpoint that is expected to be most informative for recognition. Furthermore, exploration of the object makes it possible to separate the object from its background, something that is non-trivial when passively observing an object on a highly cluttered background (Metta & Fitzpatrick, 2003).

Like many current approaches to object recognition, our model describes the objects by a set of local interest points (Harris & Stephens, 1988; Lowe, 1999; Schmid & Mohr, 1997). Description in terms of local interest points has the advantage that the representation is more robust to occlusions, clutter and noise. It is also less sensitive to changes in viewpoint. In our method we use the Scale-Invariant Feature Transform (SIFT) for the detection and description of interest points (Lowe, 2004). The choice for SIFT is motivated by the fact that it is a well-known and widely-used interest-point method and the focus of this chapter lies on active vision and not on interest-point methods. Our approach is, however, not restricted to SIFT, but can also be used with other local image detectors and descriptors, like those discussed in Chapter 8 and 9.

Interest points have been successfully used for three-dimensional (3D) object recognition (Ferrari, Tuytelaars, & Van Gool, 2006; Lowe, 2001; Moreels & Perona, 2007; Rothganger, Lazebnik, Schmid, & Ponce, 2006). These studies have demonstrated the ability to learn to recognize objects from multiple viewpoints and subsequently recog-

nize these objects in cluttered scenes. However, learning in these studies takes place in well-controlled environments: the object is usually put on a turntable which carefully rotates the object, while taking pictures of the object with fixed lighting conditions — with the exception of (Moreels & Perona, 2007) — and against a uniform background. This setup reduces the amount of noise and uncertainty and makes it trivial to separate the foreground from the background. It is therefore not representative for real-world environments. A real-world environment is usually highly uncertain and cluttered with many distracting features. In this chapter, we present a method to learn objects in uncontrolled real-world environments, using active vision. We use a mobile robot to actively explore the objects and their environment.

Our approach uses active vision in multiple ways. Firstly, we use it to separate the object from its background, similar to (Fitzpatrick, 2003; Metta & Fitzpatrick, 2003). We use a method that can be described as *what-moves-together-belongs-together*. The robot observes the object while circling around it, a behavior that is comparable to rotating an object in your hand (see figure 6.6). By doing so, the observer learns the appearance of the object from different viewpoints. This solves the *object-constancy problem*, the problem that objects appear very different from different viepoints (see Figure 7.1). The exploration of the object enables the system to link the different perspectives and build a representation of the full 3D object.

Secondly, while performing the explorative behavior, interest points belonging to the object will show little displacement on the camera image, since the object is near the center of rotation. Interest points in the background, on the contrary, show relatively large displacements, with the possible exception of points on the floor close to the object. The amount of displacement of an interest point is used to classify whether it belongs to the object or to the background.

Thirdly, active vision is used to find stable interest points. By changing viewpoints, the robot can actively test whether an interest point is recognizable from nearby viewpoints. If so, the interest point can be classified as stable. This process will filter out points that are sensitive to rotation, translation, and other affine transformations. This reduces the necessity to use affine-invariant interest point detectors (e.g., Mikolajczyk & Schmid, 2002), which are not only computationally expensive, but have also been shown to perform worse on recognizing non-planar 3D objects than SIFT (Moreels & Perona, 2007). A similar approach to the one presented here was taken in (Lehrer & Bianco, 2000), where a behavior, inspired by insects, was adopted to find reliable

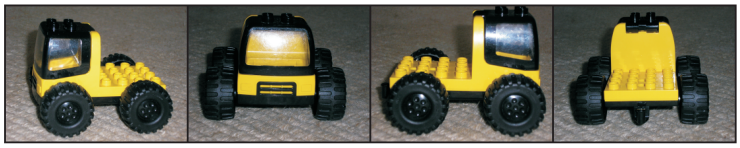

Figure 7.1: The object-constancy problem. An object appears very differently from different viewpoints. By using an active method for learning and recognition, the different perspectives can be integrated.

visual landmarks.

Finally, active vision is used to gather more evidence for recognition. This is especially important in ambiguous situations. If from one viewpoint it is not possible to recognize an object, a more promising viewpoint can be selected. Although human observers might have the impression that ambiguous situations are quite rare, we must remember that ambiguity strongly depends on the quality of the sensory system, as can be seen in (Nolfi, 1996). In 3D object recognition, gathering more evidence improves recognition (Roy, Chaudhury, & Banerjee, 2004). Some viewpoints will be more informative than others. We therefore propose a probabilistic method to select the viewpoint that is expected to be most informative as the next viewpoint. Viewpoint selection is also used by (Borotschnig, Paletta, Prantl, & Pinz, 2000; Paletta & Pinz, 2000). The difference with the model presented here is that our model uses a one-shot learning method and perform recognition in a real-world environment.

The proposed method to learn objects by circling around it requires information about the position of the object. This chicken-and-egg problem can be solved by human guidance, like done in (Van Hoof, 2008; Zwinderman, Rybski, & Kootstra, submitted). By letting a human teacher initially denote the object, the position of the object can be determined. Subsequently, the robot can explore the object. However, since we are mainly interested in the possibilities of the explorative behavior, the objects are placed at a fixed position from the robot in the presented research.

In addition to the use of active vision, we propose a method to reduce the number of points in the interest-point database. One of the reasons that SIFT is so successful in object recognition is that it uses a large number of interest points to represent one object (Lowe, 2004). This makes the system very tolerant to noise, and reduces the problem of occlusions. There is, however, an important drawback, namely that a significant

amount of computation in the recognition process is devoted to matching the observed points with the interest-points database. Nearest-neighbor search methods like kd-tree search (Friedman, Bentley, & Finkel, 1977) that are efficient in low-dimensional spaces, do not do better than exhaustive search in the high-dimensional space of the SIFT features. An improvement in computation time can be achieved by an approximate best-bin-first method (Beis & Lowe, 1997). But even then, the computation time increases with the number of stored interest points, while the success in finding the nearest neighbor decreases. It is therefore very useful for 3D object recognition to reduce the number of stored interest points in an efficient way.

In this chapter, growing-when-required (GWR) network (Marsland et al., 2002) is used for efficient clustering of interest points. When performing 3D object recognition, many of the acquired interest points look very similar. There are several reasons for this. First of all, these are interest points belonging to the same point on the object seen from different angles. Secondly, there are similarly looking points on repeating structures on the same object, and finally, different objects can have ambiguous interest points. The GWR network clusters these similar interest points to attain efficient database use.

7.2 Object Recognition by Exploration

In this section the active approach to object recognition is discussed first. Then the method to select the next viewpoint. The section ends with a description of the method to cluster the SIFT interest points.

7.2.1 The Scale-Invariant Feature Transform

The SIFT detector and descriptor (Lowe, 2004), described in Appendix B, are used as the basis for the 3D object recognition. The method used for matching the observed interest points with the database is somewhat different. Firstly, the method focuses solely on the individual matching of interest points, and therefore does not use the geometric matching of sets of points as used by Lowe. Secondly, a threshold on the distance to the nearest neighbor is used, instead of a best to second-best ratio to determine a match, since this yielded better performance in the experiments. Details on the matching and recognition processes used are described further on.

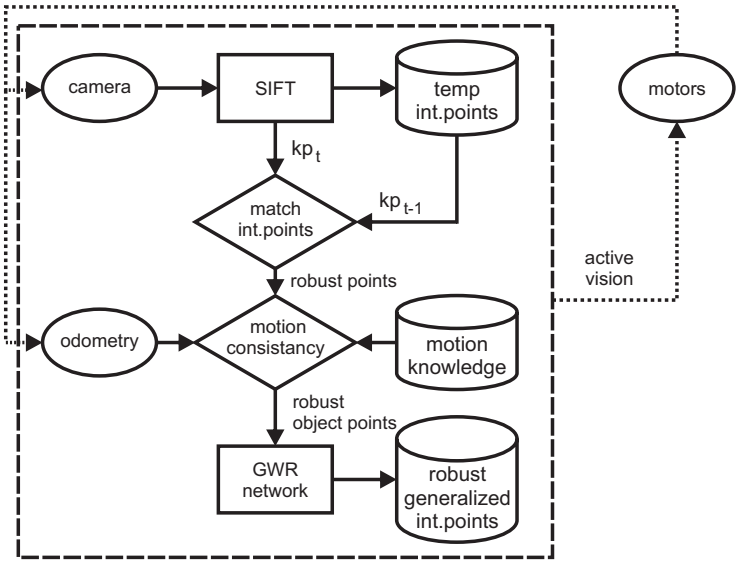

Figure 7.2: The active vision model to select stable object interest points. While the robot explores the object, SIFT interest points are detected in the current camera image. The interest points are stored in a temporal database. The current interest points are compared with the previous ones. An interest point is considered stable if it is observed in the current and in the previous and next observation. The stable interest points are then divided into background and foreground by comparing their image motion with knowledge about the expected motion. After clustering the interest points, the robust interest points belonging to the object are stored in the database.

7.2.2 Active Vision

By actively changing viewpoint, the robot gathers new information that we use in two different ways: to detect stable interest points and classify them as object or background, and to explore the object in order to gather more evidence to resolve ambiguous situations. Both methods are described in the following paragraphs. The general architecture is depicted in Figure 7.2.

An interest point is considered stable if it is originally observed at an angle of θ degrees, and subsequently matched in the previous or next image, at $\theta \pm 10$ degrees. An

interest point $\mathbf{k}_i$ is matched to its nearest neighbor in the previous image, $\mathbf{k}_\mathrm{n}$, if the Euclidean distance between both is less than 0.6, where $\mathbf{k}$ is the 128 dimensional feature vector of the interest point. This threshold is established experimentally. This filters out all interest points that are only recognizable from one specific angle.

In the next step, we segment the stable interest points belonging to the background from those belonging to the object. Each interest point $\mathbf{k}_i$ has a position (x_i, y_i) at which it is observed in the image. Since the object is in the center of rotation, object points will move little when the robot is exploring the object, whereas the displacement will be relatively large for interest points in the background. Furthermore, since the robot moves on a flat surface, points will only move in the horizontal direction. Allowing some fluctuations, we classify a stable interest point as an object point when

$$(|x_i - x_n| < x_\mathrm{T}) \wedge (|y_i - y_n| < y_\mathrm{T}) \tag{7.1}$$

where we use $x_\mathrm{T} = 12$ and $y_\mathrm{T} = 4$ pixels (the resolution of the camera image is 360×240). Otherwise, the stable interest point is classified as background. The successful use of this simple classification model nicely illustrates the power of active vision to simplify perceptual tasks. The robot explores the objects from 36 different angles and stores the stable object points along with the object ID and pose. Doing so, the appearances of objects in a cluttered environment are learned.

Once the object database is in place, objects can be recognized. Based on the set of observed interest points, $\mathcal{O}$, and the interest-point database, $\mathcal{D}$, we determine the activation of every model, $m_{b,\theta}$, for object b, and pose θ. The activation of a model is based on the set of observations, $\mathcal{M}_{b,\theta} \subseteq \mathcal{O}$, that supports the model.

$$\mathcal{M}_{b,\theta} = \bigcup (\mathbf{p}_i \in \mathcal{O} | o_n = b \wedge \alpha_n = \theta) \tag{7.2}$$

where o_n and α_n are respectively the object ID and pose of the nearest neighbor $\mathbf{k}_\mathrm{n}$ of $\mathbf{p}_i$ in the interest-point database. Every supporting observation $\mathbf{p}_i$ in $\mathcal{M}_{b,\theta}$ gives an activation a_i:

$$a_i = \exp(-|\mathbf{p}_i - \mathbf{k}_\mathrm{n}|) \tag{7.3}$$

This form is in conformity with the activation calculation used in the growing-when-required network discussed in Section 7.3. The total activation of model $m_{b,\theta}$ given

the observed interest points, $\mathcal{O}$, form viewpoint δ, and the interest-point database $\mathcal{D}$, is given by:

$$A^{\delta}_{b,\theta} = \frac{\sum_{i \in \mathcal{M}_{b,\theta}} a_i}{\sqrt{|\mathcal{D}_{b,\theta}|}} \tag{7.4}$$

where $|\mathcal{D}_{b,\theta}|$ is the number of points in the interest-point database that are associated with object b and pose θ, and δ is the current viewpoint. Equation (7.4) gives the activation of a specific pose of an object. This activation increases with the number of supporting observations relative to the number of interest points in the database associated with that object/pose. This makes that fewer matched observations are needed for objects that have relatively few interest points. However, the square root in the denominator causes the probability to increase with the number of object points given the same ratio of matched observations to database points. This reflects the idea that there is more confidence when there are more object points.

Finally, the robot can actively gather more evidence for recognition. By rotating around the object, the robot gathers more information about the object under consideration by viewing it from different angles. When driving around the object, we accumulate the evidence by:

$$A_{b,\theta}(t) = \sum_{\delta inE} A^{\delta}_{b,\theta} \tag{7.5}$$

where $A_{b,\theta}(t)$ is the accumulated activation for object b and pose θ at time t and $E = \{\phi_0, \cdots, \phi_t\}$ is the set of viewpoints from where the observations are made. The change of viewpoint helps to disambiguate object and is therefore expected to result in more robust recognition of 3D objects. In the next section, we will discuss how the next viewpoint is selected by our model.

In the above, the activation of an object in a particular pose is calculated. To obtain the activation of the full object model, the activations of all poses, Θ, for that object are summed:

$$A_b(t) = \sum_{\theta \in \Theta} A_{b,\theta}(t) \tag{7.6}$$

Figure 7.3 gives an overview of the interest-point database during learning and during

recognition.

7.2.3 Next-Viewpoint Selection

We use the active capabilities of the robot to explore the objects and gather more information from different viewpoints. In order to select the next viewpoint, we use a probabilistic approach. In this approach, the next viewpoint ϕ_{t+1} is the angle from where we expect the maximum activation of an objects-pose model, that is:

$$\phi_{t+1} = \underset{\gamma \in (\Theta - \Phi)}{\arg\max}\, E(A_{b,\theta}(t+1)) \tag{7.7}$$

where Θ is the set of all possible viewpoints, and $\Phi = \phi_0, \ldots, \phi_t$ is the set of all previous viewpoints. The expected activation of the object-pose model when viewed from viewpoint γ at time $t+1$ is given by:

$$E(A_{b,\theta}(t+1)) = A_{b,\theta}(t) + E(A^{\gamma}_{b,\theta}|O_{b,\theta})P(O_{b,\theta}) \tag{7.8}$$

$$E(A^{\gamma}_{b,\theta}|O_{b,\theta}) = \sqrt{|\mathcal{D}_{b,\theta+\gamma}|} \tag{7.9}$$

$$P(O_{b,\theta}) = \frac{A_{b,\theta}(t)}{\sum_{i=0}^{N} A_{o_i,\alpha_i}(t)} \tag{7.10}$$

In words, the expected new activation of the model is based on the old activation. This value is increased with the expected extra activation gained from the new viewpoint given that we are looking at object b and pose θ, $E(A^{\gamma}_{b,\theta}|O_{b,\theta})$, multiplied by the probability that we are actually looking at object b at pose θ, $P(O_{b,\theta})$. Equation (7.9) can be inferred from equation (7.4) when we assume that we observe all interest points belonging to object b at pose $\theta+\gamma$. Although this is not the case in reality, we can assume that a constant proportion of the interest points are observed. The next-viewpoint choice is not influenced by this constant factor. And the probability $P(O_{b,\theta})$ is the activation of the model divided by the total activation of all object-pose models. By selecting the viewpoint that optimizes the expected activation of one of the object-pose models, we select the most informative viewpoint as the next.

7.3 Clustering Interest Points: Growing When Required

As explained in the introduction of this chapter, 3D object recognition with SIFT has the main disadvantage that the computational time needed increases with the number of interest points stored in the database. We therefore use a growing-when-required (GWR) network (Marsland et al., 2002) to efficiently cluster interest points that are highly similar. A GWR network is a clustering method, very similar to a growing-neural-gas (GNG) network (Fritzke, 1995). Both networks are based on Kohonen's self-organizing maps (SOM) (Kohonen, 1990). A SOM is an efficient and unsupervised method to cluster high-dimensional data. The disadvantage, however, is that the number of clusters (i.e., nodes in the map) needs to be set in advance. This makes a SOM highly inappropriate for object recognition with SIFT, since the number of clusters depends on the number of unique interest points. A GNG-network is an adaptation of a SOM which can dynamically change the number of nodes in the network. However, the drawback of a GNG-network is that new nodes are only added after a number of inputs. This is not desirable for object recognition, since we would like to add a node in the network when we observe a completely new interest point. A GWR network does just that, it adds nodes if this is required.

The GWR network as described in (Marsland et al., 2002) uses edges between nodes. This is based on the SOM, and results in a topology preserving network in the sense that connected nodes in the network correspond to neighboring points in the input space. Usually, the connections between nodes in a GWR network are used in the learning process to move the neighboring nodes of the winning node closer to the presented input. Although this provides a better distribution of the nodes over the input data, it is undesirable for object learning, since in that case the presentation of an input not only changes the representation of the corresponding interest point, but also of neighboring interest points. This will result in changing the interest points so much that they do no longer correspond with the original input. Since this will impair recognition, we omitted the edges from the GWR network.

For the description of our implementation of the GWR network, we follow the notation and description in (Marsland et al., 2002). Let K be the set of observed interest points when learning the objects, A be the set of nodes in the network, $\mathbf{w}_n$ be the weight vector of node n (of the same dimensionality as the SIFT interest points), and t_n be the activation counter. Furthermore, each node holds a record, R_n, of all associated objects

and poses. We initialize the network with $A = n_1$, where the weight vector of n_1 is initialized with a randomly picked interest point from K, and $t_1 = 0$. Then, for each interest point $\mathbf{k}$ from K, we do:

1. $\mathbf{k}$ and the object b in pose θ to which the interest point belongs are input to the network.

2. Select the best matching node $s \in A$, such that

$$s = \arg\max_{n \in A} |\mathbf{k} - \mathbf{w}_n|.$$

3. Calculate the activity of the winning node

$$a_s = \exp(-|\mathbf{k} - \mathbf{w}_n|)$$

4. Calculate the firing counter is calculated by the decaying function:

$$h_s = 1 - (1 - \exp(-\alpha_b t_s / \tau))/\alpha_n$$

   where the shape of the function can be set using the parameters. In the experiments, we used: $\alpha_b = 1.05$, $\alpha_n = 1.05$, and $\tau = 3.33$.

5. If the activation of the winning node is low ($a_s < a_\mathrm{T}$), that is when the input does not really match any of the existing clusters, and when the winning node is mature ($h_s < h_\mathrm{T}$), create a new node r. Add r to the network, initialize its weight with $\mathbf{k}$, and set $\langle b, \theta \rangle$ as the reference list:

$$\begin{aligned} A &= A \cup r \\ \mathbf{w}_r &= \mathbf{k} \\ R_r &= \{\langle b, \theta \rangle\} \end{aligned}$$

   where the thresholds are $a_\mathrm{T} = 0.8$ and $h_\mathrm{T} = 0.4$.

6. Else, adapt the weights of the winning node and add $\langle b, \theta \rangle$ to the reference list:

$$\begin{aligned} \mathbf{w}_s &= \mathbf{w}_s + \eta \cdot h_s \cdot (\mathbf{k} - \mathbf{w}_s) \\ R_s &= R_s \cup \{\langle b, \theta \rangle\} \end{aligned}$$

where the parameter that determines the adaptivity of the node is $\eta = 0.05$.

7. Proceed to the next cycle:

$$t_s = t_s + 1$$

When the presented interest point is sufficiently similar to the winning node, it is clustered with that node, and the description of the node is altered to better represent all associated interest points. If, on the other hand, the presented interest point differs from the existing nodes causing the activation of the winning node to be below the threshold a_T, and the firing counter of the nearest node is below the threshold h_T, the presented interest point is added as a new node. In this way, the GWR network clusters similar interest point, thus creating a smaller database for recognition.

A record is kept of all objects and poses that correspond to the nodes in the network. This allows for supporting all objects containing similarly looking interest points when such a point is observed. This is in contrast with (Lowe, 2004), where important evidence is discarded by choosing only interest points that match uniquely with one object.

Figure 7.3 gives an overview of the interest-point database both with and without the use of the GWR network.

7.4 Experiments and Results

We used seven objects placed in a cluttered environment for our image database (see figure 7.4). A mobile robot equipped with a CCD camera was used to take images from 36 different viewpoints around the objects with intervals of 10°. The image database consists of four different sets, where the orientation of the objects is respectively 0°, 90°, 180°, and 270° with respect to the environment, resulting in a different background for the objects. In the experiments, training was done on one single set, while the other three sets were used to test the performance. This resulted in 12 different cross-validation tests.

7.4.1 Active Segmentation

Figure 7.4 shows a number of examples of interest points that are filtered using active vision. The white squares are the interest points that are both stable and move accord-

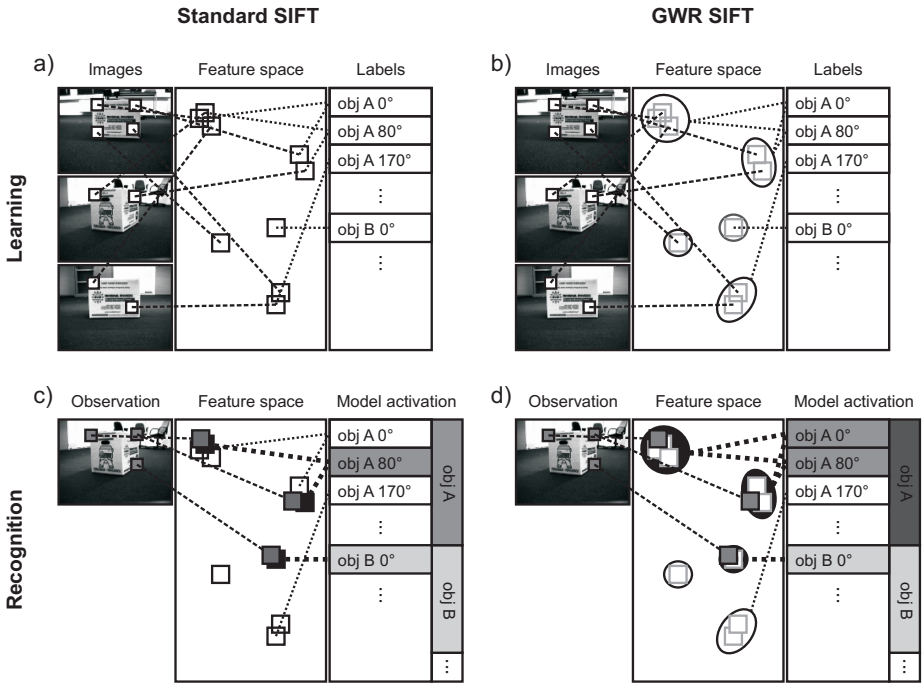

Figure 7.3: The interest-point database. a) During learning, the SIFT interest-point descriptors (squares) are stored along with the associated object and pose. b) When using the GWR network, the interest points are clustered (ellipses), and the cluster descriptor (related to the average of the clustered points) is stored with the associated object(s) and pose(s). c) During recognition, the observed interest points (gray-filled squares) are matched with the database in feature space. The nearest neighbors (black-filled squares) active the associated object-pose models. d) In case of GWR-clustering, the observations are matched with the clusters, and the nearest clusters in feature space (black-filled ellipses) support the associated object-pose models.

ing to the foreground. A point is stable when it is also observed in the previous or in the next observation. A point moves according to the foreground if it satisfies equation 7.1. The gray squares are stable interest points that move according to the background. The black squares are the interest points that are either unstable, or move according to the background. It can be appreciated from the figure that the majority of object in-

terest points lie indeed on the object and most background interest points are correctly filtered out. However, quite some interest points on the objects are also filtered out due to instability. It is a know drawback of SIFT that it produces many interest points that are not observable from slightly different viewing angles. The filtering of interest points will result in better object models. The object models will neither include background points, nor will they include unstable points that are not reobservable from small deviating points of view.

The numbers on the bottom left of the images in Figure 7.4 indicate the number of true positives, that is the number of interest points that are correctly classified as object versus the number of false positives, the number of incorrectly classified object points. The numbers on the bottom right give the precision, where the precision is calculated as: $precision = tp/(tp+fp)$.

7.4.2 *Active vs Passive Object Recognition*

In our next experiment, we tested the performance of our active approach to 3D object recognition and compared it with a passive approach that does not use active vision for robust interest points filtering and multiple viewpoints. The recognition performances are shown in Figure 7.5a. The plot shows the mean recognition rate over the 12 tests as a function of the number of viewpoints. The active methods accumulate evidence as given by Equation (7.5). The choice of the next viewpoint depends on the used next-viewpoint selection method, which is discussed in the next paragraph. Since the passive approach only uses a single viewpoint and does not accumulate evidence it is plotted as one wide bar. The error bars give the 95% confidence intervals on the mean. The active approach clearly outperforms the passive approach. Already with one viewpoint, the use of active vision to select stable object points gives significantly better performance than passively considering all visible interest points, with respectively 73% and 60% success (t-test: p-value $< 10^{-4}$). With increasing accumulation of evidence, the recognition rate rises from 73% to about 90% (t-test: p-value$\ll 10^{-4}$). Additionally, Figure 7.5b shows the decrease of the number of interest points when using the active method. This means that the active recognition system not only gives better recognition performance, it is also faster than a passive system.

