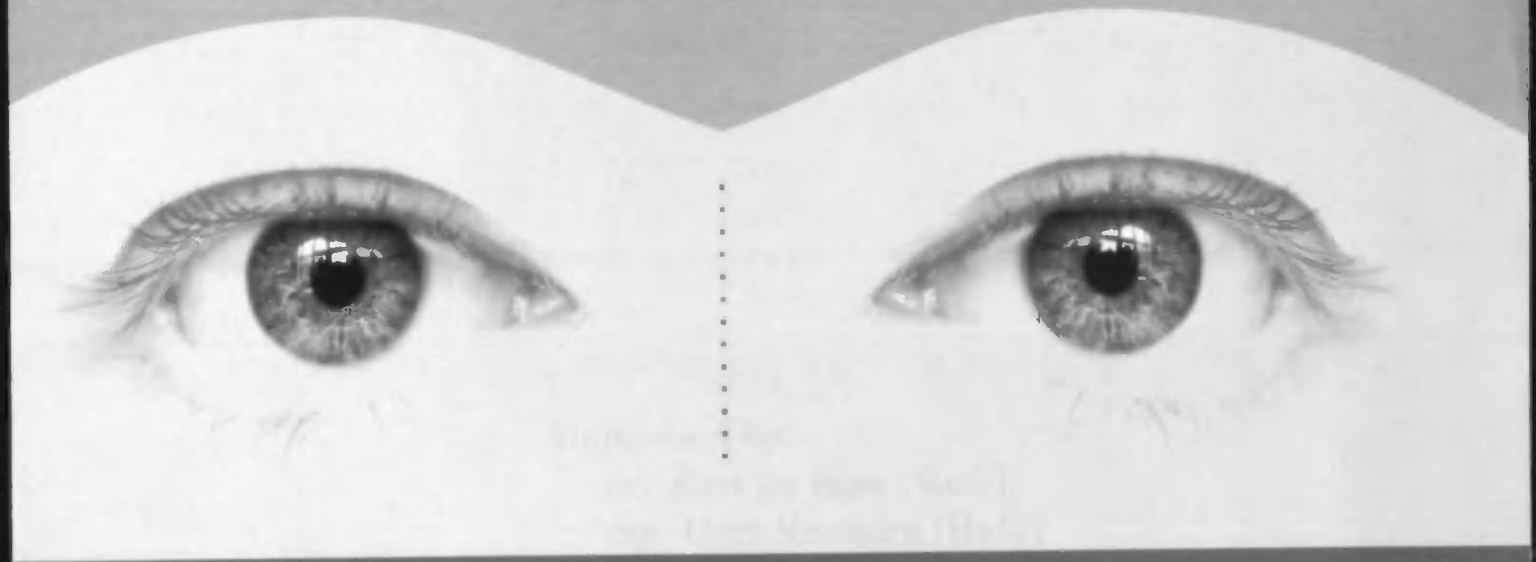
Predicting Fixations with Computational Algorithms



Arco Nederveen

Predicting Fixations with Computational Algorithms

Arco Nederveen

stud.nr.: 0899526

August 2007

Supervised by:
dr. Bart de Boer (R*u*G)
drs. Gert Kootstra (R*u*G)

Artificial Intelligence
University of Groningen

For my parents,
Jan and Ria Nederveen.

Acknowledgements

First and foremost I would like to thank my mother and father, Ria and Jan, for their patience and support. Without their help I would never have been able to finish my education.

Furthermore, I would like to thank my family and friends for encouraging me to stick with it.

Last but not least, I would like to thank my supervisors Bart de Boer and especially Gert Kootstra. Gert was always prepared to answer my questions and he guided me through making this thesis.

Abstract

Our eyes make several movements per second. When, for example, reading this line of text, our eyes constantly move to different parts of the sentence. These eye movements or saccades are interleaved by fixations. Fixations are periods of about 200 ms in which the eye has a relative fixed position, which serves to center our fovea, the most acute part of our retina, on the object of interest. Because the fovea covers only 2 degrees of our visual field, we only select a small part of a scene at once.