We furthermore compared the performance of our next-viewpoint selection method to an approach where the next viewpoint is a simple 30° interval, as well as to random-

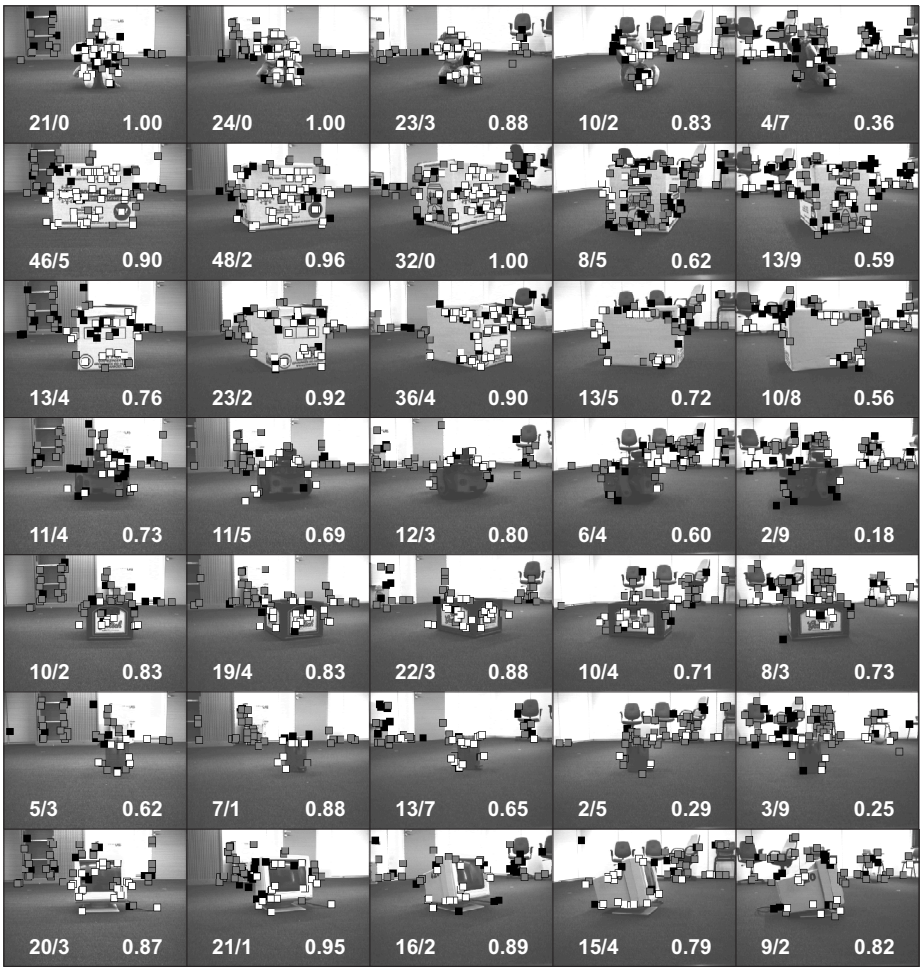

Figure 7.4: Examples of filtered interest points using active vision. The white squares are the interest points that are both stable (i.e., found in the previous or next observation) and move according to the foreground. The gray squares are stable points that move according to the background. The black squares are unstable interest points. The numbers on the bottom left give the number of correctly classified object points versus the number of incorrect classified object points. The number on the bottom right gives the precision.

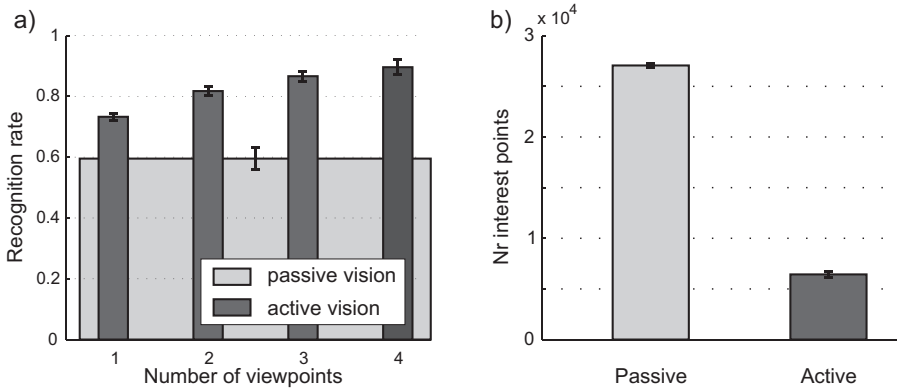

Figure 7.5: Passive versus active object recognition. In (a), the mean recognition rate is shown as a function of the number of accumulated viewpoints. Already for one viewpoint, the active-vision method has a significantly better performance than the passive method. The performance increases when more viewpoints are accumulated. The passive method dos not explore new viewpoints and is therefore plotted as one wide bar. In (b), the number of interest points are shown. It can be appreciated that the use of active vision greatly reduces the number of interest points. The error bars give the 95% confidence intervals on the means.

viewpoint selection. Figure 7.6 shows the mean recognition rates over the 12 tests. The error bars show the 95% confidence intervals. The difference of our next-viewpoint selection method with the interval method and the random selection method shows a small significant difference only for the second viewpoint (p-values of respectively 0.03 and 0.02 using a t-test). The difference in performance for more viewpoints is not significant, but is in favour of the ext-viewpoint-selection method. A significant increase in performance for the second viewpoint is important since this enables the system to recognize objects with fewer observations. Although, admittedly, the differences are small.

7.4.3 Interest-Point Clustering using the GWR Network

We first tested the influence of the thresholds a_T and h_T and the adaptation parameter η in the GWR network on the recognition performance and the reduction of interest points. The results are shown in Figure 7.7. The plots show a trade off between

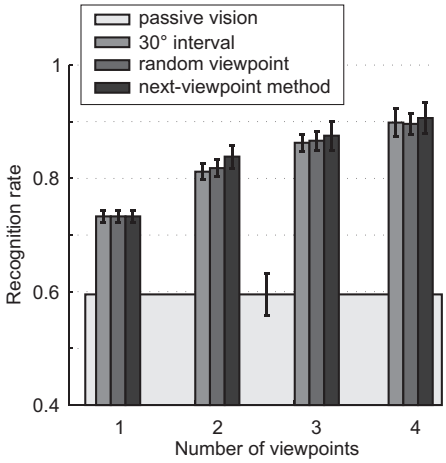

Figure 7.6: The recognition rates for the passive approach and three active methods: a fixed 30° interval, random viewpoint selection and our next viewpoint method. The error bars give the 95% confidence intervals in the means. Our viewpoint-selection method is significantly better when using two viewpoints.

the recognition performance and the number of interest-point clusters for a_T and h_T. With higher values for these parameters, the recognition rates increase, as well as the numbers of interest points. The best recognition results are obtained with $\eta = 0.15$. Lower values for η give worse recognition performance, while the number of resulting interest-point clusters is the same. Since the recognition for $\eta = 0.15$, $a_T = 0.9$ and $h_T = 0.55$ is similar to the performance of standard SIFT without interest-point clustering, we use those parameter settings to compar GWR SIFT with standard SIFT.

In the next experiment, we compared the performance of the GWR network with the standard SIFT method, both using active vision with a fixed interval of 30°. The learned interest points are presented to the GWR network in random order. We therefore performed ten different experimental runs to test the performance of the GWR network. Figure 7.8a shows the decrease of 50% in the number of interest-point clusters used by the GWR network. This results in a great improvement of the computational speed of the system. The recognition performance is shown in Figure 7.8b. We also compared the GWR network with standard SIFT that uses only 50% of the interest points, the same amount of interest points as used by the GWR network. These interest points

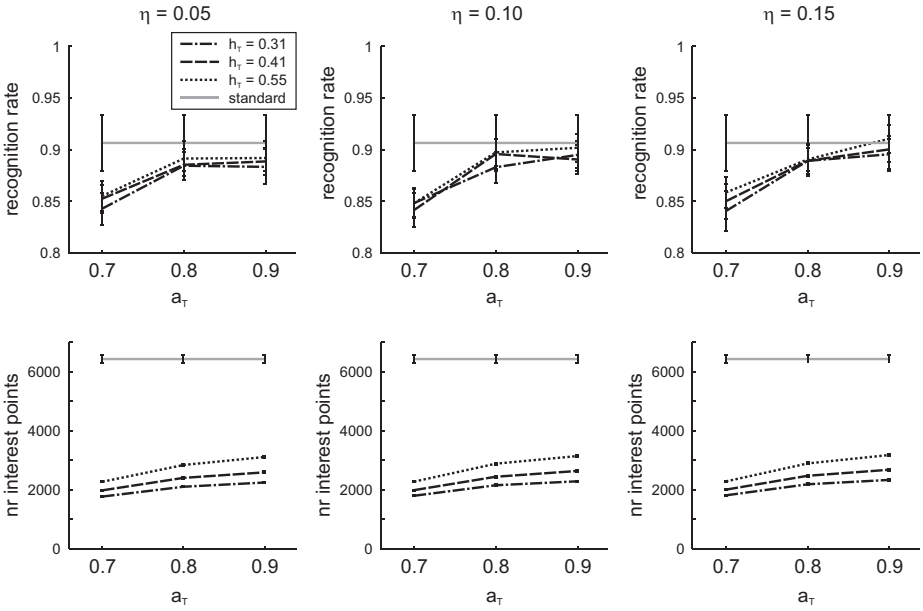

Figure 7.7: Recognition rate and number of interest-point clusters for different parameter settings of the GWR network. For larger a_T and h_T thresholds, the recognition performance improves, but also the number of interest-point clusters. A larger value for η results in better recognition without an increase of clusters. To keep up with the performance of the standard method not using clustering, we use $a_T = 0.9$, $h_T = 0.55$, and $\eta = 0.15$ in the subsequent experiment.

are selected randomly from the interest-point database. Again we performed 10 experimental runs. It can be appreciated from the plot that the GWR network performs similar to standard SIFT, but significantly better than standard SIFT using the same number of interest points. This shows that the GWR network effectively clusters the interest points, using only 50% of interest points without loss of performance.

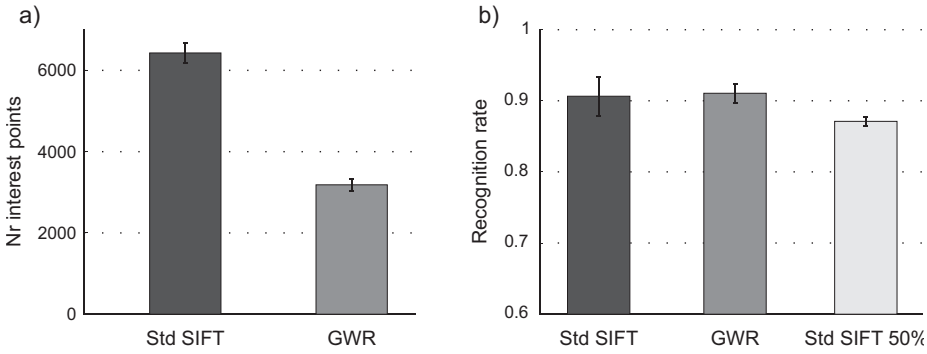

Figure 7.8: Object recognition with and without interest-point clustering using the GWR network. (a) Using GWR clustering reduces the number of interest points in the database by 50%, yielding a speed-up in computation time. (b) The mean recognition rates using the GWR network are compared with standard SIFT using all interest points and using the same amount of interest points as GWR (50%). For the GWR network and SIFT using 50% interest points, the data is acquired from 10 experimental runs. The error bars give the 95% confidence intervals on the mean. The recognition performance using the GWR network is similar to standard SIFT using all interest points, whereas it is significantly better than standard SIFT using 50% of the points.

7.5 Discussion

The experiments show the successful use of object exploration for 3D object recognition. Exploration is used in different ways, (1) to acquire evidence from multiple viewpoints, (2) to detect stable interest points, (3) to segment the object from the background, and (4) to select informative viewpoints. The active vision approach performs significantly better than passively observing. The problem of the passive method is that it not only learns to associate the object with interest points that actually lie on the object itself, but also with points in the background. Moreover, many unstable points pollute the database. These problems are solved by active exploration of the objects. By selecting the next viewpoint based on the optimization of the expected activation of object models, the system further increases its recognition performance with subsequent viewpoints.

The proposed next-viewpoint selection method results in a small but significant im-

provement of performance for the second viewpoint. However, the improvement is not significant when more viewpoints are used. There is room for improvement of the method. One of the problems is the calculation of the expected gain in activation for taking a next viewpoint. Since the proposed method is a one-shot-learning method, this expectation cannot be realistically estimated. We therefore propose to learn these probabilities during subsequent encounters with the object, to get realistic figures.

The use of a GWR network for the clustering of interest points has also been analyzed. The GWR network results in a strong reduction in the number of interest points that need to be stored in the database, while maintaining the recognition performance of standard SIFT. The GWR network performs significantly better than SIFT using the same number of interest points. Reducing the amount of interest points is important for object recognition using SIFT, especially with a growing number of objects. The results show that the GWR network is capable of effective clustering of interest points.

The use of the GWR network is very similar to the *bag-of-features* approach (Csurka et al., 2004). However, in the bag-of-features approach, the clusters are learned beforehand. Since the number of clusters and the position of the clusters cannot change during execution, the approach cannot deal with object types that are never seen before. Our approach constantly adapts to the input, and will create new clusters if necessary in new circumstances. In future work, we would like to investigate these properties of the GWR network and compare our system to the bag-of-feature approach.

In the presented study, no additional methods are used to improve the recognition rate. A good way to boost recognition is to use a geometric fit between sets of interest points, for instance the geometric verification method described in (Lowe, 2004). This method can be used both with the proposed active-vision method and with the GWR network.

Summarizing, this chapter showed the successful use of active vision to simplify complex recognition tasks. The quality of the object representations is improved by exporing the objects and better recognition is obtained by efficiently taking different viewpoints. The GWR network furthermore demonstrated the possibility to reduce the number of interest points. These methods make the implementation of object recognition in the real world more feasible.

Paying Attention to Symmetrical Interest Points

Abstract

Most visual Simultaneous Localization And Mapping (SLAM) methods use interest points as landmarks in their maps of the environment. Often the interest points are detected using contrast features, for instance those of the Scale-Invariant Feature Transform (SIFT). The SIFT interest points, however, have problems with stability and noise robustness. Taking inspiration from human vision, the use of local symmetry to select interest points is proposed. Symmetry is a stimulus that occurs frequently in everyday environments where our robots operate in, making it useful for SLAM. Furthermore, symmetrical forms are inherently redundant, and can therefore be more robustly detected. The proposed method, the MUlti-scale Symmetry Transform (MUST), has been tested on stability, robustness, and repeatability. Moreover, the method is used to select landmarks to represent the environment. To test the SLAM performance of our model, we recorded a SLAM database with a mobile robot, and annotated the database by manually adding ground-truth positions. The results show that interest points selected using symmetry are more stable and more robust to noise and contrast manipulations, have a slightly better repeatability, and above all, result in better overall SLAM performance.

This chapter is based on:

Kootstra, G., de Jong, S., & Schomaker, L. R. B. (2009). Using local symmetry to select landmarks for visual SLAM. In *The 7th International Conference on Computer Vision Systems*. Liège, Belgium.

8.1 Introduction

One of the fundamental tasks of an autonomous robot is to build a map of the environment and use it for self-localization. The problem of Simultaneous Localization and Mapping (SLAM) has therefore received much attention in the last decade (Thrun et al., 2005). Nowadays, approaches using laser range finders are very successful. SLAM using vision, however, remains a challenging research topic, (e.g., Frintrop & Jensfelt, 2008; Davison et al., 2007).

Using a camera has the advantage over a laser-range finder that it is a passive sensor that is low cost, low power, and lightweight. A camera furthermore provides a rich source of information, which enables the use of sophisticated detection and recognition methods. The difficulty, however, is to extract relevant information from the high-dimensional visual data in real time. In this chapter, the use of local symmetry to select relevant visual information is proposed.

Most visual SLAM systems use visual landmarks to create a map of the robot's environment. It is important to select robust, stable landmarks that will be recognizable in future encounters. Furthermore, the number of selected landmarks should be limited, since the computational complexity of the SLAM algorithms strongly depends on the number of landmarks.

A common approach for the selection of landmarks in current visual-SLAM systems is to detect interest point in the camera images (Mozos et al., 2008). Most approaches use contrast features to detect interest points. Examples are the Scale-Invariant Feature Transform (SIFT) (Lowe, 2004; Se et al., 2002), Speeded-Up Robust Features (SURF) (Bay et al., 2006; Murillo et al., 2007), and Harris corners (Davison & Murray, 2002). In this chapter, the use of local symmetry to detect interest points is suggested. The proposed MUlti-scale Symmetry Transform (MUST) is compared with SIFT, since it is among the best performing interest point detectors in SLAM (Mozos et al., 2008), as well as in object recognition (Moreels & Perona, 2007).

Although systems using SIFT have successfully been applied to SLAM, there are three important drawbacks. The interest points are very susceptible to noise, not all selected landmarks are recognized when the robot returns to a previously visited location, and too many interest points are found in an image, but only few points are so stable that they can be tracked over a number of successive frames. Thus SIFT has problems with robustness, repeatability, and stability.

In this chapter, a landmark-selection mechanism is proposed that uses local symmetries instead of local contrast. As argued in Section 6.4, symmetry methods have fewer problems with noisy conditions than contrast methods. The use of local symmetry is therefore hypothesized to perform better regarding the problems mentioned above. Our choice for symmetry is motivated by human behavior. As has been discussed in Chapter 2, symmetry is detected very rapidly, especially when patterns have multiple axes of symmetry (Palmer & Hemenway, 1978). Moreover, humans pay attention to locally symmetric parts of images (see Chapter 3). Furthermore, humans use symmetry to segregate a figure from its background (Driver et al., 1992). These findings suggest that symmetry is used in preattentive vision, and can therefore be used for context-free object segmentation and landmark selection. Assuming that the human visual system has evolved to be as effective as possible in the kinds of environments that humans have to operate in, this suggests that symmetry detection might be useful in robots that have to operate in similar environments.

Using symmetry to select landmarks exploits the fact that most man-made indoor environments contain many symmetrical objects and forms. Since symmetry is a strong non-accidental property, we believe that its use will result in the selection of valuable landmarks for the visual SLAM system.

To detect local symmetry in an image, a number of symmetry operators exist. Reisfeld et al. (1995), for instance, developed a mirror-symmetry operator by comparing the gradients of neighboring pixels. Heidemann (2004) extended this work to the color domain. Reisfeld et al. also proposed a radial-symmetry operator that promotes patterns that are symmetric in multiple symmetry axes. A faster operator to detect radial symmetry, the Fast Radial Symmetry Transform (FRST), is proposed in (Loy & Zelinsky, 2003).

In this chapter, a novel scale- and rotation-invariant interest-point detector is proposed, which is called the MUlti-scale Symmetry Transform (MUST). The detector extends the FRST symmetry operator to a model that detects interest points on multiple scales. Techniques from SIFT are used to obtain a scale- and rotation-invariant representation of the interest points. The use of MUST for selecting landmarks for visual SLAM is discussed. The results show that landmarks selected using local symmetry are more robust to noise and have a higher repeatability, and require fewer computations. Most importantly, the overall performance of the SLAM system increases when interest points are selected using local symmetry instead of SIFT.

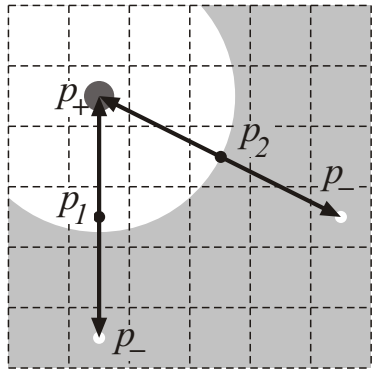

Figure 8.1: The Fast Radial Symmetry Transform (Loy & Zelinsky, 2003). Each pixel in the image votes for the existence of symmetry at a given radius r. In this example, two pixels p_1 and p_2 are shown. Using the orientation of a pixel's gradient, a vote is made for bright symmetrical forms on a dark background at p_-, and for dark symmetrical forms on a bright background at p_+. Doing this for all pixels in the image gives the symmetry transform of the image for the given radius.

8.2 Methods

The complete system consists of three parts. The first part is the selection of interest points based on local symmetry. The detected interest points are then fed to a visual buffer to select stable interest points as landmarks. The selected landmarks are finally used by the EKF-SLAM system to build a map of the environment and to estimate the position of the robot in the environment.

8.2.1 Interest Points Based on Local Symmetry

As a basis of our interest point detector, we use the Fast Radial Symmetry Transform (Loy & Zelinsky, 2003), and extended it to a multi-scale and rotation-invariant detector, MUST. The basis of the FRST is given in Figure 8.1, and MUST is depicted in Figure 8.2.

To obtain a multi-scale interest-point detector, we use a pyramid approach similar to that used in (Lowe, 2004). The symmetry response is calculated at five spatial octaves,

$\mathcal{O} = \{-1,0,1,2,3\}$. In the first octave, -1, the gray-scaled input image, I_{-1}, has twice the resolution of the original image, similar to (Lowe, 2004). For each next octave, the input image of the previous octave is smoothed with a Gaussian kernel and down-sampled by a factor of two. Within each octave, there are three scales, $s \in \{0,1,2\}$, with progressive smoothing. This gives a pyramid of gray-scaled images, $I_{o,s}$, for a given octave o and scale s.

The symmetry transform, $\Psi(o,s)$, for octave o and scale s, is calculated by investigating the local radial symmetry at a set of different radii. The size of the radii depends on the scale s. The set of radii used to calculate the symmetry response is defined as $\mathcal{R}_s = (1+s/(s_{max}-1)) \cdot \mathcal{R}$, where $s_{max} = 3$ and $\mathcal{R} = \{1,3,5\}$. The symmetry transform is then:

$$\Psi(o,s) = \frac{1}{|\mathcal{R}_s|} \sum_{r \in \mathcal{R}_s} \psi_r(o,s) \tag{8.1}$$

where $\psi_r(o,s)$ is the symmetry response at octave o and scale s for radius r.

$\psi_r(o,s)$ is determined by first calculating the gradients of the input image $I_{o,s}$ with horizontally and vertically aligned Sobel filters, resulting in a magnitude map, $G_{o,s}$, and an orientation map, $\Theta_{o,s}$. Then, each pixel p votes for the existence of symmetry based on its gradient, $\Theta_{o,s}(p)$, its magnitude, $G_{o,s}(p)$ and the given radius r. This is related to the Hough transform. A pixel votes for the existence of both a bright symmetrical form on a dark background, and a dark symmetrical form on a bright background at respectively location p_+ and p_-:

$$p_+ = p + [r \cdot (\cos\Theta_{o,s}(p), \sin\Theta_{o,s}(p))] \tag{8.2}$$

$$p_- = p - [r \cdot (\cos\Theta_{o,s}(p), \sin\Theta_{o,s}(p))] \tag{8.3}$$

where $[\ldots]$ is the nearest-integer function. Subsequently, an orientation projection map, O_r, is calculated that counts the number of symmetry votes, as well as a magnitude projection map, M_r, which keeps track of the magnitudes of the gradients that contribute to these votes. Initially, all values in these maps are set to zero. Then, for every pixel p, the maps are updated according to:

$$O_r(p_+) = O_r(p_+) + 1 \quad , \quad O_r(p_-) = O_r(p_-) - 1 \tag{8.4}$$

$$M_r(p_+) = M_r(p_+) + G_{o,s}(p) \quad , \quad M_r(p_-) = M_r(p_-) - G_{o,s}(p) \tag{8.5}$$

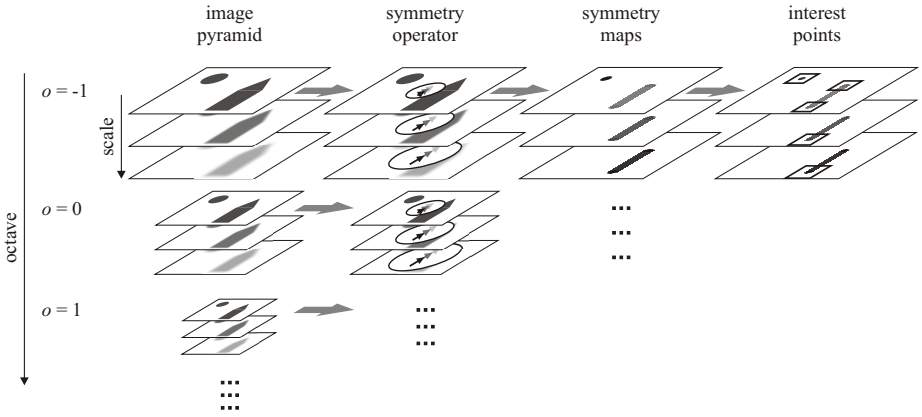

Figure 8.2: MUST, the MUlti-scale Symmetry Transform. Symmetry is calculated at multiple octaves and scales. At the first octave, $o = -1$, the image is double its original size. For every next octave, the image is down-scaled by a factor two. A total of five octaves is used. Within an octave there are three scales with progressive Gaussian blurring. For every image in the image pyramid, the symmetry response, $\psi_r(o,s)$, is calculated at three different radii by applying the symmetry operator. The average of these responses results in the symmetry map, $\Psi(o,s)$, for the given octave o and scale s. In the last step, the interest points are obtained by finding points that have a local maximal or minimal symmetry value. The local neighborhood of such an interest point is described by the SIFT descriptor. The size of the neighborhood depends on the scale.

Finally, the symmetry map for radius r at octave o and scale s is determined by:

$$\psi_r(o,s) = F_r * A_r \tag{8.6}$$

and

$$F_r(p) = M_r(p) \cdot O_r(p)/k_r \tag{8.7}$$

where $O_r(p)$ has a upper value of k_r, and a lower value of $-k_r$. k_r has been experimentally established in (Loy & Zelinsky, 2003) at $k_r = 8$ for $r = 1$, and $k_r = 9.9$ otherwise. A_r is a Gaussian kernel of size $r \times r$ with a standard deviation of $0.25r$.

Equation (8.7) weighs the votes in the orientation projection map with the values in

the magnitude projection map. This results in stronger symmetry votes for stronger gradients. The convolution with the Gaussian kernel in Equation (8.6) spreads the symmetry votes over neighboring pixels, with a larger spread for larger radii. This allows symmetrical patterns to deviations from perfect radial symmetry.

The above gives us the symmetry response $\psi_r(o,s)$ for the radius r at octave o and scale s. By averaging the symmetry responses over all radii in $\mathcal{R}_s$, according to equation (8.1), we obtain the full symmetry response at octave o and scale s, $\Psi(o,s)$.

Next, the interest points in every octave and scale are determined by finding the points in $\Psi(o,s)$ that are either a maximum or a minimum in the spatial neighborhood of 11×11 pixels. Different from SIFT, the points do not need to have an optimal value over neighboring scales, since this resulted in the rejection of too many valuable interest points. A pixel can therefore potentially hold multiple interest points at different scales, which is fine because each corresponds to a different symmetrical pattern.

For each interest point i found in the symmetry maps the following information is stored:

- the location of the interest point in the original resolution of the input image, $\mathbf{x}_i = (x_i, y_i)$,
- the scale value, $\sigma_i = 2^{o_i + s_i / s_{max}}$,
- the symmetry strength, $v_i = \Psi(o_i, s_i, \mathbf{x}_i)$,
- the orientation, γ_i, and
- the descriptor of the interest point, $\mathbf{d}_i$.

The later two are described in the next paragraph.

To calculate the orientation and the descriptor of the interest points, we use the corresponding methods of the SIFT algorithm (Lowe, 2004). The orientation of the interest point is determined by finding the dominant gradient in the local neighborhood of the interest point. This orientation is used to obtain a rotationally invariant descriptor of the local neighborhood patch. The neighborhood is described by histograms of gradients. The size of the neighborhood depends on the scale value of the interest point. This makes that the descriptor is also scale invariant. To calculate the histograms of gradients. the patch is divided in 4 by 4 squares. A histogram of gradients is then calculated for each square. Since there are 16 squares, and each histogram contains 8 bins, this gives us a feature vector with 128 values. The magnitude of the feature vector

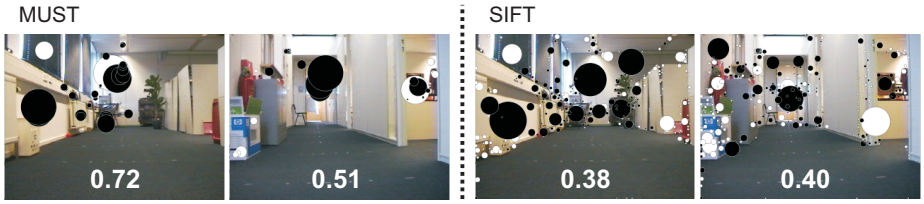

Figure 8.3: Two examples images with the MUST and SIFT interest points. The black circles present the stable interest points that are found in both the current and the previous image. The white circles depict the unstable interest points that are only found in the current image. The proportion of stable points is given at the bottom.

is normalized to 512. For more detailed information about the method to calculate the descriptor, we refer to (Lowe, 2004).

Figure 8.3 shows the interest points found by MUST on two of the images used in our experiments as compared to the SIFT interest points. Note that SIFT results in a large number of interest points, of which many are unstable. MUST, on the other hand, results in less, but more stable points.

To summarize, MUST calculates symmetry maps on multiple octaves and scales of the image using the symmetry transform. In these symmetry maps, we find points that have a locally optimal (i.e., maximal or minimal) symmetry value. These points are the interest points our method returns. We then use the SIFT descriptor to describe the interest points. We thus replace the SIFT method to find interest points based on difference of Gaussians by MUST, that finds interest points based on local symmetry.

8.2.2 The Visual Buffer

Both MUST and SIFT result in a large number of interest points when applied to the camera images taken by the robot. In our experiments, MUST detects on average 40 interest points per image, and SIFT 124. Using all these points would result in far too many landmarks in the complete map. To be practically usable, the state matrix of the Extended Kalman Filter should maximally contain a few hundred landmarks in total. Furthermore, most interest points found in one observation are not detected in subsequent observations. In other words, many interest points are unstable or only detectable

from a particular viewpoint, and are therefore useless for SLAM. We propose to use a visual buffer to retain only stable interest points, similar to (Frintrop & Jensfelt, 2008; Se et al., 2002).

The buffer contains the last N camera images. The interest points in the current image are compared to those in the $N-1$ previous images. An interest point i in the current observation is selected as landmark if it satisfies the following criteria:

1. the interest point is matched in K of the previous images in the buffer, and $K \geq M$, where M is the minimally necessary number of matching images,
2. the estimates of the position of the landmark in the environment are congruent, and
3. to avoid spurious interest points, the strength of the interest point, v_i, is at least a proportion of the strength of the maximum interest point that is successfully matched in the current image:

$$v_i \geq \lambda \cdot v_{max} \tag{8.8}$$

For criterion 1, the interest point with descriptor $\mathbf{d}_i$ is matched when there is an interest point in the previous image with a descriptor $\mathbf{d}_j$ that is sufficiently similar. This is true when the Euclidean distance is below the threshold τ_1:

$$||\mathbf{d}_i - \mathbf{d}_j|| < \tau_1 \tag{8.9}$$

Additionally, to ensure unique interest points, the *best-to-next-best ratio* should be smaller then the threshold δ_1:

$$||\mathbf{d}_i - \mathbf{d}_j||/||\mathbf{d}_i - \mathbf{d}_l|| < \delta_1 \tag{8.10}$$

where $\mathbf{d}_l$ is the descriptor of the second most similar interest point in the previous image. This prevents ambiguous landmarks in the database.

For criterion 2, the landmark's position is estimated by triangulation using the bearings of the interest point in the current image and the previous images, and the displacement of the robot. This results in a set of estimates of the range and bearing,

$$\mathrm{P} = \{\mathbf{p}_k | \mathbf{p}_k = \langle r_k, \theta_k \rangle \wedge 1 \leq k \leq K\} \tag{8.11}$$

The landmark is accepted if

$$\text{var}(R) < \rho \tag{8.12}$$

where var is the variance, and $R = r_k$.

The visual buffer tests the quality of the interest points, and adds only strong and stable points that are observable from multiple viewpoints to the map. An additional benefit is that the covariance matrix of the observation error, used in the EKF, can be initialized based upon the covariance matrix of the estimated landmark positions, cov(P).

8.2.3 Visual SLAM

We used a standard implementation of the Extended Kalman Filter (EKF) as basis of the SLAM system (Durrant-Whyte & Bailey, 2006), as discussed in Section 6.5.4. A full description of EKF-SLAM is given in Appendix C. This section focuses on the incorporation of the landmark observations in EKF-SLAM.

A landmark i that results from the buffer is classified as either a new landmark, or a previously observed landmark that is already in the map. It concerns a previously observed landmark if the landmark in the database with the most similar descriptor, j, fulfills three criteria:

1. the distance between the descriptors of landmark i and landmark j is smaller then a given threshold τ_2:
$$||\mathbf{d}_i - \mathbf{d}_j|| < \tau_2 \tag{8.13}$$
2. the best-to-next-best ratio is smaller then a given threshold δ_2:
$$||\mathbf{d}_i - \mathbf{d}_j||/||\mathbf{d}_i - \mathbf{d}_l|| < \delta_2 \tag{8.14}$$
where $\mathbf{d}_i$ is the descriptor of the currently observed landmark, $\mathbf{d}_i$ is the most similar descriptor, and $\mathbf{d}_l$ is the second most similar descriptor in the database.
3. the position of the current observation should be within the vicinity of the landmark in the database:
$$\sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{S}_j^{-1} (\mathbf{x}_i - \mathbf{x}_j)} < \eta \tag{8.15}$$

where $\mathbf{S}_j$ is the uncertainty covariance matrix, which is discussed in the next paragraph. The landmark is classified as new only if none of the three criteria is fulfilled. This ensures uniqueness of a new landmark and thus avoids ambiguous landmarks in the map.

The landmark is classified as new only if none of the three criteria is fulfilled. For a new landmark, the state matrix and the state covariance matrix are augmented using the observation, $\mathbf{z}_i$, and the uncertainty covariance matrix, $\mathbf{S}_i$, where $\mathbf{z}_i$ is set to mean(P). $\mathbf{S}_i$. $\mathbf{z}_i$ is determined using cov(P) and the uncertainty of the robot's position in the EKF. Additionally, the descriptor of the interest point is stored. A landmark that is matched with an existing landmark in the database is used in the update step of the EKF.

Because EKF is quadratic in the number of landmarks, we restrict the size of the map to a maximum of 350 landmarks, so the system can run in real time. More efficient implementations exist that can handle larger numbers of landmarks (Guivant & Nebot, 2001). This, however, lies outside of the scope of this research. Important is that the benefits of using symmetry to detect landmarks also hold for other implementations of the EKF, as well as for Particle Filters (e.g., Sim et al., 2007; Montemerlo et al., 2003) and the Information Filter (Thrun et al., 2004).

8.3 Experiments

We performed a number of experiments with a Pioneer II DX robot equipped with a Sony D31 camera to test the use of local symmetry for visual SLAM and to compare it to using standard SIFT. To be able to repeat the experiments, we created a SLAM database. This database contains camera images and odometry information, along with the ground truth position of the robot. The data was recorded in ten different runs, in which the robot drove four laps of approximately 35 meters through an office environment and hallway with an average speed of 0.3 m/s. Camera images of 320 x 240 pixels were stored at 5 Hz. At intervals of one meter, the true location of the robot was logged by hand. This enabled us to quantify the performance of the SLAM estimation.

SIFT and MUST result in different types of interest points, which might have an influence on the best settings for the other parameters in the SLAM system. We therefore optimized the parameters for the visual buffer (N, M, τ_1, δ_1, and ρ) and for the SLAM

system (τ_2 and δ_2) for both MUST and SIFT separately. Any differences in performance can therefore be subscribed to the interest-point detectors.

8.3.1 Stability

The stability of the interest points has been tested on 200 images in multiple sequences recorded by the robot. For every image, the stability of all interest points is tested. An interest point is considered stable if it is found in all previous N consecutive images. If the interest point is not found in all of the previous images, it is considered unstable. It must be noted that the visual buffer is not used in this experiment but all interest points detected by the MUST and SIFT are used.

8.3.2 Robustness

Cameras that robots use to perceive the world are usually noisy causing different kinds of perturbations of the camera images. Furthermore, most environments that a robot needs to map are subject to changing light conditions. It is therefore important that a landmark selection mechanism is robust to these confounding factors.

In this experiment, 57 camera images were used from one of the runs in the database. The images were taken at intervals of approximately 3 meters. To test the robustness to noise, we added Gaussian pixel noise, and Gaussian smoothing to the original images. In addition, we manipulated the contrast and brightness to test the robustness to changing light conditions. The visual buffer is not used and all interest points detected by MUST and SIFT are used. The functions used for the manipulations are:

1. *Gaussian pixel noise*: the original image is transformed to:

$$I'(x,y) = I(x,y) + X(\alpha_g) \tag{8.16}$$

   where $I(x,y) \in [0,1]$ is the intensity of pixel (x,y), and $X(\alpha_g)$ is a random sample from the normal distribution $\mathrm{N}(0,\alpha_g^2)$.
2. b) *Gaussian smoothing*: the original image is convolved with a Gaussian mask:

$$I' = I * G_s \tag{8.17}$$

where G_s is a Gaussian mask of size $s \times s$, with a standard deviation of $\sigma = s/6$.

3. *Contrast manipulation*:

$$I'(x,y) = I(x,y) + \alpha_c \left(I(x,y) - \bar{I}_{x,y}\right) \tag{8.18}$$

where $\bar{I}_{x,y}$ is the local average in a neighborhood of 21×21 pixels around pixel (x,y). The contrast increases with positive values for α_c, and decreases for negative values.

4. *Brightness manipulation*:

$$I'(x,y) = I(x,y)^{\log \alpha_b / log 0.5} \tag{8.19}$$

For $\alpha > 0.5$, the pixels are brightened, for $\alpha < 0.5$, the pixels are darkened.

The reason to not use the dataset in (Mikolajczyk, Tuytelaars, Schmid, Zisserman, Matas, Schaffalitzky, Kadir, & Van Gool, 2005) is that it contains images of outdoor scenes, which contain fewer symmetrical objects than indoor environments.

The robustness is measured by the proportion of matching interest points between the original and the manipulated images. Two interest points match when criterion 1 of the visual buffer is met (see section 8.2.2), where $\tau_1 = 0.6$ and $\delta_1 = 0.75$. Additionally, the spatial distance between the two interest points in the image should be fewer than 3 pixels.

8.3.3 Repeatability

For good SLAM performance, it is crucial that landmarks added to the map are observable on future encounters. We therefore test the repeatability of the interest points. For every experimental run, we selected three parts of the robot's trajectory, and aligned the sequences of images of the four consecutive laps that are approximately at the same position. The interest points in the images in the first lap are then compared to the images taken at approximately the same position in the later laps.

The average proportion of interest points that are matched in lap 2, 3, and 4 is the measure of repeatability. An interest point matches when criterion 1 of the visual buffer is met (see section 8.2.2), where $\tau_1 = 0.6$ and $\delta_1 = 0.75$. Unlike the evaluation

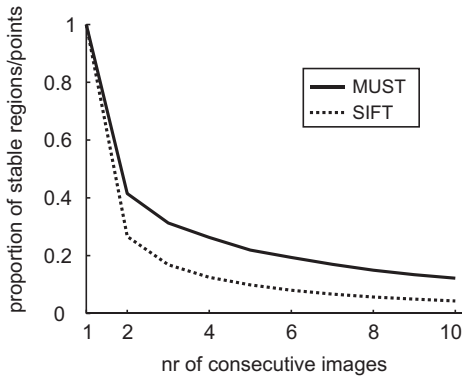

Figure 8.4: Stability of the interest points as a function of the number of consecutive images. The graphs show the proportions of interest points that are stably found in all consecutive images.

of the robustness, we do not put a constraint on the spatial distance between the two observed interest points, since there is variation in the exact position and orientation of the robot, which may cause relatively large displacements of the interest points.

8.3.4 Visual SLAM Performance

Because we logged the ground-truth position of the robot at intervals of 1 meter, we can quantify the performance of the complete visual SLAM system. For that, we compare the estimated position of the robot made by the EKF with the ground-truth position. The SLAM performance is the average Euclidean distance between the estimated and ground-truth positions in the last of the four laps.

8.4 Results

8.4.1 Stability

Figure 8.4 shows the proportion of stable interest points as a function of the number of consecutive images. It can be seen that the symmetrical interest points selected by

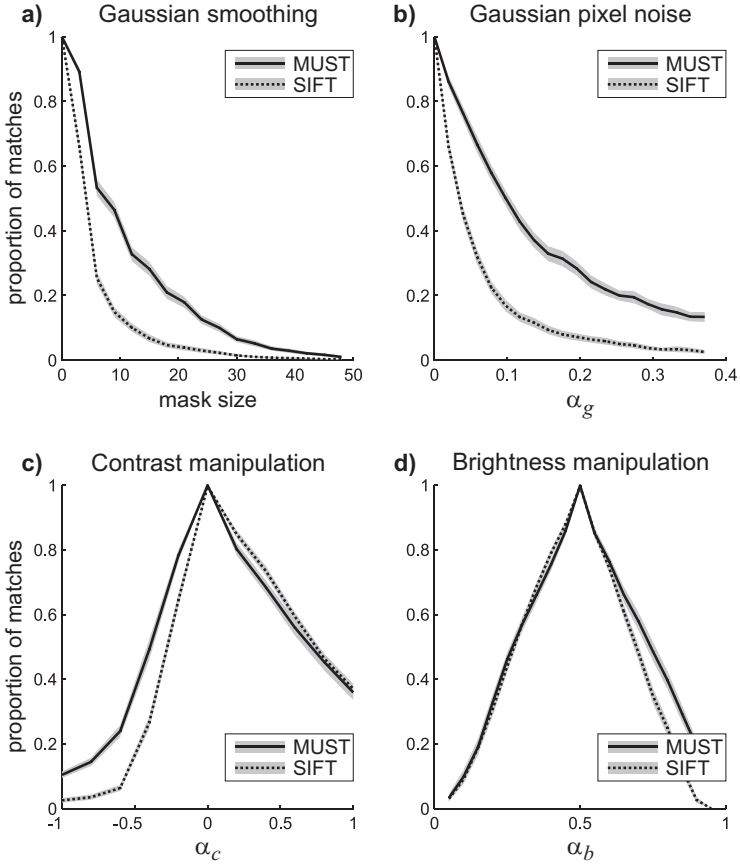

Figure 8.5: Robustness to noise and changing light condition. The lines show the mean proportion of matching interest points between the distorted and the original images over 57 runs. The gray areas around the lines display the 98% confidence intervals. Note that these intervals are small.

MUST are more stable than the points selected by SIFT. This shows that the MUST interest points are more reliably detected over small changes of viewpoint than the SIFT interest points.

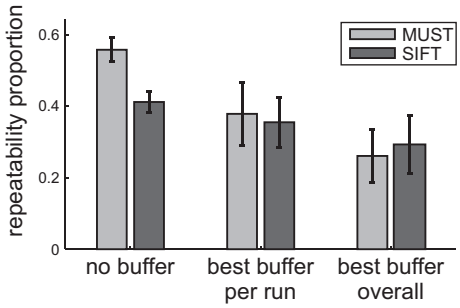

Figure 8.6: The repeatability performance. The plot shows the mean repeatability proportion over the ten runs. The error bars give the 95% confidence intervals.

8.4.2 Robustness

Figure 8.5 shows the robustness results. The proportion of matching interest points between the original and the manipulated images are displayed on the y-axis. In Figure 8.5a and 8.5b, the results for the addition of noise are given. MUST is significantly less affected by the noise than SIFT. The performance of MUST is more than twice that of SIFT, making it much more robust to noise. Figure 8.5c shows the performance with respect to contrast manipulation. Although the performance of both methods is similar for the contrast enhancement, MUST shows a significantly better performance when the contrast is reduced. Also with enhanced brightness the use of MUST results in considerably better performance (see Figure 8.5d). There is no difference with reduced brightness.

8.4.3 Repeatability

The repeatability results can be appreciated in Figure 8.6. The first two bars show the repeatability results for all detected interest points, in other words when the visual buffer is not used. The repeatability of the MUST interest points is significantly higher then that of the SIFT points. The two other groups of bars display the results when the buffer is used, where the buffer parameters are optimized both per run and overall with an additional contraint that the buffer does not result in too few (< 50) or too many (> 350) landmarks per run. MUST has a slightly better result per run, whereas overall,

SIFT has a somewhat better repeatability. The differences, however, are not significant.

The overall best settings are:

	N	M	τ_1	δ_1	v
MUST	12	4	0.3	0.8	0.6
SIFT	14	8	0.3	0.5	0.0

8.4.4 Visual SLAM Performance

The SLAM performance is displayed in Figure 8.7. The results for the best settings per run show that MUST results in a significantly lower estimation error ($p = 0.05$, paired t-test) with a medium effect size, as measured by Cohen's d ($d = .70$). Also when the overall best settings are used, MUST performs better than SIFT with a medium effect size ($d = .61$). The difference, however, is not significant ($p = .11$, paired t-test), which is due to the large variance over the ten runs and the relatively small number of runs.

The overall best settings are:

	N	M	τ_1	δ_1	v	τ_2	δ_2	η
MUST	14	8	0.3	0.8	0.8	0.3	0.7	0.5
SIFT	14	8	0.3	0.5	0.8	0.6	0.7	0.5

Not only does MUST result in better performance, it also reduces the amount of selected landmarks. MUST selects on average 40 interest points per image, and SIFT 124. These interest points are all fed to the buffer. Since the matching process is quadratic in the number of interest points, MUST greatly reduces the number of computations. Since both models are similarly fast in detecting interest points in the images, the full SLAM system is faster using MUST.

8.5 Discussion

In this chapter, a new interest point detector, MUST, is proposed for landmark selection based on local symmetry. The model exploits the presence of symmetrical forms in indoor environments. The interest points selected by MUST are more stable over small changes in perspective and are less sensitive to noise and changing light conditions

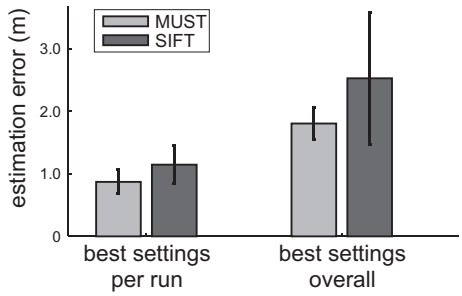

Figure 8.7: The SLAM performance. The plot shows the mean estimation error over the ten runs. The 95% confidence intervals are represented by the error bars.

than points selected on the basis of contrast by the SIFT model (Lowe, 2004). The repeatability of all selected interest points is higher for the proposed model than for SIFT if the visual buffer is not used. Most importantly, the use of symmetry results in better visual SLAM performance. This shows that exploiting symmetry results in valuable and robust landmarks for the representation of the environment. Moreover, MUST results in less interest points, which improves the processing time of the SLAM system.

There are three likely reasons for the success of using symmetry for interest-point detection. Firstly, symmetrical forms are inherently redundant. Both sides of symmetrical patterns represent the same structure. This makes a symmetrical pattern less susceptible to noise. Secondly, symmetry is a non-accidental stimulus. If there is a symmetrical pattern, it is likely that this pattern is not there by chance, but originates from something that is physical present in the environment and is thus stable and useful for SLAM. Finally, man-made environments, especially indoor environments as used in the experiments, contain many symmetrical forms, which is exploited.

The fact that repeatability drops when the visual buffer is used shows that the visual buffer does not function optimally and rejects a portion of relevant interest points. If the effectiveness of the buffer algorithm can be enhanced, the SLAM performance could be further improved. A visual buffer to select a small number of stable interest points from the high number of interest points offered by the detection methods is necessary to reduce the number of landmarks in the map.

Symmetry detection can results in the selection of complete objects. This could result

in semantically more meaningful landmarks and maps of the environment, which could be exploited, for instance, in human-robot interaction.

It can conclude that local symmetry provides robust and valuable landmarks, which result in improved visual SLAM performance.

9

Paying Attention to Symmetrical Regions of Interest

Abstract

In the previous chapter, the MUlti-scale Symmetry Transform (MUST) is proposed for the detection of interest points based on symmetry. The performance of the method is compared to the SIFT interest-point detector. The results showed that symmetrical points are more stable over small changes of viewpoint and more robust to noise and changing light conditions. Moreover, the proposed method resulted in better SLAM performance when used to select visual landmarks to represent the environment. However, Figure 8.4 reveals that even for MUST the stability of the interest points drops quickly as a function of the displacement of the robot. Regions of interest are expected to be more stably detected than points, because they are supported by larger areas and thus less susceptible to noise. Therefore, a Symmetrical Region-of-Interest Detector (SymRoID) is proposed in this chapter to improve the stability and with that the SLAM performance. The results show that symmetrical regions-of-interest are less susceptible to noise, are more stable, and above all, result in better SLAM performance.

This chapter is based on:

Kootstra, G., & Schomaker, L. R. B. (2009b). Using symmetrical regions-of-interest to improve visual {SLAM}. In *the International Conference on Intelligent RObots and Systems (IROS)*. St. Louis, USA.

9.1 Introduction

One of the challenges in visual SLAM is to find high-quality visual landmarks to represent the environment (Frintrop & Jensfelt, 2008). A good and reliable landmark is one that is detectable despite noise and changing light conditions, that is stable over a sequence of observations, and that is detectable when the robot revisits the location. In this chapter, a Symmetrical Region-of-Interest Detector (SymRoID) is proposed to improve the detection of landmarks. This research extends the work presented in the previous chapter. Like in the previous chapter, the method exploits the inherent redundancy of symmetrical patterns and the presence of many symmetrical patterns in man-made environments. However, instead of detecting interest points, SymRoID detects symmetrical regions-of-interest.

Many current approaches to visual SLAM detect interest points based on contrast features, for instance using the Scale-Invariant Feature Transform (SIFT) (Lowe, 2004; Se et al., 2002), Speeded-Up Robust Features (SURF) (Bay et al., 2006; Murillo et al., 2007) or Harris corners (Davison & Murray, 2002). SIFT has been proven to be one of the best performing interest-point detectors for SLAM (Mozos et al., 2008), as well as for object recognition (Moreels & Perona, 2007). In Chapter 8, it was shown that using local symmetry instead of contrast results in more robust interest points. Moreover, it was shown that the performance of the SLAM system using local symmetry for landmarks selection outperformed the system using SIFT interest points.

However, the problem with interest points is that a large number of points are found in the image, but many of them are unstable. This can be appreciated in Figure 8.4. The high number of points results in a high computational load. Moreover, the poor stability of the points reduces the quality of the map of the environment. Although the use of symmetry reduces the number of interest points and improves the stability compared to the use of SIFT, the large number of unstable points remains a problem. The hypothesis is that using regions-of-interest will result in fewer and more stable landmarks, since larger areas contribute to the regions-of-interest, providing more evidence and making the method less susceptible to noise. Similar findings were done by Frintrop & Jensfelt (2008) and Mikolajczyk et al. (2005). The disadvantage of using regions instead of points, however, could be that they are more susceptible to changes in viewpoint, which deteriorates the stability of the landmarks. In this chapter, the robustness and stability of the symmetrical regions-of-interest are investigated and compared to those of the

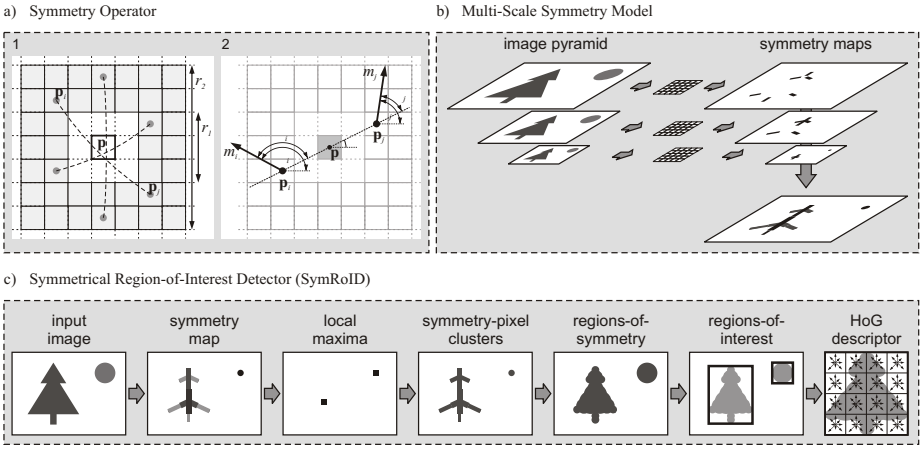

Figure 9.1: The Symmetrical Region-of-Interest Detector. The symmetry operator (a) is at the basis of the detector. All pixel pairs that lie in the symmetry kernel (the gray marked area) contribute to the symmetry value at position p (a1). The symmetry contribution of a pixel pair is determined by comparing the gradients of both pixels (a2). In the multi-scale symmetry model (b), an image pyramid is constructed. The symmetry operator is applied to all images in the pyramid. The resulting symmetry maps are rescaled, and summed up to result in the multi-scale symmetry map. The complete SymRoID model (c) first calculates the symmetry map for the input image. In this map, the local maxima are found. These local maxima serve as the seeds for the region-growing algorithm that clusters all symmetry pixels. Next, the highest contributing radius of each symmetry pixel in a cluster is found and marked with a circle in the regions-of-symmetry map. The regions-of-interest are finally determined by the bounding box of the marked regions. Subsequently, the regions are described using histograms-of-gradients (HoGs).