We are interested which part of scene is selected and to what extend we can predict those parts by using computational algorithms. We conducted an eye tracker experiment to obtain data from human participants and compared the data to several algorithms. Particular attention was given to the algorithms based on symmetry. We found that the performance of the algorithms based on symmetry compares favorable to other tested algorithms such as, among others, the saliency model by Itti, Koch, and Niebur (1998), a well known computational bottom-up model of visual attention.

Contents

Chapter 1

Introduction

How does a human create a coherent semantic view of the world? It is impossible to take all the information that is available in our surroundings into consideration. That is why we limit our self to a subset of all available information. The first selection is the consequence of the limited capacity of our senses. For example, our ears cannot hear sound frequencies above 20 kHz and our eyes are only sensitive to a limited band of the electromagnetic spectrum. After this initial filter we use specific strategies to get the best out of the available information. We selectively pick information which we consider helpful in understanding the current situation. We cannot quickly comprehend a scene by examining every little detail of that scene. We must make choices which information we consider relevant. We do this by using eye movements as an active filter whereby we process only small parts of the scene at once. This leads to the question: what information is relevant to comprehend a scene? How do we determine which information to use?

Of more specific interest to Artificial Intelligence and especially robotics is to not only understand the process of the selection of information, but also to look if this process can be used as an example to solve similar problems on artificial systems. An example of such a task is the recognition of objects by a robot. Can we learn from the human visual process to help us implement an artificial system? To gain insight on this topic we want to compare data obtained from humans to different algorithms and models.

Vision is an important part of human information processing as can be deduced by the large part of our brain which is involved with processing visual information. It is estimated that 60% of the brain receives visual information. Although this sounds like a very impressive amount of resources, we are still not able to comprehend a given scene or to recognize an object instantaneously. Yet from subjective experience the act of seeing seems a continuous stare of, and immediate comprehension of the current object of

interest, but this is an illusion. On closer inspection the eye makes several movements per second. When, for example, reading this line of text, our eyes constantly move to different parts of a sentence. Eye movements or saccades are interleaved by fixations. Fixations are periods of about 200 ms in which the eye has a relative fixed position. A fixation is to center the fovea on the part of the scene we are directing our attention to. This is useful because the fovea is the center of the retina with the highest density of photo receptors. But it only covers about 2° of the retina. Therefore, to obtain the most detailed information about a part of a scene, we have to move the fovea to cover that area.

We define eye movements as be the movement of overt visual attention. Meaning that we consider a fixated part of the image is also the object of attention. It is therefore reasonable to call parts of the scene that are fixated *regions of interest* or ROIs. We assume ROIs are selected because they contain discriminating or unique features compared to their surroundings Reinagle and Zador (1999) Krieger, Rentschler, Hauske, Schill, and Zetzsche (2000). Noton and Stark (1971) proposed in their *scanpath theory* that humans combine a set of ROIs into one mental model of a certain object or scene, and will loop over this set while that object or scene is subjected to their attention. But what defines regions of interest and how do we select them?

The process which selects information can be seen as consisting of two antagonistic parts. Firstly, a bottom-up process which is a fast and task independent. This process uses low level information which can be extracted from a scene without prior assumptions about that scene or in other words bottom-up processes are stimulus driven. Secondly, a slower top-down process which is task dependent and uses priori knowledge, such as previous experiences, about a scene to select information. As top-down processes are by definition influenced by world knowledge of the individual, top-down processes are much harder to model than bottom-up processes which only depend on the stimuli perceived in the current situation and although top-down processes play a important role in visual attention a significant part can be explained with bottom-up processes. Theeuwes (2004) shows that top down strategy to search a certain shape can be overridden by a uniquely colored distractor. Thereby showing that bottom-up influences are able to override and grab attention away from top-down strategies. Therefore, in this thesis, we are mainly interested in the possibility of predicting eye movements with a bottom-up process. In addition, we eventually want to implement a visual selection process on a artificial system. At startup, such a system, a robot for example, will not possess knowledge about the situation it is in. Therefore, the system initially can only use bottom-up information, making the

bottom-up process indispensable for artificial systems. So we have to wonder to what extend fixations can be predicted without using information about the content of the scene. This leads us to the following research question:

> Which low level properties are most suitable to predict the locations of human fixations?