MUST and SIFT interest points. Moreover, the SLAM performance is tested using the same annotated SLAM database as used in the previous chapter.

The results show that the detection of regions-of-interest based on symmetry results in landmarks that are more robust and stable then interest points detected using contrast features. Moreover, we test the performance of our landmark selection method on a SLAM database that we annotated with ground-truth positions. The results show that our model results in better SLAM performance.

9.2 Symmetrical Region-of-Interest Detector

The proposed Symmetrical Region-of-Interest Detector, SymRoID, is not based on the symmetrical interest-point detector presented in Chapter 8, but is an extension of the symmetry-saliency model that we proposed in Chapter 3 to predict human eye fixations. In Section 9.2.1, the symmetry operator of Reisfeld et al. (1995) is described, which forms the basis of the symmetry-saliency model. The developed multi-scale symmetry model is described in Section 9.2.2. Next, Section 9.2.3 describes the SymRoID model, which uses the saliency maps that result from the symmetry-saliency model to find symmetrical regions-of-interest in the image. Finally, the descriptor to create a scale-invariant representation of the regions is discussed in Section 9.2.4. The description of the symmetry operator and the multi-scale symmetry model is largely a recapitulation of the description of the symmetry-saliency model in Chapter 3. However, there are a few differences.

9.2.1 The Symmetry Operator

For every position, $\mathbf{p} = (x, y)$, in the image, a symmetry kernel is applied that calculates the amount of symmetry by comparing the intensity gradients of the surrounding pixels. Pixel pairs in the neighborhood contribute to the symmetry value. A pixel pair consists of pixel $\mathbf{p}_i$ and $\mathbf{p}_j$, so that $\mathbf{p} = (\mathbf{p}_i + \mathbf{p}_j)/2$ (see Figure 9.1a). The contribution of the pair is calculated by comparing the intensity gradient $\overrightarrow{g_i}$ at $\mathbf{p}_i$ and gradient $\overrightarrow{g_j}$ at $\mathbf{p}_j$ according to:

$$s(i,j) = d(i,j,\sigma) \cdot c(i,j) \cdot \log(1+m_i) \cdot \log(1+m_j) \tag{9.1}$$

where m_i is the magnitude of the gradient, and $d(i,j,\sigma)$ is a Gaussian weighting function on the distance between p_i and p_j with a standard deviation of σ. The multiplication with the gradient magnitudes assures that only strong edges contribute. The logarithm attenuates the influence of large magnitude values, which might be a result from noise. The symmetry measurement is:

$$c(i,j) = (1 - \cos(\gamma_i + \gamma_j)) \cdot (1 - \cos(\gamma_i - \gamma_j)) \tag{9.2}$$

where $\gamma_i = \theta_i - \alpha$ is the angle between the orientation of gradient, θ_i, and the angle, α, of the line between $\mathbf{p}_i$ and $\mathbf{p}_j$ (see Figure 9.1a). The first term in (9.2) has a maximum value when $\gamma_i + \gamma_j = \pi$, which is true for gradients that are mirror symmetric with respect to $\mathbf{p}$. Using only this term would result in high value for points that lie on a straight edge. Since we are not interested in edge detection, but in finding the centers of symmetrical patterns, the second term demotes pixel pairs with similar gradient orientations.

The symmetry value at position $\mathbf{p}$ is calculated by summing up the contributions of all pixel pairs in the neighborhood. Differently from the model in Chapter 3, this neighborhood is defined by an inner and an outer square centered around $\mathbf{p}$. The size of the sides of the squares are respectively r_1 and r_2 (see Figure 9.1a). All pixels that lie inside the outer square, but outside the inner square are considered. $\Gamma(\mathbf{p})$ is the set of contributing pairs. In our experiments we used $r_1 = 5$ and $r_2 = 17$. Smaller values of r_1 result in too small symmetry patterns, and larger values of r_2 are too computationally expensive, and make the operator view the image with too much detail. The total symmetry value at $\mathbf{p} = (x, y)$ is then:

$$S_l(x,y) = \sum_{(i,j)\in\Gamma(\mathbf{p})} s(i,j) \tag{9.3}$$

where S_l is the symmetry map at scale l. The different scales are discussed in the next section.

9.2.2 *The Multi-Scale Symmetry Model*

A region-of-interest detector for visual SLAM needs to be able to detect structures of various sizes since the appearance of landmarks changes drastically when the robot moves around in the environment. Although the symmetry operator can detect symmetry within the neighborhood radius, it cannot detect patterns on larger scales. Increasing the radius is not a good idea due to the quadratic complexity of the operator. Moreover, at larger radii, the operator takes into account too much detail, making the operator more susceptible to noise. Therefore, we propose a multi-scale symmetry model, similar to that used in Chapter 3.

In Figure 9.1b, the multi-scale symmetry model is depicted. The scale space consists of an image pyramid that is built by progressively applying a Gaussian filter to the

image, followed by a downscaling of the image by a factor of two, where scale zero is the image in its original resolution. Secondly, the symmetry operator is applied to all images in the pyramid, resulting in a paramid of symmetry maps. Finally, the symmetry maps at the different scales are resized to the size of the first scale, and then summed up to result in the overall symmetry map:

$$S(x,y) = \bigoplus_{l=L_1}^{L_2} S_l(x,y) \tag{9.4}$$

where L_1 is the first, and L_2 is the last scale. The operator $\oplus$ rescales all maps to the first scale, and subsequently sums the values of the different scales.

Since we are interested in all symmetrical regions in our robotic system, we do not apply the normalization that we used in Chapter 3, because that promotes symmetry maps with only one dominant salient point. Instead, the values in the saliency map are normalized between 0 and 1.

9.2.3 The SymRoID Model

A simplified flow chart of the complete SymRoID model is given in Figure 9.1c. It consists of a number of steps:

1. The symmetry map is calculated by the multi-scale symmetry model as described earlier.

2. Local maxima. A pixel $(\mathrm{x_m}, \mathrm{y_m})$ is a local maximum if it has the highest value in its 3×3 pixels neighborhood, and its symmetry value $S(\mathrm{x_m}, \mathrm{y_m}) \geq \tau$, where we used $\tau = 0.5$ in our experiments.

3. The local maxima are seeds for a region-growing algorithm. The flood-fill algorithm that we applied, takes a local maximum, and grows the area to add all neighboring pixels that have a symmetry value of $S(x,y) \geq \lambda \cdot S(\mathrm{x_m}, \mathrm{y_m})$, where the threshold is a ratio, λ, of the symmetry value of the local maximum. Connecting regions are merged. The region growing results in clusters of symmetry pixels. In our experiments, we used $\lambda = 0.5$.

4. The symmetry-pixel clusters contain the pixels that are the centers of symmetry. Since we are interested in symmetrical regions-of-interest, the complete symmetrical pattern that contributed to these symmetry centers needs to be found. To do so, the radius that has the highest contribution to its symmetry value is stored for every pixel. If $\mathbf{p}_i$ and $\mathbf{p}_j$ form the pixel pair with the highest symmetry contribution $s_{\max}(i,j)$, then $r_s = \|\mathbf{p}_i - \mathbf{p}_j\|/2$ is the maximally contributing symmetry radius. A circle with center $\mathbf{p}$ and radius r_s is then marked in the regions-of-symmetry map.

5. Finally, the regions-of-interest are determined by taking the bounding box of the different regions in the regions-of-symmetry map. The regions-of-interest can overlap.

Some examples of regions-of-interest found by SymRoID can be found in Figure 9.2. It shows two pairs of subsequent images from the SLAM database.

9.2.4 The Region-of-Interest Descriptor

The detected regions-of-interest are described using a histograms-of-gradients (HoGs) descriptor, similar to the SIFT descriptor (Lowe, 2004). A region is first resampled to a 16×16 pixels descriptor window. This window is then divided into 16 squares (see Figure 9.1c). For each square, a histogram-of-gradients is calculated from the intensity gradients of the 4×4 pixels that are in the square. Such a histogram contains 8 bins, for the different gradient orientations, i.e., $[0, \frac{1}{4}\pi), [\frac{1}{4}\pi, \frac{2}{4}\pi)$, etc.

The values of the 8 bins in each of the 16 histograms form the 128-dimensional region-of-interest descriptor. Following (Lowe, 2004), the descriptor is normalized to achieve invariance to changes in intensity, resulting in a vector of unit length.

Since the descriptor window adapts to the size of the region-of-interest, and the size of the region-of-interest itself is determined by the observed symmetrical pattern, the SymRoID model is scale invariant. This makes it possible to detect a landmark from different distances. Moreover, the descriptor is relatively invariant to small translational changes due to noise and affine transformations due to change of perspective, as discussed in (Lowe, 2004). Unlike the standard SIFT descriptor (Lowe, 2004) and the symmetrical interest-point detector proposed in Chapter 8, we did not add rotational

invariance, since our robot drives on flat surfaces, and will therefore not encounter rotational transformations of the stimulus.

9.3 The Visual SLAM System

To ensure that only landmarks are added to the map of the environment that are stably detectable over a number of sequential observations, we use a *visual buffer* that tests the stability of the regions-of-interest. When a region passes the buffer, it is added as a landmark to the *SLAM system*.

9.3.1 The Visual Buffer

Like in the work presented in the Chapter 8, a visual buffer is used to test the stability of the regions-of-interest to make sure that only stable landmarks are added to the map of the environment. Although the functionality of the visual buffer is the same as that discussed in Section 8.2.2, it is recapitulated here for reasons of completeness.

The visual buffer contains the N most recent camera images. The regions in the current image are compared to those in the $N-1$ previous images. A region, i, passes the buffer if it is matched in at least M of the previous images. Two regions, i and j, match when the descriptors, $\mathbf{d}_i$ and $\mathbf{d}_j$, are sufficiently similar. This is true when the Euclidean distance is below the threshold τ_1,

$$\|\mathbf{d}_i - \mathbf{d}_j\| < \tau_1, \tag{9.5}$$

and when the *best-to-next-best ratio* is smaller than the threshold δ_1,

$$\|\mathbf{d}_i - \mathbf{d}_j\| / \|\mathbf{d}_i - \mathbf{d}_l\| < \delta_1, \tag{9.6}$$

where $\mathbf{d}_l$ is the descriptor of the second most-similar region in the previous image. This ratio ensures uniqueness.

The parameter settings used for SymRoID compared with those of MUST and SIFT are:

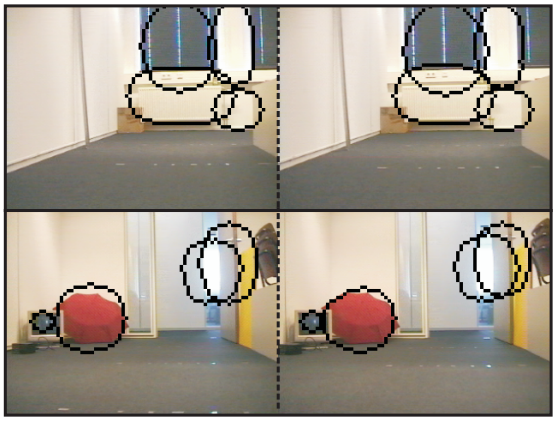

Figure 9.2: Examples of regions-of-interest found in the images. Both pairs contain two sequential images. Note that the bounding box of a region is used to calculate the descriptor.

	N	M	τ_1	δ_1	v
SymRoID	7	5	0.6	0.8	-
MUST	12	4	0.3	0.8	0.6
SIFT	14	8	0.3	0.5	0.0

v is not used in the visual buffer for SymRoID.

9.3.2 The SLAM system

We use a standard implementation of the Extended Kalman Filter (EKF) as basis of the SLAM system (see Appendix C). Our methods and results, however, should also be valid for other SLAM approaches. In this section, the incorporation of the landmark observations in EKF-SLAM is discussed. The method is largely similar to that presented in Section 8.2.3.

The position of a landmark in the environment that pass the visual buffer is obtained by comparing the different observations of the region. Estimates of the position are made by triangulation using the bearings of the observations and the displacement of

the robot, and by inferring depth information from the change in area of the regions-of-interest and the displacement of the robot. This results in a set of K estimations:

$$\mathrm{P} = \{\mathbf{p}_k | \mathbf{p}_k = \langle r_k, \theta_k \rangle \wedge 1 \leq k \leq K\}, \tag{9.7}$$

where r_k and θ_k are respectively the range and bearing of the estimation. The position of the landmark is then determined by the mean of P, and the uncertainty by its covariance matrix.

A landmark i with descriptor $\mathbf{d}_i$ that results from the buffer is classified as either a new landmark, or a previously observed landmark that is already in the map. It concerns a previously observed landmark if the landmark in the database with the most similar descriptor, $\mathbf{d}_j$, fulfills three criteria:

1. Similarity in descriptors:
$$||\mathbf{d}_i - \mathbf{d}_j|| < \tau_2 \tag{9.8}$$

2. A small best-to-next-best ratio to only match unique landmarks:
$$||\mathbf{d}_i - \mathbf{d}_j|| / ||\mathbf{d}_i - \mathbf{d}_l|| < \delta_2 \tag{9.9}$$
where $\mathbf{d}_l$ is the second most similar descriptor in the database.

3. A small distance in the EKF map, measured by the Mahalanobis distance:
$$\sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{S}_j^{-1} (\mathbf{x}_i - \mathbf{x}_j)} < \eta \tag{9.10}$$
where $\mathbf{S}_j$ is the uncertainty covariance matrix, discussed in the next paragraph.

The parameter values used for SymRoID compared to those used for MUST and SIFT are:

	N	M	τ_1	δ_1	v	τ_2	δ_2	η
SymRoID	7	5	0.6	0.8	-	0.5	0.5	0.5
MUST	14	8	0.3	0.8	0.8	0.3	0.7	0.5
SIFT	14	8	0.3	0.5	0.8	0.6	0.7	0.5

The landmark is classified as new only if none of the three criteria is fulfilled. For a new landmark, the state matrix and covariance matrix are augmented using the observation,

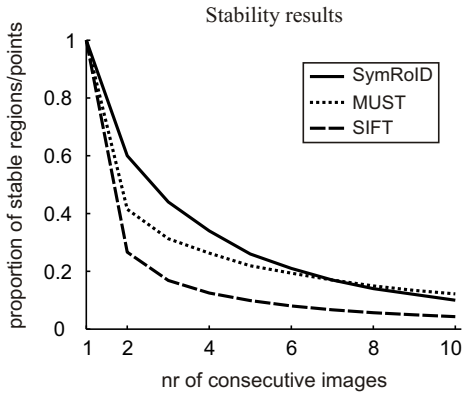

Figure 9.3: Stability of the regions-of-interest and interest points as a function of th number of consecutive images. The graphs show the proportion of regions (for SymRoID) or points (for MUST and SIFT) that are stably found in all observed images.

$\mathbf{z}_i$, and the uncertainty covariance matrix, $\mathbf{S}_i$, where $\mathbf{z}_i$ is set to mean(P), and $\mathbf{S}_i$ is determined using the uncertainty of the observation, cov(P), and the uncertainty of the robot's position in the EKF. When a landmark is matched with an existing landmark in the database, $\mathbf{z}_i$ and $\mathbf{S}_i$ are used to update the EKF.

9.4 Experiments and Results

9.4.1 Experimental Setup

To test the stability and robustness of the symmetrical regions-of-interest, and to test the SLAM performance using SymRoID, the same SLAM database is used as introduced in Chapter 8. The database is recorded with a Pioneer II DX robot, equipped with a Sony D31 camera. The database contains the camera images and the odometric information of ten different runs, in which the robot drove four laps through an office environment. Each lap was approximately 35 meters, and the robot drove at an average speed of 0.3 m/s. Camera images of 320×240 were stored at 5Hz. At intervals of one meter, the true location of the robot was logged by hand. This enabled us to quantify the SLAM

performance.

In the experiments, we tested the performance of SymRoID, and compared it to MUST (Chapter 8) and SIFT Lowe (2004). The stability, robustness, and SLAM-performance experiments are setup identically to those discussed in the previous chapter in Section 8.3.

9.4.2 Stability

To test the stability of the regions-of-interest, multiple sequences of images recorded by the robot are taken. A region or point in the last image of the sequence is considered stable if it is matched in all the previous images of the sequence. Two regions match when the Euclidean distance between the two descriptors, $\mathbf{d}_i$ and $\mathbf{d}_j$, is $\|\mathbf{d}_i - \mathbf{d}_j\| < 0.6$.

Figure 9.3 shows the stability of the symmetrical regions-of-interest compared to the stability of interest points obtained by MUST (Chapter 8) and SIFT (Lowe, 2004). It can be appreciated that the stability of the symmetry methods is higher than that of SIFT. Moreover, the use of symmetrical regions-of-interest gives considerably higher stability than the use of interest points. These results show that the symmetrical regions-of-interest are stably tracked over a number of sequential images, suggesting that they are more robust to small changes in viewpoint than MUST and SIFT.

9.4.3 Robustness

In Mikolajczyk et al. (2005) a benchmark for testing the robustness of detectors is presented. Unfortunately, it is not suited for testing landmark selection methods for SLAM in indoor environments since it contains outdoor scenes. To test the robustness of SymRoID in indoor environments, a subsample of our SLAM database is used. Images are taken from one of the runs in the database with intervals of 3 meters, so that the complete environment is represented in the subsample. To test the noise robustness of SymRoID, we smoothed the image with a Gaussian kernel and added Gaussian noise to the pixels. In addition, we manipulated the contrast and brightness of the images to test the robustness to changing light conditions (see Section 8.3.2 for a full description of the experiments).

The robustness is measured by the proportion of matching regions between the original and the manipulated images. Two regions match when (9.5) and (9.6) are met, where

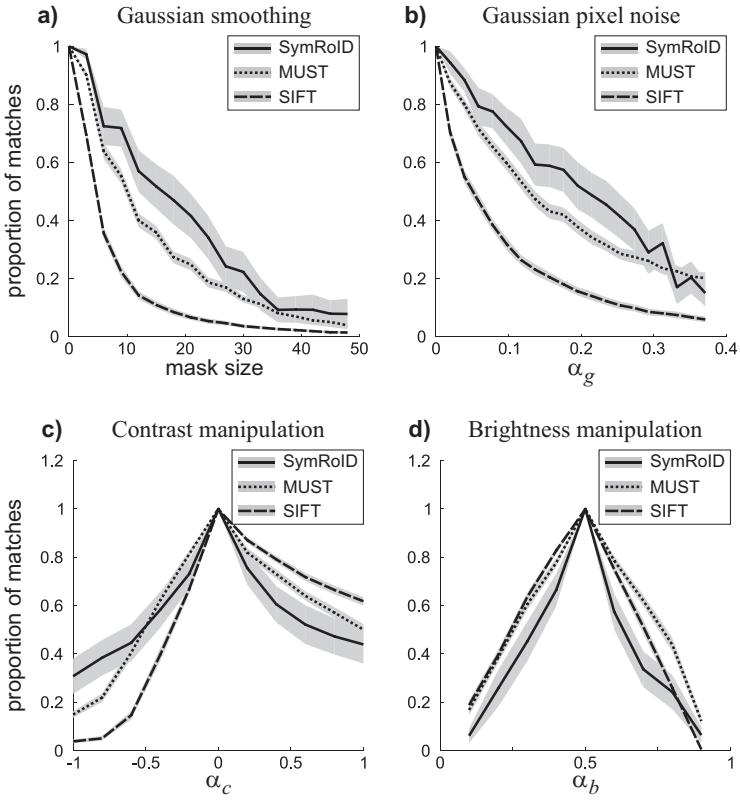

Figure 9.4: Robustness to noise and changing light conditions. The lines give the mean proportion of matched regions-of-interest or interest points between the distorted and the original images of 57 runs. The gray areas around the lines depict the 95% confidence intervals on the mean. Note that these intervals for MUST and SIFT are sometimes small, and therefore hardly visible.

$\tau_1 = 0.6$ and $\delta_1 = 0.75$. Additionally, the distance between the two regions in the image should be less than 3 pixels.

The results in Figure 9.4 show the robustness results. The lines give the mean performance over the 57 images, with the 95% confidence intervals on the mean given by the gray areas. The symmetry models are significantly less affected by the two types

of noise than SIFT (Figure 9.4a and b). Moreover, using regions-of-interest instead of interest points gives a significant improvement. The reason for this success is likely to be the larger area that is used as evidence for the presence of symmetry. SymRoID is therefore less vulnerable to noise.

For the contrast manipulation, SymRoID performs significantly better when there is low contrast. With enhances contrast, on the other hand, SIFT performs better (Figure 9.4c). Using regions-of-interest results in worse performance for the brightness manipulation. MUST is the best method when the brightness is enhanced (Figure 9.4d). The main influence of the contrast manipulations is on the magnitudes of the image gradients. This influences the symmetry value in Equation (9.1). Since a threshold is used on the symmetry values to determine which pixels are part of the symmetrical region and which not, the change of light conditions can thus result in larger or smaller regions. This explains the reduced performance of SymRoID in changing light conditions.

9.4.4 SLAM Performance

To test the SLAM performance, the first three laps of each run are used to train the EKF. The fourth and last lap is used to determine the estimation error for that run. This is done by calculating the Euclidean distance between the EKF estimation of the robot's position and the ground-truth position. The estimation error is the average distance in the last lap.

The parameters for the buffer and for the matching of regions with the landmark $(\tau_1, \tau_2, \delta_1, \delta_2)$ are optimized for all three landmark selection methods. The overall best settings for SymRoID, MUST, and SIFT are given in section 9.3.

Figure 9.5 shows the estimation error. The bars give the mean over the 10 runs, and the error bars depict the 95% confidence intervals. The horizontal dashed line and the horizontal gray bar show the mean and 95% confidence interval of the odometry error. It can be appreciated that the use of symmetry by SymRoID and MUST gives significantly better SLAM performance. Moreover, the use of symmetrical regions-of-interest significantly outperforms the use of interest points. This is true for both the best settings per run and the overall best settings.

Also computationally, SymRoID outperforms the other models. SIFT selects on average 120 interest points, MUST 40, but SymRoID selects only half a dozen regions per

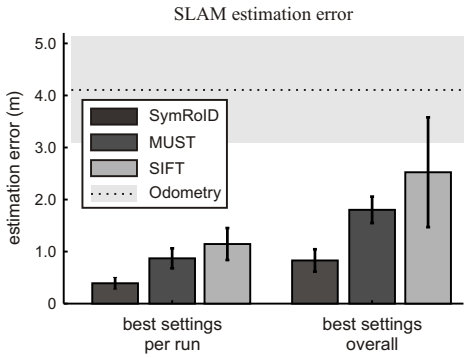

Figure 9.5: The slam performance. The bars give the mean estimation error in the last lap through the environment, after a map has been established in the previous three laps. The error bars are the 95% confidence intervals on the mean. The first group of bars give the results where the best buffer and matching settings per run are used, the second group shows the performance for the overall best settings. The horizontal dashed line gives the mean error when only the odometry information is used. The height of the horizontal gray bar represents the 95% interval.

image, thereby greatly reducing the computations in the buffer. Moreover, both SIFT and MUST take in the order of a second to calculate interest points from an image, while SymRoID finds regions about four times as fast. This improvement is due to the coarse scale space used by SymRoID, in contrast with the detailed scale space used by SIFT and MUST.

9.5 Discussion

In this chapter, a symmetrical region-of-interest detector, SymRoID, is presented. The model selects regions-of-interest in the image using local symmetry. The model is used to select landmarks for a visual SLAM system. The stability, robustness, and SLAM performance of the model is tested and compared with that of MUST, a symmetrical interest-point detector proposed in Chapter 8, and SIFT, an interest-point detector based on contrast features Lowe (2004). The results showed that the use of symmetry improves the stability and robustness to noise, and yields significantly better SLAM

performance. Moreover, the use of symmetrical regions-of-interest outperforms the use of interest points. SymRoID also is more robust to decreased contrast in the image. However, for enhanced contrast and brightness manipulation, the model scores worse than the others.

The higher stablity of the symmetrical regions of interest shows that the regions are more robust to small changes in perspective. This is probably due to the descriptor window, which is adapted to the size of the region of interest. When there is a change in perspective, the size of the region will change and the descriptor region will adapt. This makes it more robust to perspective changes than the interest points, which always have a squared descriptor window.

The improved robustness to noise of the symmetry model can be explained by the fact that symmetrical regions are intrinsically redundant. Both sides of a mirror-symmetrical pattern contain the same information. This redundancy makes the detection less susceptible to noise. Furthermore, by using regions, instead of points, more evidence of the existence of symmetry is gathered, making the model more robust. Robustness to noise is an important property, since robots usually operate under noisy conditions.

Due to the fact that symmetrical patterns are redundant, the descriptor of the symmetrical regions of interest is redundant as well. It should therefore be possible to decrease the dimensionality of the descriptor by disregarding one half of the symmetrical pattern, while maintaining the same descriptive power. Future work needs to show if this is the case.

The worse robustness to changing light conditions can be explained by the role of the gradient magnitudes in the SymRoID model. When light conditions change, the gradient magnitudes in the image are influenced. This results in different symmetry values calculated by Equation (9.1), which has a consequences for the size of the symmetrical regions, due to the fixed threshold used to determine which pixels are member of the symmetrical regions. More research is required to overcome this problem, since robustness to changing light conditions is an important property when a robot needs to operate in an environment over extended periods.

As can be seen in Figure 9.2, the symmetrical regions-of-interest often coincide with objects. The umbrella, radiator, and blinds, for instance, are detected as interesting regions. This shows that symmetry can be used as a bottom-up object detector. This confirms the Gestalt notion that symmetry is a cue for figure-ground segregation (see Chapter 5 for a full discussion). However, it can also be seen in the figure that non-

objects are selected. A clear example is the area between the two blinds. This area is indeed symmetrical, however, it does not coincide with an object, but rather with the background. To solve these kind of situations, other Gestalt principles need to be incorporated in the region-of-interest detector. The principle of *closure*, for instance, will reject the area in between the blinds as an object. We propose to use more Gestalt principles for figure-ground segregation to develop bottom-up object detectors.

9.6 *Visual Attention and Active Vision in Machines*

Part II of the dissertation discussed visual attention and active vision in artificial vision systems. In Chapter 7, an active approach was proposed for object recognition. The results showed that recognition greatly improves when a robot can explore an object. The active exploration gives it the possibility to observe the object from different viewpoints. This not only builds better three-dimensional object models, it also greatly simplifies the segmentation of the object from its background and enables the robot to test the stability of interest points so that only robust points are included in the object representation.

Inspired by the findings in Part I, symmetry is used in Chapter 8 and Chapter 9 to detect respectively interest points and regions of interest in camera images to serve as visual landmarks, thereby exploiting the many symmetrical patterns present in most indoor environments. The results show that landmarks selected by symmetry are more stable, have a higher repeatability, and are more robust to noise and changing light conditions than those selected by contrast features. Moreover, the SLAM performance improves when symmetry is used, showing that valuable landmarks are selected.

The next chapter concludes the thesis with a discussion on the results and the main insights gained from the multi-disciplinary study to visual attention and active vision in natural and in artificial systems.

General Discussion

10

Natural and Artificial Vision Systems

This dissertation presents a multi-disciplinary study to active vision and visual attention. The topics have been discussed in natural systems in Part I and in artificial systems in Part II. The main theme, the role of symmetry in visual attention, has been studied from both perspectives, which has led to a better understanding of the topic. This chapter first gives an overview of the main results and conclusions discussed in the thesis. This is followed by a discussion dealing with the benefits of a multi-disciplinary approach, the role of symmetry in bottom-up object detection, and the use of other Gestalt principles in saliency models and computer-vision systems.

10.1 Summary and Conclusions of the Thesis

Part I dealt with overt visual attention and human eye fixations. In the experiments described in Chapter 3, the role of symmetry in the prediction of human eye fixations has been pointed out. Saliency models using local symmetry in the image have been proposed and compared with human eye fixation gathered in an eye-tracking experiment. The saliency maps produced by the symmetry-saliency model correlated well with the locations of the fixations, not only for images selected to contain symmetrical forms, but for a wide variety of different photographic images. Moreover, the amount of local symmetry is higher at the fixation points than in the other parts of the image. Especially early fixations are on locally symmetrical parts. The proposed symmetry-saliency models outperform the saliency model based on contrast. This is mainly because the symmetry models more specifically focus on the symmetrical centers of objects, like what is observed in the human data. The contrast-saliency model, on the other hand, gives a more spread out activation. Local symmetry has been shown to be a good predictor of human visual attention.

The symmetry pop-out experiment discussed in Chapter 4 showed that the search for a target that differs from the distractors in the presence of symmetry is very efficient. This shows that symmetry results in a pop-out effect. Moreover, the results showed a search asymmetry that is congruent with other visual-search experiments. The search for a non-symmetrical target among symmetrical distractors is faster than the search for a symmetrical target among non-symmetrical distractors. Both findings suggest that the detection of symmetry is preattentive. Symmetry is therefore likely to play a role in the guidance of overt visual attention. The scene-memory experiment, also presented in the chapter, showed that the participants do not pay extra attention to the symmetrical

objects in the display when asked to remember a scene. Instead, attention is paid to objects in general. It can be concluded that symmetry is perceived preattentively, and that human visual attention is primarily object oriented.

The finding that visual attention is oriented towards objects sheds a different light on the results of Chapter 3. This has been discussed in Chapter 5. The literature reviewed in the chapter shows that visual attention is object oriented, which is in congruence with the findings of the scene-memory experiment. It was furthermore concluded from Gestalt research that symmetry is a strong cue for the presence of an object. This explains the good correlations of the symmetry-saliency model with the human eye fixations. However, not only symmetry, but also other Gestalt principles are of importance for figure-ground segregation.

It can thus be concluded from Part I of the dissertation that visual attention is object oriented and that symmetry can be used to preattentively detect objects in the scene. This conclusion has been exploited in Part II to guide the visual attention of a mobile robot.

Chapter 7 demonstrated the importance of active vision in robotics. By letting a robot actively explore objects, the problems of selecting stable interest points and segregating points belonging to the object from those belonging to the background are greatly simplified. Moreover, the robot acquires more perspectives of the object in this way, which improves the recognition and solves the object-constancy problem. The results showed a great improvement in object-recognition performance compared to a passive approach. Moreover, a growing-when required (GWR) network was presented that effectively clusters interest points. The GWR network greatly reduces the amount of interest points, thereby reducing the amount computations, while maintaining the good recognition performance.

The main focus of Chapters 8 and 9 was the use of symmetry for the selection of visual landmarks for simultaneous localization and mapping. Using symmetry exploits the presence of many symmetrical patterns in the indoor environments that robots operate in. Chapter 8 introduced the MUlti-scale Symmetry Transform, MUST, an interest-point detector based on local symmetry in the image. MUST was compared to the scale-invariant feature transform (SIFT), which select points based on contrast. The experiments showed that the interest points selected using MUST are more robust to noise, are more stable, and result in a higher repeatability. Most importantly, using MUST resulted in improved SLAM performance, which shows that MUST selects

valuable landmarks to represent the environment.

In Chapter 9, the idea to use symmetry was extended by proposing a symmetrical region-of-interest detector, SymRoID. This model has been based on the same model that is used to predict human eye fixations in Chapter 3. Using the symmetry-saliency maps, the model detects the parts in the image with high local symmetry. In subsequent steps, SymRoID detects the symmetrical regions of interest by finding the image patterns that contributed to the local symmetry. To test the model, SymRoID was applied in a robotic SLAM system. The experiments showed that the robustness and stability of the symmetrical regions outperforms that of the interest points detected by both MUST or SIFT. Moreover, the SLAM performance greatly improved when using SymRoID to select the symmetrical regions as landmarks.

To summarize, this thesis has shown the importance of active vision and the role of symmetry in visual attention. Local symmetry has been shown useful in finding interesting points in the visual field, which cannot only be used to predict human eye fixations, but also to select robust and stable visual landmarks in the environment of a mobile robot.

10.1.1 Insights From a Multi-Disciplinary Study

The natural and artificial perspectives taken to study active vision and visual attention have led to different insights on visual attention and active vision, as illustrated in Figure 10.1. Some of the findings have had meaning only within the particular discipline, but others have influenced the study in the other discipline:

The main insights about natural vision systems that have been obtained are:

1. Symmetry is a good predictor for human eye fixations.
2. Symmetry is perceived preattentively.
3. Visual attention is object oriented.
4. Symmetry is a cue for the presence of an object.

Inspiration taken from natural systems for artificial systems:

5. Perception is active, not passive.

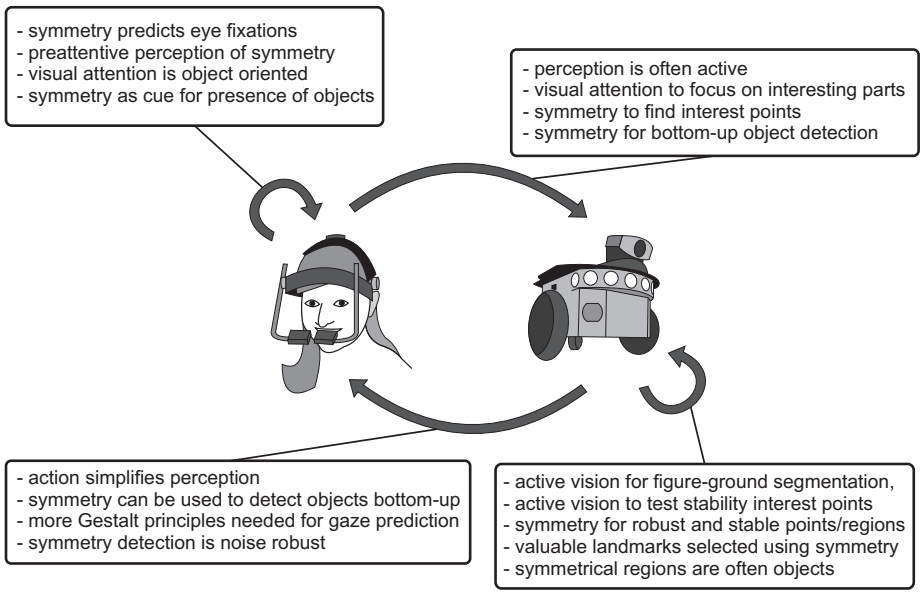

Figure 10.1: The most important insights and inspirations from the multi-disciplinary study to active vision and visual attention.

6. Visual attention is an efficient method to scan the interesting parts of the visual field.
7. Symmetry can be used to detect interest points in an image.
8. Symmetry can be used for bottom-up object detection.

Insights about artificial vision systems:

9. Active vision simplifies figure-ground segmentation.
10. Active vision can be used to obtain good interest points by testing their stability.
11. Interest points detected using symmetry are noise robust and stable.
12. Symmetrical regions-of-interest are even more robust and stable.
13. SLAM performance is improved when symmetry is used to select landmarks instead of contrast.
14. Symmetrical regions often correspond with objects.

Inspiration for natural systems from artificial systems:

15. Perception is active because it simplifies many perceptual tasks in that way.
16. It is indeed possible to preattentively detect objects using symmetry.
17. The use of more Gestalt principles is likely to improve saliency models for the prediction of eye fixations.
18. Detection of symmetry is noise robust. This is a possible reason for the use of symmetry in human vision.

10.2 Discussion

10.2.1 Symmetry for Bottom-Up Object Detection

One of the issues that is raised in both parts of the thesis is the possibility to use symmetry as a cue for figure-ground segregation. As has been discussed in Chapter 5, symmetry is one of the Gestalt principles for figure-ground segregation by humans. This is suggested by the studies presented in Chapter 3 and 4. After all, when visual attention is object oriented and symmetry correlates so well with human visual attention, then symmetry is likely to be a cue for the presence of an object. The possibility of figure-ground segregation using symmetry is also supported by the study of symmetry in artificial vision systems. The symmetrical region-of-interest detector (SymRoID), proposed in Chapter 9, detected regions-of-interest that often corresponded with objects in the environment.

Although the results presented in the thesis strongly suggest the symmetry is a cue for figure-ground segregation, more research needs to be done to investigate the possibilities of applying symmetry for bottom-up object segmentation in computer vision. It is furthermore likely that applying more Gestalt principles will improve the object segmentation. However, it is to be expected that there is a limit on the performance of bottom-up segmentation models. To boost segmentation, the addition of top-down methods to incorporate knowledge about the objects is a necessary step to take.

10.2.2 Gestalt Principles in Saliency Models and Computer Vision

Although symmetry seems to be usable for object detection and segmentation, it is clear from the scene-memory experiment in Chapter 4 that symmetry is not the only cue for the presence of an object. There are other, possibly stronger, cues that point out an object. Also the SymRoID model in Chapter 9 resulted in many regions-of-interest that do not correspond with objects. The open space between two objects, such as for instance between two chairs, is often a symmetrical region as well, and therefore detected as region of interest. This shows that symmetry might be a necessary feature for figure-ground segregation, but not a sufficient one. The Gestalt theory proposes a number of additional principles for preattentive bottom-up figure-ground segmentation, which have been outlined in Section 5.2. We therefore propose to develop bottom-up object detection models based on all the Gestalt principles for improved segmentation performance.

Fowlkes, Martin, & Malik (2003) analysed many figures segmented by humans and found that the properties of these figures coincide with a number of Gestalt principles. Figural regions are generally smaller, more convex, and occur below ground regions. A number of models exist that utilize some of the Gestalt principles to segment objects from their background. The principle of convexity, for instance, was utilized in (Jacobs, 1996; Pao, Geiger, & Rubin, 1999). The principles of symmetry, convexity, and parallelism are combined in (Kikuchi & Fukushima, 2003) to group contours that belong to the foreground. Ren, Fowlkes, & Malik (2006) proposed the use of shapemes, which exploit contextual cues such as parallelism and convexity. As future work, we propose to integrate the symmetry model proposed in this thesis with other models based on Gestalt principles to improve bottom-up object segmentation.

Since human visual attention is object oriented, we hypothesize that such a Gestalt-based object-detection model will be a good predictor of human gaze. When the model can determine the likely locations of objects in an image, it will correspond well with the fixation locations of humans. An appealing property of the proposed object-detection model is that it is purely bottom-up and context-free, that is, no prior knowledge about the scene or the objects in the scene is needed to detect the likely locations of objects.

10.2.3 Hybrid Visual-Attention Models

Although in the previous section it was claimed that the bottom-up context-free property of a Gestalt-based object-detection model is appealing, it also has some clear limitations. As discussed in Chapter 2, we know that human visual attention is often context and task dependent. A purely bottom-up visual-attention model is therefore not capable of good predictions in cases where human visual attention is top-down controlled. To improve performance, visual-attention models therefore need to combine bottom-up and top-down influences.

A number of hybrid approaches have been proposed in the literature. The models of Navalpakkam & Itti (2005, 2006a); Wolfe (2007), for instance, include target knowledge to modulate the bottom-up processes, in order to facilitate attention for certain image features. Other models add a separate target-directed attention module, which calculates the similarity with the target for all parts of the image (Frintrop, 2006; Zelinsky et al., 2006), or bias the bottom-up saliency model towards likely locations to find the target object (Torralba et al., 2006). These models have been discussed in more detail in Section 2.4.2.

The above-mentioned hybrid models include contextual and target knowledge. This is a good starting point to develop a task-specific visual-attention model. However, more top-down knowledge can be incorporated to facilitate search. With knowledge about the task at hand, for instance, the model can use information about the subsequent targets to focus attention on the relevant objects. Search can furthermore be speeded-up if semantic relations between objects are considered in the visual-attention model. A bias for all objects that are semantically related to the current target will facilitate search for related objects. Similarly, spatial relations between objects can be employed as well to facilitate search for the next target.

Publications and Bibliography

Publications

Kootstra, G., & de Boer, B. (2009). Tackling the premature convergence problem in monte-carlo localization. *Robotics and Autonomous Systems*, *57*(11), 1107–1118.

Kootstra, G., de Jong, S., & Schomaker, L. R. B. (2009). Using local symmetry for landmark selection. In M. Fritz, B. Schiele, & J. H. Piater (Eds.) *Computer Vision Systems*, vol. 5815 of *Lecture Notes in Computer Science*, (pp. 94–103). Springer.

Kootstra, G., Nederveen, A., & de Boer, B. (2008a). Paying attention to symmetry. In M. Everingham, C. Needham, & R. Fraile (Eds.) *British Machine Vision Conference (BMVC2008)*, (pp. 1115–1125). Leeds, UK.

Kootstra, G., & Schomaker, L. R. B. (2009a). Prediction of human eye fixations using symmetry. In *Cognitive Science Conference (CogSci)*. Amsterdam, The Netherlands.

Kootstra, G., & Schomaker, L. R. B. (2009b). Using symmetrical regions-of-interest to improve visual SLAM. In *the International Conference on Intelligent RObots and Systems (IROS)*. St. Louis, USA.

Kootstra, G., & Schomaker, L. R. B. (submitted). Prediction of eye fixations on complex visual stimuli using local symmetry. *Journal of Vision*.

Kootstra, G., Ypma, J., & de Boer, B. (2007). Exploring objects for recognition in the real world. In *IEEE International Conference on Robotics and Biomimetics (ROBIO*

'07). Sanya, China.

Kootstra, G., Ypma, J., & de Boer, B. (2008b). Active exploration and keypoint clustering for object recognition. In *International Conference on Robotics and Automation (ICRA)*. Pasadena, CA.

Niessen, M. E., Kootstra, G., de Jong, S., & Andringa, T. C. (2009). Expectancy-based robot localization through context evaluation. In *Proceedings of the International Conference on Artificial Intelligence (ICAI)*, (pp. 371–377).

Zwinderman, M., Rybski, P. E., & Kootstra, G. (submitted). A human-assisted approach for a mobile robot to learn 3D object models using active vision. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. Submitted.

Bibliography

Alata, O., & Quintard, L. (2009). Is there a best color space for color image characterization or representation based on multivariate gaussian mixture model? *Computer Vision and Image Understanding*, *113*(8), 867–877. 120

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, *61*, 183–193. 112

Attneave, F. (1955). Symmetry, information, and memory for patterns. *The American Journal of Psychology*, *68*(2), 209–222. 50

Backer, G., Mertsching, B., & Bollmann, M. (2001). Data- and model-driven gaze control for an active-vision system. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(12), 1415–1429. 129

Bahnsen, P. (1928). Eine untersuchung uber symmetrie und asymmetrie bei visuellen wahrnehmungen. *Zeitschrift für Psychologie*, *108*, 129–154. 110

Bailey, T., & Durrant-Whyte, H. (2006). Simultaneous localization and mapping (SLAM): Part ii. *IEEE Robotics & Automation Magazine*, *13*(3), 108–117. 134, 135, 137, 265

Ballard, D. H. (1991). Animate vision. *Artificial Intelligence*, *48*, 57–86. 139, 147

Barlow, H. B., & Reeves, B. C. (1979). The versatility and absolute efficiency of detecting mirror symmetry in random dot displays. *Vision Research*, *19*, 783–793. 47, 87

Bauer, B., Jolicoeur, P., & Cowan, W. B. (1996). Visual search for color targets that are or are not linearly-separable from distractors. *Vision Research*, *36*(10), 1439–1465. 33, 34, 37, 88, 101

Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, *110*(3), 346–359. 127, 129

Bay, H., Tuytelaars, T., & Van Gool, L. (2006). SURF: Speeded up robust features. In *ninth European Conference on Computer Vision (ECCV)*. Graz, Austria. 123, 127, 134, 169, 189

Baylis, G. C., & Driver, J. (1993). Visual attention and objects: evidence for hierarchical coding of location. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(3), 451–470. 106

Baylis, G. C., & Driver, J. (1994). Parallel computation of symmetry but not repetition within single visual shapes. *Visual Cognition*, *1*, 377–400. 47, 87

Beck, D. M., Pinsk, M. A., & Kastner, S. (2005). Symmetry perception in humans and macaques. *Trends in Cognitive Sciences*, *9*(9), 405–406. 47, 51

Beis, J., & Lowe, D. G. (1997). Shape indexing using approximate nearest-neighbour search in high-dimensional spaces. In *Conference on Computer Vision and Pattern Recognition*, (pp. 1000–1006). Puerto Rico. 150

Belongie, S., Malik, J., & Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Transations of Pattern Analysis and Machine Intelligence*, *2*(4), 509–522. 125

Bindemann, M., Scheepers, C., & Burton, A. M. (2009). Viewpoint and center of gravity affect eye movements to human faces. *Journal of Vision*, *9*(2), 1–16. 48, 57

Bornstein, M. H., & Stiles-Davis, J. (1984). Discrimination and memory for symmetry in young children. *Developmental Psychology*, *20*(4), 637–649. 48

Borotschnig, H., Paletta, L., Prantl, M., & Pinz, A. (2000). Appearance-based active object recognition. *Image and Vision Computing*, *18*, 715–727. 149

Borst, A., & Egelhaaf, M. (1993). Detecting visual motion: Theory and models. In F. A. Miles, & J. Wallmann (Eds.) *Visual Motion and its Role in the Stabilization of Gaze.*, (pp. 3–27). Elsevier Science. 141

Brooks, R. A. (1999). *Cambrian Intelligence: The Early History of the New AI*. The MIT Press. 11

Bruce, H. B., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, *9*(3), 1–24. 39, 56

Carmi, R., & Itti, L. (2006a). The role of memory in guiding attention during natural vision. *Journal of Vision*, *6*(9), 898–914. 27

Carmi, R., & Itti, L. (2006b). Visual causes versus correlates of attentional selection in dynamic scenes. *Vision Research*, *46*, 4333–4345. 39

Carmody, D. P., Nodine, C. F., & Locher, P. J. (1977). Global detection of symmetry. *Perceptual and Motor Skills*, *45*, 1267–1273. 87

Chun, M. M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, *36*, 28–71. 27

Civera, J., Davison, A. J., & Montiel, J. M. M. (2008). Inverse depth parametrization for monocular SLAM. *IEEE Transactions on Robotics and Autonomous Systems*, *24*(5), 932–945. 134

Corballis, M. C., & Roldan, C. E. (1975). Detection of symmetry as a function of angular orientation. *Journal of Experimental Psychology: Human Perception and Performance*, *1*(3), 221–230. 47, 48

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*, 201–215. 29, 30

Csurka, G., Dance, C. R., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. In *European Conference on Computer Vision, Workshop on Statistical Learning in Computer Vision*, (pp. 1–22). 125, 126, 165

Davison, A. J., & Murray, D. W. (2002). Simultaneous localization and map-building using active vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 865–880. 134, 169, 189

Davison, A. J., Reid, I. D., Molton, N. D., & Stasse, O. (2007). MonoSLAM: Real-time single camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(6), 1–16. 133, 169

De Graef, P., Christiaens, D., & d'Ydewalle, G. (1990). Perceptual effects of scene context on object identification. *Psychological Research*, *52*, 317–329. 27

de Jong (2008). *Top-down selective visual attention during SLAM*. Master's thesis, University of Groningen, The Netherlands. 265

De Kuijer, J., Deregowski, J. B., & McGeorge, P. (2004). The influence of visual symmetry on the encoding of objects. *Acta Psychologica*, *116*(1), 75–91. 50

Delius, J. D., & Nowak, B. (1982). Visual symmetry recognition by pigeons. *Psychological Research*, *44*, 199–212. 47

Deregowski, J. B. (1971). Symmetry, gestalt and information theory. *The Quarterly Journal of Experimental Psychology*, *23*(4), 381–385. 50

Dick, M., Ullman, S., & Sagi, D. (1987). Parallel and serial processes in motion detection. *Science*, *237*, 400–402. 34

Dissanayake, G., Newman, P., Durrant-Whyte, H. F., Clark, S., & Csobra, M. (2001). A solution to the simultaneous localisation and mapping (SLAM) problem. *IEEE Transactions on Robotics and Automation*, *17*(3), 229–241. 136

Driver, J., & Baylis, G. C. (1995). One-sided edge assignment in vision: 2. part decomposition, shape description, and attention to objects. *Current Directions in Psychological Science*, *4*(6), 201–206. 106, 112

Driver, J., Baylis, G. C., & Rafal, R. D. (1992). Preserved figure-ground segregation and symmetry perception in visual neglect. *Nature*, *360*, 73–75. 49, 57, 107, 110, 170

Durrant-Whyte, H., & Bailey, T. (2006). Simultaneous localization and mapping: Part i. *IEEE Robotics & Automation Magazine*, *13*(2), 99–108. 134, 135, 137, 177, 265

D'Zmura (1991). Color in visual search. *Vision Research*, *31*(6), 951–966. 33

Egeth, H., & Dagenbach, D. (1991). Parallel versus serial processing in visual search: Further evidence from subadditive effects of visual quality. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(2), 551–560. 38

Einhäuser, W., Rutishauser, U., Frady, E. P., Nadler, S., Köning, P., & Koch, C. (2006). The relation of phase noise and luminance contrast to overt attention in complex visual stimuli. *Journal of Vision*, *6*, 1148–1158. 28

Einhäuser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, *8*(2), 1–19. 28

Enns, J. T., & Rensink, R. A. (1993). A model for the rapid interpretation of line drawings in early vision. In D. Brogan, A. Gale, & K. Carr (Eds.) *Visual Search 2*, (pp. 73–90). London: Taylor and Francis. 105, 106

Evans, C. S., Wenderoth, P., & Cheng, K. (2000). Detection of bilateral symmetry in complex biological images. *Perception*, *29*, 31–42. 47

Farmer, E. W., & Taylor, R. M. (1980). Visual search through color displays: Effects of target-background similarity and background uniformity. *Perception and Psychophysics*, *27*, 267–272. 33

Ferrari, V., Tuytelaars, T., & Van Gool, L. (2006). Simultaneous object recognition and segmentation from single or multiple model views. *International Journal of Computer Vision*, *67*(2), 159–188. 147

Findlay, J. M. (1982). Global visual processing for saccadic eye-movements. *Vision Research*, *22*(8), 1033–1045. Pb047 Times Cited:240 Cited References Count:26. 48, 57, 82

Findlay, J. M., & Gilchrist, I. D. (2003). *Active Vision: The Psychology of Looking and Seeing*. Oxford, UK: Oxford University Press. 25, 26

Fisher, C. B., Ferdinandsen, K., & Bornstein, M. H. (1981). The role of symmetry in infant form discrimination. *Child Development*, *52*(2), 457–462. 48, 112

Fitzpatrick, P. (2003). First contact: an active vision approach to segmentation. In *the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Las Vagas, Nevada. 148

Foster, D. H., & Ward, P. A. (1991). Asymmetries in oriented-line detection indicate two orthogonal filters in early vision. *Proceedings of the Royal Society (London B)*, *243*, 75–81. 34, 36

Foulsham, T., & Underwood, G. (2007). How does the purpose of inspection influence the potency of visual salience in scene perception? *Perception*, *36*(8), 1123–1138. 224GH Times Cited:2 Cited References Count:38. 58

Fowlkes, C., Martin, D., & Malik, J. (2003). On measuring the ecological validity of local figure - ground cues. In *European Conference on Visual Perception*. 216

Friedman, J. H., Bentley, J. L., & Finkel, R. A. (1977). An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software*, *3*(3), 209–226. 150

Frintrop, S. (2006). *VOCUS: A Visual Attention System for Object Detection and Goal-directed Search*, vol. 3899/2006 of *Lecture Notes in Artificial Intelligence (LNAI)*. Springer Berlin/Heidelberg. 40, 217

Frintrop, S., & Jensfelt, P. (2008). Attentional landmarks and active gaze control for visual SLAM. *IEEE Transactions on Robotics*, *24*(5), 1054–1065. 133, 134, 169, 176, 189

Fritzke, B. (1995). A growing neural gas network learns topologies. In *Advances in Neural Information Processing Systems (NIPS'94)*, vol. 7. Denver. 155

Gao, D., Mahadevan, V., & Vasconcelos, N. (2008). On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of Vision*, *8*(7), 1–18. 39

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton-Mifflin. 10, 139, 147

Grammer, K., & Thornhill, R. (1994). Human (homo sapience) facial attractiveness and sexual selection: The role of symmetry and averageness. *Journal of Comparative Psychology*, *108*(3), 233–242. 46

Grill-Spector, K., Kourtzi, Z., & N., K. (2001). The lateral occipital complex and its role in object recognition. *Vision Research*, *41*(10-11), 1409–1422. 50

Guivant, J. E., & Nebot, E. M. (2001). Optimization of the simultaneous localization and map-building algorithm for real-time implementation. *IEEE Transactions on Robotics and Automation*, *17*(3), 242–257. 138, 178, 264

Hargittai, M., & Hargittai, I. (2009). *Visual Symmetry*. Singapore: World Scientific Publishing Co. 44

Harris, C., & Stephens, M. (1988). A combined corner and edge detector. In *The Fourth Alvey Vision Conference*, (pp. 147–151). Manchester, UK. 121, 123, 147

He, P. Y., & Kowler, E. (1989). The role of location probability in the programming of saccades - implications for center-of-gravity tendencies. *Vision Research*, *29*(9), 1165–1181. Ap318 Times Cited:71 Cited References Count:29. 48, 57, 82

Heidemann, G. (2004). Focus-of-attention from local color symmetries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(7), 817–830. 59, 62, 129, 130, 170

Henderson, J. M., Brockmole, J. R., Castelhano, S., Monica, & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In R. v. Gompel, M. Fischer, W. Murray, & R. Hill (Eds.) *Eye movements: A window on mind and brain*, (pp. 537–562). Oxford: Elsevier. 27

Henderson, J. M., & Castelhano, M. S. (2005). Eye movements and visual memory for scenes. In G. Underwood (Ed.) *Cognitive Processes in Eye Guidance*, (pp. 213–235). Oxford University Press. 27

Hochberg, J., & McAlister, E. (1953). A quantitative approach to figural goodness. *Journal of Experimental Psychology*, *46*, 361–364. 112

Hoffman, D. D., & Singh, M. (1997). Salience of visual parts. *Cognition*, *63*, 29–78. 112

Huk, A. C., & Heeger, D. J. (2000). Task-related modulation of visual cortex. *Journal of Neurophysiology*, *83*, 3525–3536. 27

Itti, L., Dhavale, N., & Pighin, F. (2003). Realistic avatar eye and head animation using a neurobiological model of visual attention. In B. Bosacchi, D. B. Fogel, & J. C. Bezdek (Eds.) *SPIE 48th Annual International Symposium on Optical Science and Technology*, vol. 5200, (pp. 64–78). Bellingham, WA: SPIE Press. 39

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10-12), 1489–1506. 39, 56, 63, 247, 252

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203. 56, 247

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259. 33, 35, 39, 40, 54, 56, 57, 58, 63, 64, 80, 122, 129, 247, 248, 249

Jacobs, D. (1996). Robust and efficient detection of salient convex groups. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *18*(1), 23–37. 216

Julesz, B. (1971). *Foundations of Cyclopean Perception*. Chicago: University of Chicago Press. 87

Julesz, B. (1981). Figure and ground perception in briefly presented isodipole textures. In M. Kubovy, & J. R. Pomerantz (Eds.) *Perceptual Organization*, (pp. 27–54). Hillsdale, NJ: Erlbaum. 87

Julesz, B. (1984). A brief outline of the texton theory of human vision. *Trends in Neuroscience*, *7*, 41–45. 34, 35

Kanizsa, G., & Gerbino, W. (1976). Convexity and symmetry in figureground organization. In M. Henle (Ed.) *Vision and artifact*, (pp. 25–32). Springer. 112

Karn, K. S., & Hayhoe, M. M. (2000). Memory representations guide targeting eye movements in a natural task. *Visual Cognition*, *7*(6), 673–703. 27

Kaufman, L., & Richards, W. (1969). Spontaneous fixation tendencies for visual forms. *Perception & Psychophysics*, *5*(2), 85–88. 48, 57, 82

Ke, Y., & Sukthankar (2004). PCA-SIFT: A more distinctive representation for local image descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 506–513). 124, 127

Kienzle, W., Franz, M. O., Schölkopf, B., & Wichmann, F. A. (2009). Center-surround patterns emerge as optimal predictors for human saccade targets. *Journal of Vision*, *9*(5), 7:1–15. 56

Kikuchi, M., & Fukushima, K. (2003). Assignment of figural side to contours based on symmetry, parallelism, and convexity. In *Lecture Notes in Computer Science*, vol. 2774. Heidelberg: Springer Berlin. 216

Kimchi, R., & Peterson, M. A. (2008). Figure-ground segmentation can occur without attention. *Psychological Science*, *19*(7), 660–668. 112

Kimchi, R., Yeshurun, Y., & Cohen-Savransky, A. (2007). Automatic, stimulus-driven attention capture by objecthood. *Psychonomic Bulletin & Review*, *14*(1), 166–172. 112

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, *4*, 219–227. 56, 247

Koffka, K. (1935). *Principles of Gestalt Psychology*. London: Lund Humphries. 49, 107, 108, 112

Köhler, W. (1947). *Gestalt psychology : an introduction to new concepts in modern psychology*. New York: Liveright. Reissued, 1992. 49, 107, 108

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464–1480. 155

Kootstra, G. (2002). *Selection of Landmarks for Visual Landmark Navigation on a Mobile Robot*. Master's thesis, University of Groningen, The Netherlands. 141

Kootstra, G., & de Boer, B. (2009). Tackling the premature convergence problem in monte-carlo localization. *Robotics and Autonomous Systems*, *57*(11), 1107–1118. 139

Kootstra, G., de Jong, S., & Schomaker, L. R. B. (2009). Using local symmetry for landmark selection. In M. Fritz, B. Schiele, & J. H. Piater (Eds.) *Computer Vision*

Systems, vol. 5815 of *Lecture Notes in Computer Science*, (pp. 94–103). Springer. 128, 130, 134, 141, 167

Kootstra, G., Nederveen, A., & de Boer, B. (2008a). Paying attention to symmetry. In M. Everingham, C. Needham, & R. Fraile (Eds.) *British Machine Vision Conference (BMVC2008)*, (pp. 1115–1125). Leeds, UK. 53, 71

Kootstra, G., & Schomaker, L. R. B. (2009a). Prediction of human eye fixations using symmetry. In *Cognitive Science Conference (CogSci)*. Amsterdam, The Netherlands. 53

Kootstra, G., & Schomaker, L. R. B. (2009b). Using symmetrical regions-of-interest to improve visual SLAM. In *Proceedings of the International Conference on Intelligent RObots and Systems (IROS)*. St. Louis, USA. 127, 128, 130, 134, 141, 187

Kootstra, G., & Schomaker, L. R. B. (submitted). Prediction of eye fixations on complex visual stimuli using local symmetry. *Journal of Vision*. 39, 53

Kootstra, G., Ypma, J., & de Boer, B. (2007). Exploring objects for recognition in the real world. In *IEEE International Conference on Robotics and Biomimetics (ROBIO '07)*. Sanya, China. 126, 142, 145

Kootstra, G., Ypma, J., & de Boer, B. (2008b). Active exploration and keypoint clustering for object recognition. In *International Conference on Robotics and Automation (ICRA)*. Pasadena, CA. 126, 127, 142, 145

Kourtzi, Z., & Kanwisher, N. (2001). Representation of perceived object shape by the human lateral occipital complex. *Science*, *293*(5534), 1506–1509. 50

Lakoff, G., & Johnson, M. (1999). *Philosophy in the Flesh: The embodied mind and its challenge to Western thought*. New York: Basic Books. 11

Land, M. F., & Hayhoe, M. M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research*, *41*, 3559–3565. 82

Land, M. F., & McLeod, P. (2000). From eye movements to actions: how batsmen hit the ball. *Nature Neuroscience*, *30*, 1340–1345. 82

Le Meur, O., Le Callet, P., Barba, D., & Thoreau, D. (2006). A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(5), 802–817. 39, 56, 69, 71, 82

Le Meur, O., Le Callet, P., Barba, D., Thoreau, D., & Francois, E. (2004). From low-level perception to high-level perception, a coherent approach for visual attention modelling. In B. E. Rogowitz, T. N. Pappas, & T. N. (Eds.) *Human Vision and Electronic Imaging IX*, vol. 5292 of *Proceedings of the SPIE*, (pp. 284–295). San Jose, CA. 39

Lehrer, M. (1993). Why do bees turn back and look? *Journal of Comparative Physiology A*, *172*, 549–563. 141

Lehrer, M., & Bianco, G. (2000). The turn-back-and-look behaviour: bee versus robot. *Biological Cybernetics*, *83*(3), 211229. 141, 148

Levi, D. M., & Saarinen, J. (2004). Perception of mirror symmetry in amblyopic vision. *Vision Research*, *44*, 24752482. 51

Lewis, M. A., & Nelson, M. E. (1998). Look before you leap: Peering behavior for depth perception. In *Proceedings of the fifth international conference on simulation of adaptive behavior on From animals to animats 5*, (pp. 98–103). Zurich, Switzerland. 140

Lindeberg, T. (1998). Feature detection with automatic scale selection. *International Journal of Computer Vision*, *30*(2), 79–116. 123

Locher, P. J., & Nodine, C. F. (1987). Symmetry catches the eye. In J. O'Regan, & A. Lévy-Schoen (Eds.) *Eye Movements: From Physiology to Cognition*. North-Holland: Elsevier Science Publishers B.V. 48, 57

Locher, P. J., & Nodine, C. F. (1989). The perceptual value of symmetry. *Computers and Mathematics with Applications*, *17*(4-6), 475–484. 87

Locher, P. J., & Wagemans, J. (1993). The effects of element type and spatial grouping on symmetry detection. *Perception*, *22*, 565–587. 87

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, (pp. 1150–1157). Corfu, Greece. 121, 123, 125, 147

Lowe, D. G. (2001). Local feature view clustering for 3d object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. Kauai, Hawaii. 147

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110. 122, 123, 124, 125, 126, 129, 134, 147, 149, 150, 157, 165, 169, 171, 172, 174, 175, 185, 189, 194, 199, 202, 253

Loy, G., & Zelinsky, A. (2003). Fast radial symmetry for detecting points of interest. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(8), 959–973. 130, 170, 171, 173

Machilsen, B., Pauwels, M., & Wagemans, J. (2009). The role of vertical mirror symmetry in visual shape detection. *Journal of Vision*, *9*(12), 1–11. 49

Mannan, S., Ruddock, K. H., & Wooding, D. S. (1995). Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-d images. *Spatial vision*, *9*(3), 363–386. 28

Marola, G. (1989). Using symmetry for detecting and locating objects in a picture. *Computer Vision, Graphics, and Image Processing*, *46*, 179–195. 129

Marsland, S., Shapiro, J., & Nehmzow, U. (2002). A self-organising network that grows when required. *Neural Networks*, *15*, 1041–1058. 126, 150, 155

Maybeck, P. S. (1979). *Stochastic Models, Estimation and Control*, vol. 1. New York: Academic. 262

Metta, G., & Fitzpatrick, P. (2003). Early integration of vision and manipulation. *Adaptive Behavior*, *11*(2), 109–128. 142, 147, 148

Metzger, F. (1953). *Gesetze des Sehens*. Frankfurt-am-Main: Waldemar Kramer. 112

Mikolajczyk, K., & Schmid, C. (2001). Indexing based on scale invariant interest points. In *IEEE Int. Conf. on Computer Vision (ICCV)*. Vancouver, BC. 123

Mikolajczyk, K., & Schmid, C. (2002). An affine invariant interest point detector. In *the 7th European Conference on Computer Vision*, vol. 1, (pp. 128–142). Copenhagen, Denmark. 124, 148

Mikolajczyk, K., & Schmid, C. (2004). Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, *60*(1), 63–86. 124

Mikolajczyk, K., & Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(10), 1615–1630. 125, 127

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., & Van Gool, L. (2005). A comparison of affine region detectors. *International Journal of Computer Vision*, *65*(1/2), 43–72. 180, 189, 199

Moller, A. P., & Thornhill, R. (1998). Bilateral symmetry and sexual selection: a meta-analysis. *The American Naturalist*, *151*(2), 174–192. 46

Montemerlo, M., Thrun, S., Koller, D., & Wegbreit, B. (2002). FastSLAM: A factored solution to the simultaneous localization and mapping problem. In *Proceedings of the AAAI National Conference on Artificial Intelligence*. Edmonton, Canada: AAAI. 136, 137

Montemerlo, M., Thrun, S., Koller, D., & Wegbreit, B. (2003). FastSLAM 2.0: An improved particle filtering algorithm for simultaneous localization and mapping that provably converges. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI)*. Acapulco, Mexico. 136, 137, 178

Moravec, H. (1988). *Mind Children*. Harvard University Press. 11

Moreels, P., & Perona, P. (2007). Evaluation of features detectors and descriptors based on 3D objects. *International Journal of Computer Vision*, *73*(3), 263–284. 125, 147, 148, 169, 189

Mozos, O. M., Gil, A., Ballesta, M., & Reinoso, O. (2008). Interest point detectors for visual SLAM. In D. Borrajo, L. Castillo, & J. M. Corchado (Eds.) *Lecture Notes in Computer Science*, vol. 4788, (pp. 170–179). Springer-Verlag. 125, 169, 189

Murillo, A. C., Guerrero, J. J., & Sagues, C. (2007). SURF features for efficient robot localization with omnidirectional images. In *IEEE International Conference on Robotics and Automation (ICRA)*, (pp. 3901–3907). Rome, Italy. 134, 169, 189

Nagy, A. L., & Sanchez, R. R. (1990). Critical color differences determined with a visual search task. *Journal of the Optical Society of America*, *7*(7), 1209–1217. 33, 36, 37

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research*, *45*, 205–231. 40, 217

Navalpakkam, V., & Itti, L. (2006a). An integrated model of top-down and bottom-up attention for optimal object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 2049–2056). 40, 217

Navalpakkam, V., & Itti, L. (2006b). Top-down attention selection is fine grained. *Journal of Vision*, *6*, 1180–1193. 40

Neider, M. B., & Zelinsky, G. J. (2006). Scene context guides eye movements during visual search. *Vision Research*, *46*, 614621. 27

Nolfi, S. (1996). Adaptation as a more powerful tool than decomposition and integration. In *the Workshop on Evolutionary Computing and Machine Learning, 13th International Conference on Machine Learning*. Bari, Italy. 149

Nothdurft, H. C. (1991). The role of local contrast in pop-out of orientation, motion, and color. *Investigative Ophthalmology and Visual Science*, *32*(4), 714. 34

Noton, D., & Stark, L. W. (1971a). Scanpaths in eye movements during pattern perception. *Science*, *171*(3968), 308–311. 27, 58

Noton, D., & Stark, L. W. (1971b). Scanpaths in saccadic eye movements while viewing and recognizing patterns. *Vision Research*, *11*, 929–942. 27, 58

Odekar, A., Hallowell, B., Kruse, H., Moates, D., & Lee, C. Y. (2009). Validity of eye movement methods and indices for capturing semantic (associative) priming effects. *Journal of Speech, Language, and Hearing Research*, *52*, 31–48. 28

Olivers, C. N., & van der Helm, P. A. (1998). Symmetry and selective attention: a dissociation between effortless perception and serial search. *Perceptual & Psychophysics*, *60*(7), 1101–1116. 88, 91, 93, 100

Olmos, A., & Kingdom, F. A. A. (2004). Mcgill calibrated colour image database, http://tabby.vision.mcgill.ca. 66

Ottes, F. P., Van Gisbergen, J. A. M., & Eggermont, J. J. (1984). Metrics of saccade responses to visual double stimuli: Two different modes. *Vision Research*, *24*(10), 1169–1179. 48, 57, 82

Ouerhani, N., von Wartburg, R., Hügli, H., & Müri, R. (2004). Empirical validation of the saliency-based model of visual attention. *Electronic Letters on Computer Vision and Image Analysis*, *3*(1), 13–14. 39, 56, 69, 71

Paletta, L., & Pinz, A. (2000). Active object recognition by view integration and reinforcement learning. *Robotics and Autonomous Systems*, *31*, 71–86. 149

Palmer, S. E. (1991). Goodness, gestalt, groups, and garner: Local symmetry subgroups as a theory of figural goodness. In G. R. Lockhead, & J. R. Pomerantz (Eds.) *The Perception of Structure. Essays in Honor of Wendell R. Garner*, (pp. 23–40). Washington, DC: American Psychological Association. 47, 112

Palmer, S. E. (1992). Modern theories of gestalt perception. In G. W. Humphreys (Ed.) *Understanding Vision: An Interdisciplinary Perspective – Readings in Mind and Language*, (pp. 39–70). Oxford, England: Blackwell. 108

Palmer, S. E., & Hemenway, K. (1978). Orientation and symmetry: Effects of multiple, rotational, and near symmetries. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(4), 691–702. 47, 48, 57, 87, 170

Pao, H., Geiger, D., & Rubin, N. (1999). Measuring convexity for figure/ground separation. In *Proceedings of the Seventh International Conference on Computer Vision (ICCV)*, vol. 2. 216

Parkhurst, D. J., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*, 107–123. 39, 56, 58, 69, 79, 82

Parkhurst, D. J., & Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision*, *16*(2), 125–154. 82

Parkhurst, D. J., & Niebur, E. (2004). Texture contrast attracts overt visual attention in natural scenes. *European Journal of Neuroscience*, *19*, 783–789. 39

Pfeifer, R., & Scheier, C. (1999). *Understanding Intelligence*. Cambridge, MA: MIT Press. 11, 18, 147

Pomerantz, J. R. (2006). Colour as a gestalt: Pop out with basic features and with conjunctions. *Visual Cognition*, *14*(4-8), 619–628. 41, 106

Pomerantz, J. R., Sager, L. C., & Stoever, R. J. (1977). Perception of wholes and of their component part: Some configural superiority effects. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(3), 422–435. 42, 106

Privitera, C. M., & Stark, L. W. (1998). Evaluating image processing algorithms that predict regions of interest. *Pattern Recognition Letters*, *19*, 1037–1043. 39

Privitera, C. M., & Stark, L. W. (2000). Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(9), 970–982. 39, 56

Rao, R. P. N., Zelinsky, G. J., Hayhoe, M. M., & Ballard, D. H. (2002). Eye movements in iconic visual search. *Vision Research*, *42*, 1447–1463. 40

Raphael, B. (1976). *The Thinking Computer: Mind Inside Matter*. W.H.Freeman & Co Ltd. 11

Reisfeld, D., Wolfson, H., & Yeshurun, Y. (1995). Context-free attentional operators: The generalized symmetry transform. *International Journal of Computer Vision*, *14*, 119–130. 59, 61, 130, 170, 191

Ren, X., Fowlkes, C. C., & Malik, J. (2006). Figure/ground assignment in natural images. In *Proceedings of the European Conference of Computer Vision (ECCV)*. 216

Rensink, R. A., & Enns, J. T. (1995). Pre-emption effects in visual search: Evidence for low-level grouping. *Psychological Review*, *102*(1), 101–130. 105

Rhodes, G., Proffitt, F., Grady, J. M., & Sumich, A. (1998). Facial symmetry and the perception of beauty. *Psychonomic Bulletin & Review*, *5*(4), 659–669. 46

Rothganger, F., Lazebnik, S., Schmid, C., & Ponce, J. (2006). 3D object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *International Journal of Computer Vision*, *66*(3), 231–259. 147

Rothkopf, C. A., Ballard, D. H., & Hayhoe, M. M. (2007). Task and context determine where you look. *Journal of Vision*, *7*(14), 1–20. 27

Roy, S. D., Chaudhury, S., & Banerjee, S. (2004). Active recognition through next view planning: a survey. *Pattern Recognition*, *37*, 429–446. 149

Royden, C. S., Wolfe, J. M., Konstantinova, E., & Hildreth, E. C. (1996). Search for a moving object by a moving observer. *Investigative Ophthalmology and Visual Science*, *37*. 34

Royer, F. L. (1981). Detection of symmetry. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(6), 1186–1210. 47, 57, 87

Saarinen, J., & Levi, D. M. (2000). Perception of mirror symmetry reveals long-range interactions between orientation-selective cortical filters. *Neuroreport*, *11*, 21332138. 51

Sasaki, Y., Vanduffel, W., Knutsen, T., Tyler, C., & Tootell, R. (2005). Symmetry activates extrastriate visual cortex in human and nonhuman primates. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(8), 3159–3163. 50, 51

Schill, K., Umkehrer, E., Beinlich, S., Krieger, G., & Zetzsche, C. (2001). Scene analysis with saccadic eye movements: top-down and bottom-up modeling. *Journal of Electronic Imaging*, *10*, 152–160.
URL `http://adsabs.harvard.edu/abs/2001JEI....10..152S` 40

Schmid, C., & Mohr, R. (1997). Local greyvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(5), 530–535. 121, 147

Scholl, B. J. (2001). Objects and attention: the state of the art. *Cognition*, *80*(1-2), 1–46. 433NB Times Cited:184 Cited References Count:184. 49, 105

Schumann, F., Einhäuser-Treyer, W., Vockeroth, J., Bartl, K., Schneider, E., & König, P. (2008). Salient features in gaze-aligned recordings of human visual input during free exploration of natural environments. *Journal of Vision*, *8*(14), 1–17. 82

Se, S., Lowe, D. G., & Little, J. (2002). Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. *International Journal of Robotics Research*, *21*(8), 735–758. 134, 169, 176, 189

Sela, G., & Levine, M. D. (1997). Real-time attention for robotic vision. *Real-Time Imaging*, *3*, 173–194. 129

Shapiro, L. G., & Stockman, G. C. (2003). *computer vision*. Prentice Hall. 120

Sim, R., Elinas, P., & Little, J. J. (2007). A study of the rao-blackwellised particle filter for efficient and accurate vision-based SLAM. *International Journal of Computer Vision/International Journal of Robotics Research Special Joint Issue on Vision in Robotics*, *74*(3), 303–318. 136, 137, 178

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception*, *28*, 1059–1074. 27

Sobel, E. C. (1990). The locust's use of motion parallax to measure distance. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*, *167*(5), 1432–1351. 140

Stark, L. W., & Privitera, C. M. (1997). Top-down and bottom-up image processing. *Neural Networks*, *4*(9-12), 2294–2299. 39

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*(14), 1–17. 69

Theeuwes, J. (1991). Exogenous and endogenous control of attention: The effect of visual onsets and offsets. *Perception & Psychophysics*, *49*(1), 83–90. 28

Theeuwes, J. (1992). Perceptual selectivity for color and form. *Perception & Psychophysics*, *51*(6), 599–606. 28, 34

Theeuwes, J. (1994). Stimulus-drive capture and attentional set: Selective search for color and visual abrupt onsets. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(4), 799–806. 28

Theeuwes, J., & Kooi, J. L. (1994). Parallel search for a conjunction of shape and contrast polarity. *Vision Research*, *34*(22), 3013–3016. 37

Thornhill, R., & Gangestad, S. W. (1993). Human facial beauty: Averageness, symmetry and parasite resistance. *Human Nature*, *4*, 237–269. 46

Thrun, S., Burgard, W., & Fox, D. (2005). *Probabilistic Robotics*. Cambridge, Massachusetts: The MIT Press. 133, 137, 169, 265

Thrun, S., Liu, Y., Koller, D., Ng, A. Y., Ghahramani, Z., & Durrant-Whyte, H. (2004). Simultaneous localization and mapping with sparse extended information filters. *The International Journal of Robotics Research*, *23*(7-8), 693–716. 138, 178

Torralba, A., Oliva, A., Castelhano, S., Monica, & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search. *Psychological Review*, *113*, 766–786. 41, 217

Townsend, J. T. (1990). Serial and parallel processing: Sometimes they look like tweedledum and tweedledee but they can (and should) be distinguished. *Psychological Science*, *1*, 46–54. 37

Treisman, A. M. (1985). Preattentive processing in vision. *Computer Vision, Graphics, and Image Processing*, *31*(2), 156–177. 31

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136. 31, 33, 35, 38, 56, 247

Treisman, A. M., & Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review*, *95*, 15–48. 34, 88, 95, 101

Treisman, A. M., & Sato, S. (1990). Conjunction search revisited. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(3), 459–478. 37, 38

Treue, S. (2003). Visual attention: the where, what, how and why of saliency. *Current Opinion in Neurobiology*, *13*(428-432), 428–432. 28

Tsotsos, J. K. (1990). Analyzing vision at the complexity level. *Behavioral and Brain Sciences*, *13*, 423469. 27

Tsotsos, J. K. (1997). Limited capacity of any realizable perceptual system is a sufficient reason for attentive behavior. *Consciousness and Cognition*, *6*(2-3), 429–436. 25

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, *3*(1), 7186. 120

Tuytelaars, T., & Mikolajczyk, K. (2008). Local invariant feature detectors: A survey. *Foundations and Trends in Computer Graphics and Vision*, *3*(3), 177–280. 124

Tuytelaars, T., & Van Gool, L. (2004). Matching widely separated views based on affine invariant regions. *International Journal of Computer Vision*, *59*(1), 61–85. 124

Tyler, C. W. (2000). The human expression of symmetry: Art and neuroscience. In A. Bogdan (Ed.) *ICUS Symmetry Symposium*. Seoul. 46

Van Hoof, H. (2008). *Using Different Methods to Direct a Robot's Attention*. Master's thesis, University of Groningen. 149

Van Maanen, L. (2009). *Context Effects On Memory Retrieval: Theory And Applications*. Ph.D. thesis, University of Groningen. 122

Van Zoest, W., & Donk, M. (2004). Bottom-up and top-down control in visual search. *Perception*, *33*, 927–937. 28

Van Zoest, W., Donk, M., & Theeuwes, J. (2004). The role of stimulus-driven and goal-driven control in saccadic visual selection. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(4), 746–759. 28

Varela, F., Thompson, E. T., & Rosch, E. (1991). *The Embodied Mind*. The MIT Press. 11

Wagemans, J. (1993). Skewed symmetry: a nonaccidental property used to perceive visual forms. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(2), 364–380. 82

Wagemans, J. (1995). Detection of visual symmetries. *Spatial Vision*, *9*(1), 9–32. 47

Wagemans, J. (1997). Characteristics and models of human symmetry detection. *Trends in Cognitive Sciences*, *1*, 346–352. 47

Wagemans, J. (1999). Parallel visual processes in symmetry perception: Normality and pathology. *Documenta Ophthalmologica*, *95*, 359–370. 47, 57, 87

Wagemans, J., Van Gool, L., & d'Ydewalle, G. (1991). Detection of symmetry in tachistoscopically presented dot patterns: Effects of multiple axes and skewing. *Perception and Psychophysics*, *50*(5), 413–427. 87

Walter, W. G. (1950). An imitation of life. *Scientific American*, *182*(5), 42–45. 11

Walter, W. G. (1951). A machine that learns. *Scientific American*, *185*(2), 60–63. 11

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*, *19*, 1395–1407. 249

Wang, D., Kristjansson, A., & Nakayama, K. (2005). Efficient visual search without top-down or bottom-up guidance. *Perception & Psychophysics*, *67*(2), 239–253. 42

Wedema, D. (2009). *Comparing the EKF and FastSLAM solutions to the problem of Simultaneous Localization and Mapping.*. Master's thesis, University of Groningen. 139

Wertheimer (1945). *Productive Thinking*. New York: Harper & Row. 108

Wertheimer, M. (1923). Untersuchungen zur lehre von der gestalt ii. *Psychologische Forschung*, *4*, 301–350. Translation published in Ellis, W. (1938). A source book of Gestalt psychology (pp. 71-88). London: Routledge & Kegan Paul. 108

Winn, J., Criminisi, A., & Minka, T. (2005). Object categorization by learned universal visual dictionary. In *Tenth IEEE International Conference on Computer Vision*, vol. 2, (pp. 1800–1807). Beijing, China. 126

Wolfe, J. M. (1992). "effortless" texture segmentation and "parallel" visual search are not the same thing. *Vision Research*, *32*, 757–763. 37

Wolfe, J. M. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, *1*(2), 202–238. 33, 34, 36, 38, 39, 41

Wolfe, J. M. (1996). Extending guided search: Why guided search needs a preattentive "item map". In . G. D. L. A. Kramer, G. H. Cole (Ed.) *Converging operations in the study of visual selective attention*, (pp. 247–270). Washington, DC: American Psychological Association. 105

Wolfe, J. M. (1998). Visual search. In H. Pashler (Ed.) *Attention*. University College London Press. 33, 37, 38, 88, 95

Wolfe, J. M. (2001). Asymmetries in visual search: an introduction. *Percept Psychophys*, *63*(3), 381–389. 88

Wolfe, J. M. (2007). Guided search 4.0: Current progress with a model of visual search. In W. Gray (Ed.) *Integrated Models of Cognitive Systems*, (pp. 99–119). New York: Oxford. 36, 38, 41, 217

Wolfe, J. M., Butcher, S. J., Lee, C., & Hyle, M. (2003). Changing your mind: On the contributions of top-down and bottom-up guidance in visual search for feature singletons. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(2), 483–502. 28

Wolfe, J. M., & Friedman-Hill, S. R. (1992). On the role of symmetry in visual search. *Psychological Science*, *3*(3), 194–198. 34, 36, 88

Wolfe, L., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *15*, 419–433. 37, 38

Yarbus, A. (1967). *Eye Movements and Vision*. New York: Plenum Press. 27

Yeshurun, Y., Kimchi, R., Sha'shoua, G., & Carmel, T. (2009). Perceptual objects capture attention. *Vision Research*, *49*, 1329–1335. 49, 105

Zelinsky, G. J., Zhang, W., Yu, B., Chen, X., & Samaras, D. (2006). The role of top-down and bottom-up processes in guiding eye movements during visual search. In Y. Weiss, B. Schlkopf, & J. Platt (Eds.) *Advances in Neural Information Processing Systems (NIPS)*, vol. 18, (pp. 1569–1576). Cambridge, MA: MIT Press. 40, 217

Zwinderman, M., Rybski, P. E., & Kootstra, G. (submitted). A human-assisted approach for a mobile robot to learn 3D object models using active vision. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. Submitted. 149

Appendices

APPENDIX A

The Contrast-Saliency Model

Figure A.1 gives an overview of the contrast-saliency model of Itti et al. (1998); Itti & Koch (2001). The model calculates the saliency of an image on the basis of contrast in three different feature channels: intensity, color, and orientation. The model is based on an biologically-plausible architecture developed by Koch & Ullman (1985). It is an implementation of the Feature-Integration Theory of human visual search (Treisman & Gelade, 1980). The model correctly predicts human behavior in pop-out experiments such as discussed in Section 2.3 (Itti & Koch, 2000).

The contrast-saliency model consists of two parts. In the first part, a saliency map is calculated by finding contrast in the image. The saliency map locates the conspicuous parts of the image, that are potentially interesting to pay attention to. In the second part, fixation points are generated based on the saliency map, using a winner-takes-all mechanism to select the most salient location, followed by an inhibition-of-return mechanism to prevent focussing on the same region twice. Since in the study discussed in Chapter 3 we do not utilize the fixation-generation part, but only the calculation of the contrast-saliency map, we only discuss that element of the model.

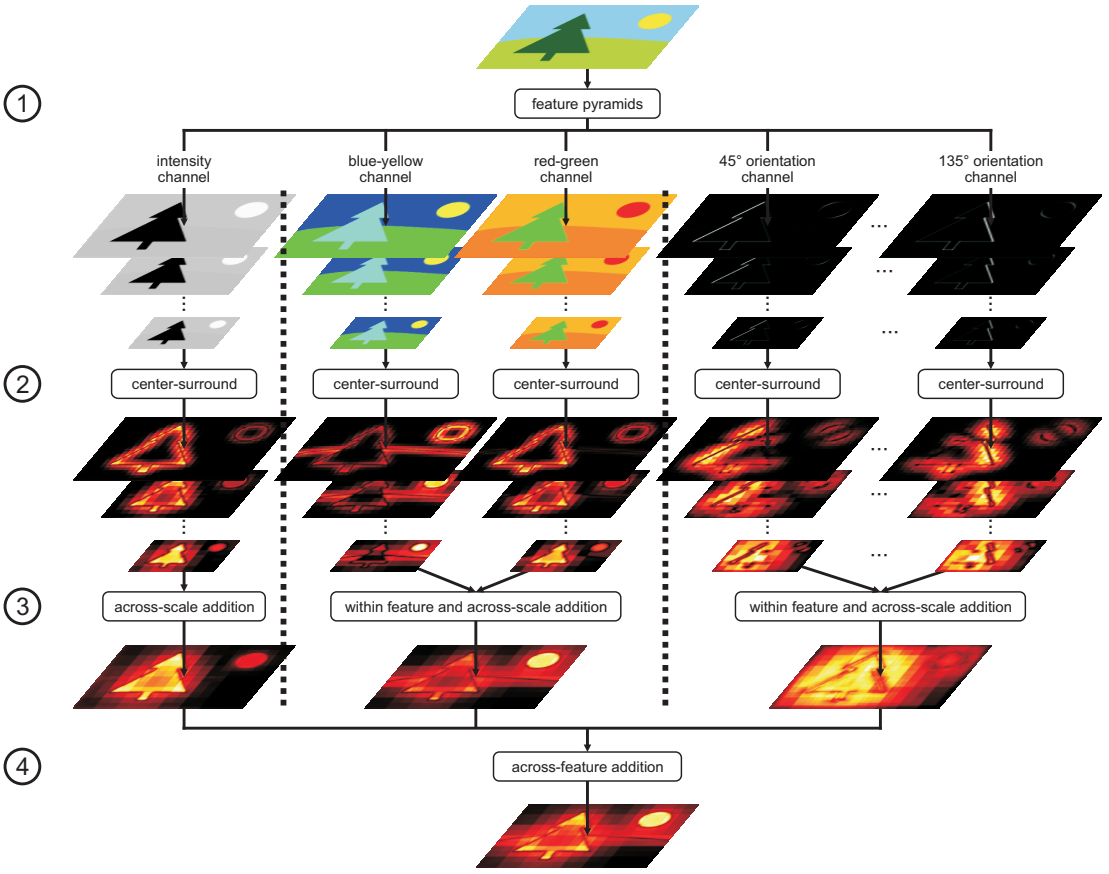

Figure A.1: The contrast-saliency model of Itti et al. (1998). In step 1, The original image is split into different feature channels: one for intensity, two for color (a red-green and a blue-yellow channel), and four for orientation (0°, 45°, 90°, and 135°). Furthermore, for each of the channels, a Gaussian pyramid is built. In step 2, the center-surround calculations are made to get feature maps at different scales. The different scales and channels within a feature are normalized and added together to form the conspicuity maps in step 3. Finally, in step 4, the three conspicuity maps, intensity, color, and orientation, are normalized and summed to result in the saliency map.

A.1 Calculating the saliency map

The contrast-saliency map is calculated based on contrasts in intensity, color, and orientation. Contrast is determined using *center-surround differences*. This is biologically motivated by human behavior in pop-out experiments, where the contrast of an object with its surroundings seems to determine the conspicuity of that object. A white square stands out among surrounding black squares, as does a red square among green squares, and a tilted bar among vertical bars. Center-surround cells are furthermore widely present in early visual processing in the human brain. What follows is a description of the saliency model as implemented on the basis of (Itti et al., 1998) with a few modifications, mainly corrections of some errors in the equations used in that paper. The important processing steps of the contrast-saliency model are shown in Figure A.1.

A.1.1 Step 1:

The original image is split into intensity and chromatic maps. The intensity map I is calculated by:

$$I = (R+G+B)/3, \tag{A.1}$$

where R, G, and B are respectively the red, green, and blue channel in the original image. The chromatic maps are constructed for the color opponents red-green, RG, and blue-yellow, BY:

$$RG(x,y) = \frac{R(x,y)-G(x,y)}{\max(R(x,y),\ G(x,y),\ B(x,y))} \tag{A.2}$$

$$BY(x,y) = \frac{B(x,y)-\min(R(x,y),\ G(x,y))}{\max(R(x,y),\ G(x,y),\ B(x,y))}, \tag{A.3}$$

where max and min are operators that return respectively the maximum and minimum value of their arguments. This method to calculate the opponent-color channels is different from that described in (Itti et al., 1998), but used in the current implementation of the model at Laurent Itti's iLab at the University of Southern California (Walther & Koch, 2006)[1].

[1] The Saliency Toolbox can be downloaded from: http://www.saliencytoolbox.net

To be able to detect contrast on different scales, image pyramids of the different features are constructed. A Gaussian intensity pyramid, I_l, is created from the intensity map at its original scale I_0, where $l \in [0 \dots 8]$. At subsequent scales, the map is first convolved with a Gaussian kernel, G, for low-pass filtering, and then downsampled to obtain a map that is half the width and height of the previous scale:

$$I'_{l-1} = I_{l-1} * G \tag{A.4}$$

$$I_l(x,y) = I'_{l-1}(2x,2y). \tag{A.5}$$

The opponent-color Gaussian pyramids, RG_l and BY_l, where $l \in [0 \dots 8]$, are constructed similar to the intensity pyramid.

The Oriented Gabor pyramids,$O_{l,\theta}$ are finally constructed by convolving I_l with oriented Gabor filters, $\mathcal{G}_\theta$:

$$O_{l,\theta} = I_l * \mathcal{G}_\theta, \tag{A.6}$$

where $l \in [0 \dots 8]$ and $\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$.

A.1.2 Step 2:

Once the image pyramids are constructed, the center-surround contrasts are calculated at different scales by taking the difference between a feature map at a fine scale $c \in \{2,3,4\}$ and at a coarser scale $s = c + \delta$, $\delta \in \{3,4\}$. This is done by using the across-scale difference operator, $\ominus$, which first interpolates the coarser scale to the finer scale and then subtracts the two maps point by point. δ gives the distance in scale between the center and surround. The higher this value, the larger the surrounding area. Thus, the intensity center-surround feature maps are calculated:

$$\mathcal{I}_{c,s} = \|I_c - I_s\|. \tag{A.7}$$

By applying the absolute operator $\|\|$, both dark-center-light-surround and light-center-dark-surround are assigned high saliency values.

Chromatic double-opponency feature maps are calculated by subtracting the red-green opponency at the center from the red-green opponency at the surround and similar for

blue-yellow:

$$\mathcal{RG}_{c,s} = \|RG_c - RG_s\| \quad \text{(A.8)}$$

$$\mathcal{BY}_{c,s} = \|BY_c - BY_s\|. \quad \text{(A.9)}$$

Again, by taking the absolute value, both a red-green center and green-red surround and the reverse are considered salient.

Similarly, orientation-contrast feature maps are calculated:

$$\mathcal{O}_{c,s,\theta} = \|O_{c,\theta} - O_{s,\theta}\|. \quad \text{(A.10)}$$

A.1.3 Step 3:

The above results in feature-contrast maps on different scales, with different sizes of the surrounding area, and for the color and orientation channel, with different subfeature channels. To achieve across-scale conspicuity maps of the three different features, the feature maps are normalized and added. For the intensity conspicuity map, that is:

$$\bar{\mathcal{I}} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N(\mathcal{I}_{c,s}), \quad \text{(A.11)}$$

where $\oplus$ is the summation operator that first resizes all elements to the same scale, and then sums the maps pixelwise. The normalization operator N is used to promote feature maps that have only a few outstanding points and demote maps that contain many similarly salient points. N is:

1. Normalize the values in the map to the range [0,1].
2. Multiply all values in the map with $(1-\bar{m})^2$, where $\bar{m}$ is the average value of all local maxima in the map that have a value greater than or equal to 0.10.

If there are many similarly symmetrical patterns, $\bar{m}$ will be large, and the map will thus be multiplied by a small value. If, on the other hand, there is one clear global maximum, $\bar{m}$ will be small, and the map will be promoted. By applying this normalization in Equation A.11, scales with only a few salient points will be emphasized. The normalization is one of the important elements that make the model detect singleton pop outs.

Another normalization procedure based on lateral inhibition is discuss in (Itti & Koch, 2000). However, in the experience performed for this thesis, that procedure resulted in sparse saliency maps with too few salient locations. Although that method is interesting for predicting eye fixations in a visual-search task, it is less useful for predicting eye fixations in a free-view experiment with complex photographic stimuli. Since the stimuli are complex and the participants could make multiple eye fixations, there are many potentially interesting locations to focus on in the image. The normalization method described above compares better with the human behavior.

The color conspicuity map is calculated by:

$$\bar{\mathcal{C}} = N\left(\bigoplus_{c=2}^{4}\bigoplus_{s=c+3}^{c+4} N(\mathcal{RG}_{c,s})\right) + N\left(\bigoplus_{c=2}^{4}\bigoplus_{s=c+3}^{c+4} N(\mathcal{BY}_{c,s})\right) \tag{A.12}$$

and the orientation conspicuity map:

$$\bar{\mathcal{O}} = \sum_{\theta \in Theta} N\left(\bigoplus_{c=2}^{4}\bigoplus_{s=c+3}^{c+4} N(\mathcal{O}_{c,s,\theta})\right). \tag{A.13}$$

A.1.4 Step 4:

Finally, the contrast-saliency map is obtained by normalizing and summing the three conspicuity maps:

$$S^{\text{con}} = \frac{1}{3}\left(N(\bar{\mathcal{I}}) + N(\bar{\mathcal{C}}) + N(\bar{\mathcal{O}})\right). \tag{A.14}$$

APPENDIX B

The Scale-Invariant Feature Transform

The Scale-Invariant Feature Transform (SIFT) is a method for detecting and describing interest points in images. These interest points are largely invariant to changes in scale and rotation. The features are furthermore shown to be robust for other distortions such as affine transformations, small changes in viewpoint, and change in illumination (Lowe, 2004). The detected interest points are represented by a 128-dimensional feature vector. The discriptor space on the one hand gives room for the description of many unique interest points, while on the other hand, different observations of an interest point can be reliably matched.

SIFT consists of two distinct parts: the detector and the descriptor. The SIFT detector is based upon difference-of-Gaussian calculations, very similar to the center-surround implementation in the contrast-saliency model described in Appendix A. The SIFT descriptor utilizes histograms of oriented gradients to describe the local neighborhood of the interest points. The description of both parts below is based on (Lowe, 2004).

B.1 The SIFT interest-point detector

The interest-point detector consists of three steps. Firstly, interest points are proposed by detecting extrema in scale space using a Gaussian-image pyramid and difference-of-Gaussian calculations. Secondly, the proposed interest points are accurately localized in the image. Finally, points that lie on edges or have low contrast are eliminated.

B.1.1 Detecting scale-space extrema

To detect points in the image that can be reliable found under different views and at different scales, stable features across the image and across scales are found. This is done by building a scale-space using Gaussian-image pyramids (see Figure B.1). The stable points are those that are extrema in both local space and local scale in the image convolved by different-of-Gaussian functions. These difference-of-Gaussian images, $D(x,y,\sigma)$, are efficiently computed as the difference of two Gaussian images, $L(x,y,\sigma)$, at nearby scales:

$$D(x,y,\sigma) = L(x,y,k\sigma) - L(x,y,\sigma) \tag{B.1}$$

$$L(x,y,\sigma) = G(x,y,\sigma) * I(x,y), \tag{B.2}$$

where σ is the current scale, k is a constant multiplicative scale factor, I is the gray-scaled input image, $*$ is the convolution operator, and G is a Gaussian kernel with standard deviation σ:

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} \cdot e^{-(x^2+y^2)/2\sigma^2} \tag{B.3}$$

A point is proposed as interest point if it is a local extremum, that is, a local minimum or maximum, in its $3 \times 3 \times 3$ neighborhood in scale space (see Figure B.1).

The size of the Gaussian kernel, G, can be limited to $6\sigma \times 6\sigma$. The values outside that area are close to zero and have virtually no influence on the outcome of the convolution. Moreover, the convolution with the two-dimensional kernel can be efficiently performed since it is separable in two iterative convolutions with a single-dimensional Gaussian kernel, first in the horizontal and then in the vertical direction. However,

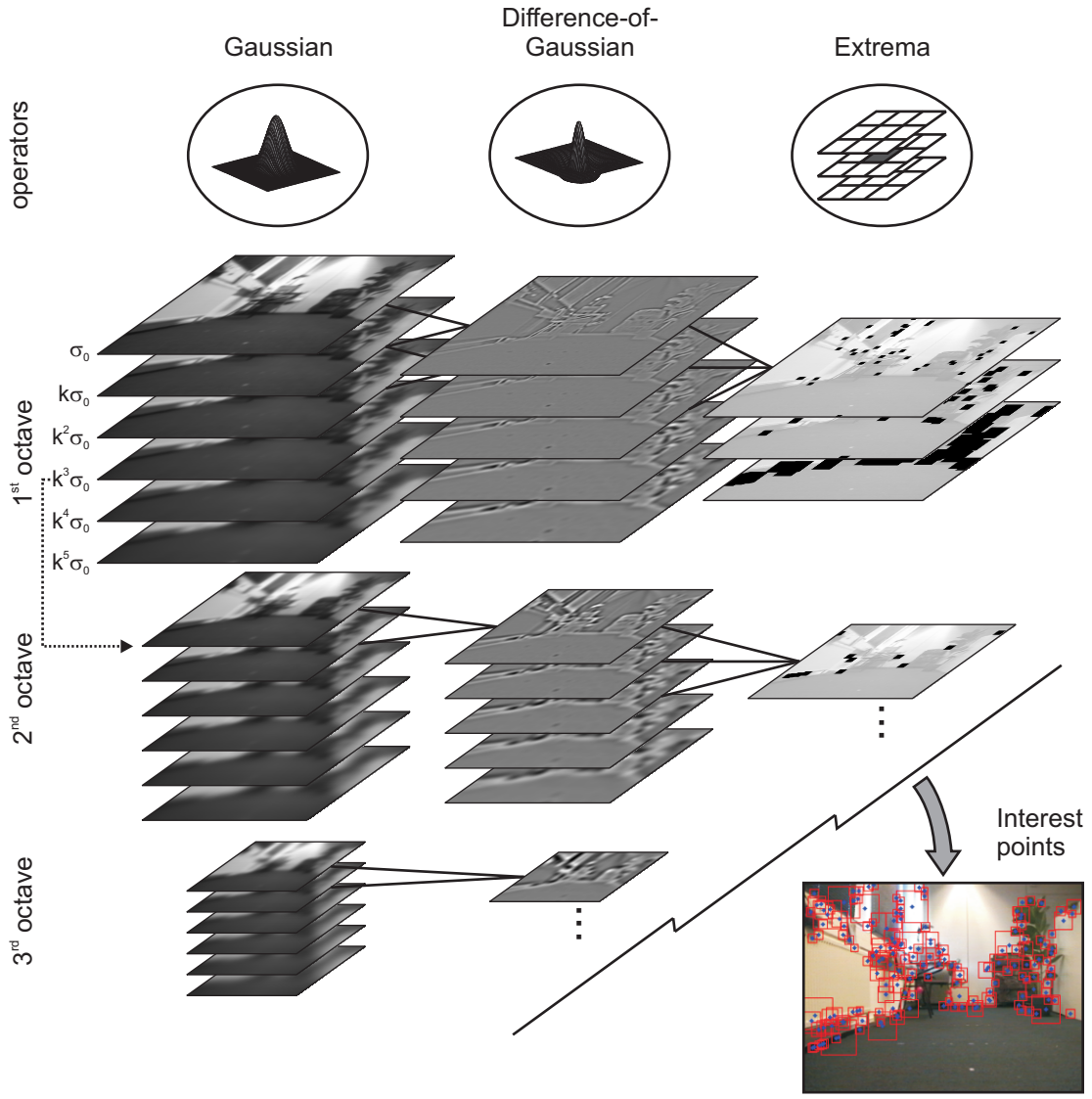

Figure B.1: The SIFT detector. The input image is transformed into a Gaussian pyramid by progressive Gaussian filtering. At subsequent octaves, the image is down scaled by a factor two. Contrast in the image is calculated by subtracting two Gaussian images at adjacent scales. This difference of Gaussians (DOG) is an approximation of the second derivative of the image. Local extrema are found in the DoG images. After accurately locating the extrema and filtering the weak points and those lying on an edge, the location and scale of the interest points in the image are detected.

since σ is constantly increased by a factor k, the kernel size grows. For reasons of efficiency, the image is therefore down-sampled by a factor two when σ is twice its original value, σ_0. This separates the scale space into different octaves. Instead of convolving the original image, I_0, with a large kernel, $G(x,y,2\sigma)$, the down-sampled image, I_1, is convolved with $G(x,y,\sigma)$. This results in a Gaussian scale space that consists of Gaussian-filtered images, I_l, at different octaves l. Within an octave, the images are convolved with Gaussian kernels $G(x,y,\sigma)$ at different scales $\{\sigma, k\sigma, k^2\sigma, \ldots\}$ (see Figure B.1).

To obtain interest points at s different scales per octave, there need to be $s+2$ different scales of difference-of-Gaussian images, since the extrema are determined using a lower and higher scale. Therefore, $s+3$ different scales of Gaussian images are needed. To have evenly distributed scales of interest points over the octaves, $k = 2^{1/s}$. Throughout the dissertation, we use $s = 3$ and three different octaves.

The scale-space extrema detection results in a number of proposed interest points. In subsequent steps of SIFT, the locations of these points are established more accurately. Next, unstable points are eliminated.

B.1.2 Accurate localization of interest points

The locations of the proposed interest points are rough estimations. Especially on higher levels of the scale-space pyramid, it is important to more accurately estimate the position, since there the resolution is low due to subsequent downsampling. The location in scale-space is more accurately estimated by fitting a three-dimensional quadratic function to the local sample point. The location of the extrema can then be found by setting the derivative of the function to zero. This is done using the Taylor expansion up to the quadratic term of $D(x,y,\sigma)$, shifted so that the origin is at the proposed interest point $\mathbf{x} = (x,y,\sigma)^T$:

$$D(\mathbf{x}) = D + \frac{\partial D}{\partial \mathbf{x}}^T \mathbf{x} + \frac{1}{2}\mathbf{x}^T \frac{\partial^2 D}{\partial \mathbf{x}^2}\mathbf{x}, \tag{B.4}$$

where D and its first and second derivative are evaluated at the interest point. The re-estimated location, $\hat{\mathbf{x}}$, of the interest point is determined by setting the derivative of $D(\mathbf{x})$ with respect to $\mathbf{x}$ to zero. This gives:

$$\hat{\mathbf{x}} = -\frac{\partial^2 D}{\partial \mathbf{x}^2}^{-1} \frac{\partial D}{\partial \mathbf{x}}. \tag{B.5}$$

The Hessian $\frac{\partial^2 D}{\partial \mathbf{x}^2}$ and derivative $\frac{\partial D}{\partial \mathbf{x}}$ can be efficiently approximated using differences of neighboring sample points.

B.1.3 Elimination of unstable interest points

In the last step of the SIFT interest-point detector, the proposed interest points are further investigated, in order to eliminate unstable points. Points that have a weak contrast are disregarded in order to avoid noisy detections. Also points lying on an edge are removed, because these are less unique, since all points on the edge appear similar.

The contrast of an interest point is taken as the difference-of-Gaussian value at its re-estimated location:

$$D(\hat{\mathbf{x}}) = D + \frac{1}{2} \frac{\partial D}{\partial \mathbf{x}}^T \hat{\mathbf{x}}. \tag{B.6}$$

All extrema with $|D(\hat{\mathbf{x}})| < \tau_c$ are rejected as interest point. We used $\tau_c = 0.02$ in our experiments.

Finally, points that lie on an edge are disregarded as well, since these points are not uniquely identifiable on the edge. Therefore, only points that lie on a peak or an inversed peak in D should be selected as interest points. This can be tested by calculating the principal curvatures of D in spatial scale using the 2×2 Hessian matrix, $\mathbf{H}$, computed at the location of the point (x, y, σ):

$$\mathbf{H} = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix}, \tag{B.7}$$

where the derivatives D_{aa} is the partial second derivative of D with respect to a, and D_{ab} is the mixed partial second derivative with respect to a and b. These can be estimated by taking differences of neighboring sample points:

$$D_{xx} = D(x-1, y) + D(x+1, y) - 2 \cdot D(x, y) \tag{B.8}$$

$$D_{xy} = \frac{1}{4}\left(D(x-1,y-1)+D(x+1,y+1)- D(x-1,y+1)-D(x+1,y-1)\right) \tag{B.9}$$

To check if the proposed interest point is on a peak, the two eigenvalues of $\mathbf{H}$, which are proportional to the principal curvature of D, should be similar, that is, the ratio between the highest eigenvalue and the second should be close to 1.0. This ratio can be efficiently calculated using the determinant and the trace of the $\mathbf{H}$. The check then boils down to:

$$\frac{\mathrm{Tr}(\mathbf{H})^2}{\mathrm{Det}(\mathbf{H})} < \frac{(\tau_r+1)^2}{\tau_r}, \tag{B.10}$$

where τ_r is the threshold on the ratio between the two principle curvatures. We used $\tau_r = 10$ in the experiments presented in this dissertation.

B.2 The SIFT interest-point descriptor

The SIFT detector reliably and repeatable detects interest points in the image. To be able to store and recognize the detected points, the SIFT descriptor computes a description that is invariant to rotations, and robust to small changes in viewpoint and change in illumination. The same descriptor is used in the MUlti-scale Symmetry Transform (MUST) discussed in Chapter 8. Also the Symmetrical Region-of-Interest Detector (SymRoID), discussed in Chapter 9, uses a descriptor that is largely based on the SIFT descriptor. The computation consists of two parts. Firstly, the orientation of each interest point is determined, so that the descriptor can be calculated relative to this orientation. Secondly, the descriptor is created using histograms of gradient orientations.

B.2.1 Orientation assignment

The orientation of the interest point is determined by the dominant gradient orientation in the local neighborhood of the interest point. To calculate this, a histogram of all gradient orientations in the neighborhood patch is created. The magnitude, $m(x,y)$,

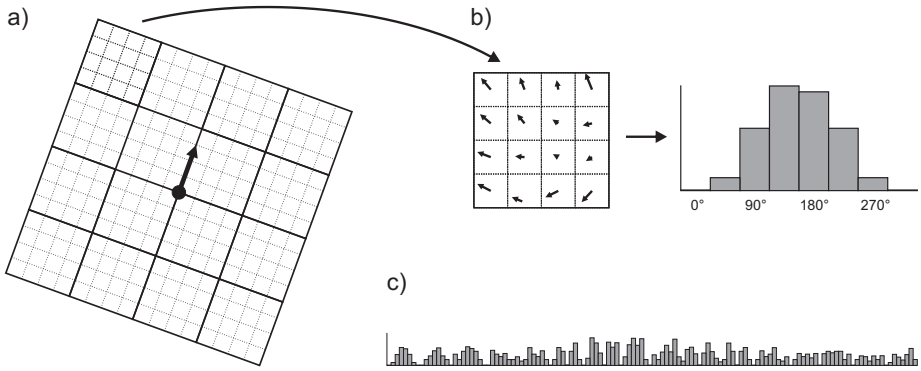

Figure B.2: The SIFT descriptor. a) The local neighborhood is determined based on the orientation and the scale of the interest point. The neighborhood patch is split into 4×4 subregions. b) Each subregion consists of 4×4 sample points. At each sample point, the intensity gradient is determined. The 16 gradient samples fill the histogram of gradient orientations. The histogram contains 8 bins. c) The 8 bins of all 16 subregions result in a 128-dimensional feature vector.

and orientation, $\theta(x,y)$ of the gradient at sample point (x,y) is determined by:

$$m(x,y) = \sqrt{(L(x+1,y)-L(x-1,y))^2+(L(x,y+1)-L(x,y-1))^2} \quad \text{(B.11)}$$

$$\theta(x,y) = \tan^{-1}\left(\frac{L(x,y+1)-L(x,y-1)}{L(x+1,y)-L(x-1,y)}\right), \quad \text{(B.12)}$$

where L is the Gaussian image at the scale closest to the scale, σ_i, of the interest point. The histogram consists of 36 bins, covering the 360° range of orientation. Each sample point contributes to the histogram proportional to its magnitude and weighted by a Gaussian function on the distance from the interest-point location. This distance function has a standard deviation of $1.5 \cdot \sigma_i$. Subsequently, the bin with the highest value plus all bins with a value $> 80\%$ of the highest value are selected as dominant orientations. In case that there are multiple dominant orientations, an interest point is created for all dominant orientations. Finally, the dominant orientations are more accurately estimated using interpolation by fitting a parabola to the 3 histogram values closest to each dominant orientation.

B.2.2 Descriptor representation

The local neighborhood of every interest points is described using histograms of gradient orientations. The neighborhood is split up into 4×4 subregions (see Figure B.2a). This way, important spatial information is preserved. A histogram of gradient orientation is calculated for each subregion based on the 4×4 sample points in that subregion. The locations of the sample points are determined based on the scale and orientation of the interest point (see Figure B.2a) to gain rotation invariance. Moreover, the gradients are rotated over the dominant orientation of the interest point. All 16 gradients in the subregion contribute to the histogram, which contains 8 bins (see Figure B.2b) The contribution is proportional to the magnitudes of the gradients weighted by a Gaussian function on the distance towards the location of the interest point. The standard variation of this function is 1.5 times the size of the neighborhood.

To promote robustness to small shifts in location and orientation, every gradient also contributes to the adjacent bins in the histogram with a weight inversely proportional to the angular distance between the gradient orientation and the bin center. Furthermore, every gradient also contributes to the histograms of adjacent subregions with a weight inversely proportional to the Euclidean distance between the sample point and the center of the subregion.

This results in 4×4 histograms containing 8 bins, and therefore to a feature vector of size 128 (see Figure B.2b). To increase the robustness of the descriptor to changes in illumination, the feature vector is normalized in three steps. Firstly, the vector is normalized to unit length. Then, all bins with a value higher than 0.2 are set to 0.2, to decrease the influence of large magnitudes, possibly due to camera saturation. Finally, this saturated vector is normalized again to unit length.

APPENDIX C

The Extended Kalman Filter

The standard Kalman filter assumes linearity in the motion and observation model. These assumptions, however, are usually not satisfied in robotic SLAM. The *extended* Kalman filter solves this problem by linearizing the motion model and the observation model by computing their Jacobians or partial derivatives with respect to position.

The motion model of the robot in EKF-SLAM is defined as:

$$P(\mathbf{x}_k|\mathbf{x}_{k-1},\mathbf{u}_k) \Leftrightarrow \mathbf{x}_k = \mathbf{f}(\mathbf{x}_{k-1},\mathbf{u}_k) + \mathbf{w}_k, \tag{C.1}$$

where $\mathbf{f}(\cdot)$ is the motion estimation based on the action of the robot (estimated by its odometry) $\mathbf{u}_k$ at timestep k and the previous estimate of the robot's position $\mathbf{x}_{k-1}$, and $\mathbf{w}_k$ is zero mean, uncorrelated Gaussian motion noise with covariance $\mathbf{U}_k$.

The observation model is:

$$P(\mathbf{z}_k|\mathbf{x}_k,\mathbf{m}) \Leftrightarrow \mathbf{z}_k = \mathbf{h}(\mathbf{x}_k,\mathbf{m}) + \mathbf{v}_k, \tag{C.2}$$

where $\mathbf{h}(\cdot)$ is a function that predicts the observation based on the current pose of the

robot and the map $\mathbf{m}$. The map consists of the positions of all landmarks. $\mathbf{v}_k$ is zero mean, uncorrelated Gaussian observation noise with covariance $\mathbf{R}_k$.

The standard EKF method (Maybeck, 1979) is used to compute the mean and the covariance of the joint posterior distribution $P(\mathbf{x}_k, \mathbf{m} | \mathbf{Z}_{0:k}, \mathbf{U}_{0:k}, \mathbf{x}_0)$. The mean, or *state vector*, $\mathbf{s}_k$ at timestep k is:

$$\mathbf{s}_k = \begin{bmatrix} \mathbf{x}_{k|k} \\ \mathbf{m}_k^T \end{bmatrix} = \begin{bmatrix} \mathbf{x}_{k|k} \\ \mathbf{m}_{1k} \\ \cdots \\ \mathbf{m}_{Nk} \end{bmatrix}, \tag{C.3}$$

where $\mathbf{x}_{k|k}$ is the current estimation of the robot's position and $\mathbf{m}_{1k}, \ldots, \mathbf{m}_{Nk}$ are the estimated positions of the N landmarks in the map at time k. $\mathbf{x}_{k|k}$ and $\mathbf{m}_{ik}$ are vectors of length 2. The covariance matrix $\mathbf{P}_{k|k}$ at timestep k is:

$$\mathbf{P}_{k|k} = \begin{bmatrix} \mathbf{P}_{xx} & \mathbf{P}_{xm} \\ \mathbf{P}_{xm}^T & \mathbf{P}_{mm} \end{bmatrix}_{k|k} = \begin{bmatrix} \mathbf{P}_{xx} & \mathbf{P}_{xm_1} & \cdots & \mathbf{P}_{xm_n} \\ \mathbf{P}_{m_1x}^T & \mathbf{P}_{m_1m_1} & \cdots & \mathbf{P}_{m_1m_n} \\ \cdots & \cdots & \cdots & \ddots \quad \cdots \\ \mathbf{P}_{m_nx}^T & \mathbf{P}_{m_nm_1} & \cdots & \mathbf{P}_{m_nm_n} \end{bmatrix}_{k|k}, \tag{C.4}$$

where $\mathbf{P}_{xx}$ is the 3×3 covariance matrix of the position and orientation of the robot $\mathbf{x}$, $\mathbf{P}_{mm}$ is the $2N \times 2N$ covariance matrix of the positions of all N landmarks in the map, and $\mathbf{P}_{xm}$ is the $3 \times 2N$ covariance of the position of the robot with respect to the position of all landmarks in the map. $\mathbf{P}_{m_im_i}$ is the individual 2×2 covariance matrix for landmark i, and the $\mathbf{P}_{xm_i}$ matrices are 3×2.

The EKF is used to estimate the state vector and its covariance matrix based on the previous position of the robot, the robot's action, the robot's observation and the current map of the environment. This is iteratively done by applying the prediction and the update step. The notation of the position estimation are as followed. At the previous timestep $k-1$, the estimation of the robot's position is $\mathbf{x}_{k-1|k-1}$. After the prediction step the estimation is adjusted by using the current action of the robot and becomes $\mathbf{x}_{k|k-1}$. When also the current observation is applied in the update step, the current estimate becomes $\mathbf{x}_{k|k}$.

C.1 Prediction step

In the prediction step, only the location and orientation of the robot need to be updated, assuming that the landmarks are stationary. This results in:

$$\mathbf{s}_{k|k-1} = \begin{bmatrix} \mathbf{f}(\mathbf{x}_{k-1|k-1}, \mathbf{u}_k) \\ \mathbf{m}_k^T \end{bmatrix} \tag{C.5}$$

$$\mathbf{P}_{k|k-1} = \begin{bmatrix} \nabla\mathbf{f}_x \, \mathbf{P}_{xx,k-1|k-1} \nabla\mathbf{f}_x^T + \mathbf{Q}_k & \nabla\mathbf{f}_x \, \mathbf{P}_{xm,k-1|k-1} \\ \mathbf{P}_{xm,k-1|k-1}^T \nabla\mathbf{f}_x^T & P_{mm,k-1|k-1} \end{bmatrix} \tag{C.6}$$

$$\mathbf{Q}_k = \nabla\mathbf{f}_u \, \mathbf{U}_k \nabla\mathbf{f}_u^T \tag{C.7}$$

where $\mathbf{Q}_k$ is the covariance characterizing the uncertainty in position and orientation of the robot and $\mathbf{U}_k$ that of the distance travelled and change made by the robot. $\nabla\mathbf{f}_x$ and $\nabla\mathbf{f}_u$ are the Jacobians of $\mathbf{f}(\mathbf{x}_{k-1|k-1}, \mathbf{u}_k)$ evaluated at respectively $\mathbf{x}_{k-1|k-1}$ and $\mathbf{u}_k$:

$$\nabla\mathbf{f}_x = \frac{\partial \mathbf{f}}{\partial \mathbf{x}_{k-1|k-1}} = \begin{bmatrix} 1 & 0 & -d_k \sin(\theta_{k-1} + \varphi_k) \\ 0 & 1 & d_k \cos(\theta_{k-1} + \varphi_k) \\ 0 & 0 & 1 \end{bmatrix} \tag{C.8}$$

$$\nabla\mathbf{f}_u = \frac{\partial \mathbf{f}}{\partial \mathbf{u}_k} = \begin{bmatrix} \cos(\theta_{k-1} + \varphi_k) & -d_k \sin(\theta_{k-1} + \varphi_k) \\ \sin(\theta_{k-1} + \varphi_k) & d_k \cos(\theta_{k-1} + \varphi_k) \\ 0 & 1 \end{bmatrix}, \tag{C.9}$$

where d_k is the distance travelled and φ_k is the change in orientation made by the robot, as estimated by the odometry.

C.2 Update step

The full state vector and its covariance matrix are updated using the current observations of knwon landmarks by:

$$\begin{bmatrix} \mathbf{x}_{k|k} \\ \mathbf{m}_k^T \end{bmatrix} = \begin{bmatrix} \mathbf{x}_{k|k-1} \\ \mathbf{m}_{k-1} \end{bmatrix} + \mathbf{W}_k [\mathbf{z}_k - \mathbf{h}(\mathbf{x}_{k|k-1}, \mathbf{m}_{k-1})] \tag{C.10}$$

$$\mathbf{P}_{k|k} = \mathbf{P}_{k|k-1} - \mathbf{W}_k \mathbf{S}_k \mathbf{W}_k^T, \tag{C.11}$$

where

$$\mathbf{S}_k = \nabla\mathbf{h}\,\mathbf{P}_{k|k-1}\nabla\mathbf{h}^T + \mathbf{R}_k \tag{C.12}$$

$$\mathbf{W}_k = \mathbf{P}_{k|k-1}\nabla\mathbf{h}^T\mathbf{S}_k^{-1}, \tag{C.13}$$

where $\mathbf{R}$ is covariance of the observation noise and $\nabla\mathbf{h}$ is the Jacobian of $\mathbf{h}$ evaluated at $\mathbf{x}_{k|k-1}$ and $\mathbf{m}_{k-1}$. For reasons of efficiency, the size of the Jacobian varies with the number of observed landmarks, since only these landmarks need to be used to update the state vector and its covariance. There is no need to update the state vector and its covariance matrix for landmarks in the map that are currently not observed. Guivant & Nebot (2001) define the Jacobian for a single observed landmark as:

$$\nabla\mathbf{h} = \begin{bmatrix} \frac{\Delta x}{\Delta} & \frac{\Delta y}{\Delta} & 0 & \cdots & -\frac{\Delta x}{\Delta} & -\frac{\Delta y}{\Delta} & \cdots \\ -\frac{\Delta y}{\Delta^2} & \frac{\Delta x}{\Delta^2} & -1 & \cdots & \frac{\Delta y}{\Delta^2} & -\frac{\Delta x}{\Delta^2} & \cdots \end{bmatrix}, \tag{C.14}$$

where $\Delta x = (x - m_{ix})$ and $\Delta y = (y - m_{iy})$ are the differences between respectively the x- and y-position of the robot and the observed landmark i. $\Delta = \sqrt{(\Delta x)^2 + (\Delta y)^2}$ is the distance between robot and landmark position. The values on $\cdots$ are all zero. Except for the columns that are associated with the robot position and with the observed landmark position, all values in the matrix are zero. When there are multiple landmarks in the map observed, the above calculations are done sequentially for each of the observed landmarks.

C.3 State-augmentation step

The prediction and update step implement the localization part of SLAM. The steps estimate the state vector and its covariance matrix based on the previous position of the robot, its action, its observation, and the map of the environment. To build this map, observations of new landmarks need to be added to the map. This is done in the *state-augmentation step*.

When a new landmark is observed, it is initialized using the pose of the robot and the observation:

$$\mathbf{m}_{n+1} = \mathbf{g}(\mathbf{x}_k, \mathbf{z}_k). \tag{C.15}$$

The augmented state vector becomes:

$$\mathbf{s}_k^+ = \begin{bmatrix} \mathbf{x}_k \\ \mathbf{m} \\ \mathbf{g}(\mathbf{x}_k, \mathbf{z}_k) \end{bmatrix} \tag{C.16}$$

and the augmented covariance matrix is extended as:

$$\mathbf{P}_k^+ = \begin{bmatrix} \mathbf{P}_{xx} & \mathbf{P}_{xm} & \mathbf{P}_{rr}\nabla\mathbf{g}_\mathbf{x}^T \\ \mathbf{P}_{xm}^T & \mathbf{P}_{mm} & \mathbf{P}_{xm}^T\nabla\mathbf{g}_\mathbf{x}^T \\ \nabla\mathbf{g}_\mathbf{x}\mathbf{P}_{xx} & \nabla\mathbf{g}_\mathbf{x}\mathbf{P}_{xm} & \nabla\mathbf{g}_\mathbf{x}\mathbf{P}_{xx}\nabla\mathbf{g}_\mathbf{x}^T + \nabla\mathbf{g}_{\mathbf{m}_{n+1}}\mathbf{R}_k\nabla\mathbf{g}_{\mathbf{m}_{n+1}}^T \end{bmatrix}, \tag{C.17}$$

where

$$\nabla\mathbf{g}_\mathbf{x} = \begin{bmatrix} 1 & 0 & -r\sin\beta \\ 0 & 1 & r\cos\beta \end{bmatrix} \quad \nabla\mathbf{g}_{\mathbf{m}_{n+1}} = \begin{bmatrix} \cos\beta & -r\sin\beta \\ \sin\beta & r\cos\beta \end{bmatrix}, \tag{C.18}$$

where r is the distance and β is the absolute bearing towards the new observation.

In summary, the EKF-SLAM builds a map of the environment, and localizes the robot using the map in three steps. In the update step, the robot updates its own position based on the odometric information. In the update step, the robot updates its own position and the positions of landmarks based on observations of known landmarks. Finally, whenever the robot encounters a new landmark, this landmark is added to the map in the augmentation step. This results in an estimation of the robot's position and orientation, $\mathbf{x}$, with covariance $\mathbf{P}_{xx}$, and a map of the environment including all observed landmarks, $\mathbf{m}$, with covariance $\mathbf{P}_{mm}$. When the robot re-observes a known landmark, the uncertainty of the robot and the landmark location decreases. Based on the correlations between landmarks in the covariance matrix $\mathbf{P}_{mm}$, also the positions of all other landmarks become more certain.

For more information on EKF-SLAM, we refer to (Thrun et al., 2005; Durrant-Whyte & Bailey, 2006; Bailey & Durrant-Whyte, 2006) and to (de Jong, 2008) for our implementation.

Dankwoord

Jaren ben ik bezig geweest met het onderzoek dat in dit proefschrift staat geschreven en maanden om het allemaal netjes op te schrijven. Nu is alles af, op het dankwoord na. En laat dat nu juist het meest gelezen deel van elk proefschrift zijn. Iedereen zal meteen vol interesse bladeren naar het dankwoord om te kijken of hij of zij ook wordt genoemd. Ondanks dat het einde voor mij in zicht is, rust er dus nog een zware taak op mijn schouders. Het is bijna een onmogelijke opgave om in een paar pagina's uit te drukken wat iedereen voor me heeft betekend en om niemand te vergeten. Als dat niet is gelukt, mijn excuses.

Allereerst wil ik graag mijn ouders hartelijk bedanken. Lies en Michiel, zonder jullie had ik dit proefschrift uiteraard nooit geschreven, maar de geweldige liefde, zorg en ondersteuning die ik altijd van jullie heb mogen ontvangen op alle facetten van mijn leven is niet iets vanzelfsprekend. Ik wil jullie daar enorm voor bedanken.

Uiteraard wil ik mijn copromotor, Bart de Boer, en mijn promotor, Lambert Schomaker, van harte bedanken voor alle hulp en het feit dat ze mij hebben gevormd tot wetenschapper. Bart, we hebben elkaar leren kennen toen ik nog docent bij Kunstmatige Intelligentie was. Vanaf het begin klikte het meteen. Ik wil je heel erg bedanken voor alle inspirerende gesprekken over mijn onderzoek en over de wetenschap in het algemeen. Ik heb veel van je geleerd en heb grote bewondering voor je. Lambert, ook jou wil ik erg bedanken voor alle ondersteuning. Het is mooi om een hoogleraar te treffen die zo vol passie is over zijn onderzoek en elke gelegenheid aangrijpt om ook zelf de handen uit de mouwen te steken. Ik vond het inspirerend om met je samen te werken.

Mijn paranimfen, Maria en Leendert, wil ik niet alleen enorm bedanken voor de ondersteuning in de voorbereidingen voor de promotiedag, maar ook voor de geweldige vriendschap en collegialiteit.

Daarnaast heb ik veel steun gehad van alle collegas in de afgelopen vier jaar. Ik wil alle collegas bij KI bedanken voor de inspirerende persoonlijkheden die jullie zijn en de interessante gesprekken, ontspannende lunches, en geweldige borrels die we samen hebben gehad. De promovendi: Anja, Axel, Bea, Ben, Dirkjan, Elske, Hedde, Jacolien, Jelmer, Karin, Leendert, Maria, Renante en Tijn, en het overige wetenschappelijk en ondersteunend personeel: Arnold, Bart, Chris, Edith, Elina, Esther, Fokie, Frank, Hanneke, Hedderik, Ingrid, Jennifer, Joep, Jolie, Marco, Mariëtte, Margriet, Marius, Nancy, Niels, Rineke, Ronald, Ronald, Sietse, Sietse, Sjoerd, Sonja, Sujata en Tjeerd. Ook promovendi bij BCN wil ik bedanken voor de gezellige en interessante retraites: Cornelis, Joanne, Jojanneke, Juha, Marleen en Sjouke. Ook alle studenten Kunstmatige

Intelligentie bedankt, met name Arco, Edwin, Jelmer, Matthijs, Paul en Sjoerd, die hun steentje hebben bijgedragen aan het onderzoek dat in dit proefschrift staat beschreven.

Natuurlijk wil ik ook alle niet-werk-gerelateerde vrienden bedanken. Marianne, ik heb erg veel gehad aan alle steun en motivatie die je me door de jaren hebt gegeven. Jacob, bedankt voor de relaxte weekendjes Antwerpen, met goede gesprekken onder genot van nog betere Belgische biertjes. Erik, je enthousiasme over alles, van sport tot politiek, is een grote inspiratie en dank dat ik altijd voor alle vluchten naar conferenties bij jou en Suzanne kon overnachten. Harm, dank je voor de muziek, maar ook voor het feit dat de deur van jouw en Juul's huis voor me open stond in moeilijke tijden. Als de werkdruk af en toe te hoog was, waren er gelukkig altijd de (winterse) weekendjes weg met Erik, Jelle, Jonnes, Harm, Pieter, Richard, Robert, Roelof, Rutger en Wieger om te ontladen. Ik ben eindelijk afgestudeerd. Ook Alma, Arnold, Carien, Catherine, Chris, Egbert, Else, Fiona, Freek, Gemma, Guus, Hanneke, Japke, Joanne, Kirsten, Loes, Marie, Marsha, Martin, Mintsje, Monja, Nelleke, Paulien, Peter, Reinier, Rick bedankt voor *all the good times*, voor de talloze goede gesprekken, etentjes, wijntjes, films en concerten.

Een gezonde geest in een gezond lichaam, en dat betekende voor mij de afgelopen jaren Ultimate Frisbee bij Gronical Dizziness. Iedereen enorm bedankt voor de fanatieke en gezellige trainingen en wedstrijden. Ans, Anton, Erik, Jaap, Koen, Leo, Maria, Marije, Marleen, Max, Michel, Rebecca, Rob, Tim, Tonny en Zhen Chih, ik vond het geweldig om samen met jullie de bronzen plak op de Europese Kampioenschappen te halen. Janneke, Joost, Marije en Zhen Chih, het was een goede ervaring om met jullie in het bestuur te hebben gezeten en Abele, Douwe, Erik en Max, ik hoop dat de Martinicup uitgroeit tot een begrip.

Klaas, Tineke, Titia, Geertje, David, Martin, Marcel, Geke, Iris, Daniëlle, Jordy, Maikel, Marijn en Corné, bedankt voor jullie onvoorwaardelijke steun en liefde als familie.

Het was een mooie en leerzame tijd.

The multi-disciplinary approach taken in this dissertation has led to new insights in visual attention and active vision in natural and artificial systems. Vision, instead of being passive, involves active processes to focus attention. This is not only true for natural systems, but it is also important for artificial vision systems. This dissertation deals with visual attention in natural and artificial systems and proposes symmetry, one of the Gestalt principles for figure-ground segregation, as an important feature.

In the first part of the dissertation, the perception of symmetry by the human visual system is studied in the context of overt visual attention. We propose a visual attention model based on symmetry. The results show that human eye fixations are predicted better by our model than by a model using center-surround contrasts of basic features such as brightness, color, and orientation. The results furthermore indicate that symmetry is detected efficiently by humans, despite being a higher-level visual feature.

In the second part, the proposed visual attention model is applied to focus the attention of a robot on interesting parts in its environment. The use of symmetry is shown to be beneficial for the selection of stable and robust visual landmarks. Based on these landmarks, the robot builds a map of the environment and uses this map to localize itself. In this context, the use of symmetry outperforms the landmark selection method based on center-surround contrasts. It is furthermore illustrated that perception is simplified using active vision.

The main conclusion of this dissertation is that symmetry is a valuable feature both for the prediction of human gaze and for focusing the attention of an autonomous robot.