This thesis investigates the predictability of these ROIs. The predictive power of several computational models will be evaluated. No classification of the content of the image is attempted. We do not, for example, separate figure-ground before processing the image, neither do we categorize the content, in order to create a top down model. Only bottom-up information will be used as saliency predictor. We expect that bottom up features should be capable of explaining a significant part of or the fixations.

The models will take an image as input and will calculate a saliency map. A saliency map indicates how likely it is that a certain part of the image will be fixated. To produce these saliency maps, several different algorithms will be used. One of the models is the *saliency model* by Itti et al. (1998) which combines several biologically plausible low level descriptors into one saliency map. A second algorithm is the SIFT algorithm by Lowe (2004). This algorithm is originally used to extract unique and stable properties from an photographed object and uses these properties to recognize the same object in another image. Stable means that the same property can still be extracted from the image even if the target object is subject to rotation, scaling or is looked at from a different angle. The locations of these properties are therefore highly informative. Since SIFT is proven to be successful in artificial object recognition. It is interesting to compare its interest point detection with human eye movements. Furthermore, other algorithms such as entropy and Michaelson contrast as published by Privitera and Stark (2000) will be investigated.

Special attention will be given to a measure of symmetry proposed by Reisfeld, Wolfson, and Yeshurun (1995) and a modification of the algorithm by Heidemann (2004). Symmetry seems to catch the immediate attention of humans and is regarded as an aesthetic property (Lochner & Nodine, 1989). And a preference for symmetry is not only found in humans. Lehrer (1999) found that bees have innate preference for symmetry.

To be able to compare the performance of the different models and algorithms we have collected data from human subjects. The subjects looked at pictures while their eye movements were recorded by an eye tracker. Thereby recording the gaze of at the subject. These recorded fixations have to be compared to the saliency maps generated by the models. To this end we adopted

the methods to compare the data found in papers by Parkhurst, Law, and Niebur (2002) and by Ouerhani et al. (2004).

One of the contributions of the research presented in this thesis is to learn us something about the human visual system. For example, the model described by Itti et al. (1998) is inspired by knowledge of physiological structures found in the human visual system and by theories about the information processing performed by these structures. The performance of such an algorithm could therefore also tell us something about the correctness of the assumptions about the human physiology and about visual information processing in humans. But well performing algorithms with no apparent biological bases might of course also give us insight what humans regard as interesting.

Besides telling us something about the human visual system the results can also be used for Artificial intelligence. The eye fixates on information rich parts of the scene. Parts from a scene without much information, like uniform areas, are rarely targeted by fixations. By selecting several small regions to understand the content of the entire scene a significant reduction in the amount data which has to processed is realized. This would give us an algorithm we could use to select and therefore reduce the information needed for scene and object recognition. This is especially useful in the field of robotics where processing power is always a bottleneck. A well performing model/algorithm could therefore be a valuable contribution to Artificial Intelligence.

What follows is an outline of the organization of the thesis. In chapter 2 we discus the theoretical background. We provide an overview of the human visual system and we tell something about symmetry. Finally, we will discus research which also deals with comparing algorithms with the human eye movements. In chapter 3 we provide detailed description of the algorithms used to construct saliency maps. In chapter 4 we provide a more detailed examination of several algorithms related to symmetry. In chapter 5 we explain the experiments we conducted and explain the methods used to analyze the data. In chapter 6 we will present and the results of the analysis. Finally, in chapter 7 we provide a discussion about the results end how our findings relate to existing research.

Chapter 2

Theoretical background

In this chapter we will provide an overview of the human visual system. Furthermore we will discuss symmetry and finally we will review research relating to comparing eye movements with algorithms.

2.1 The human visual system

This section will be dedicated to the human visual system. We will first present a general view of looking at the stages of visual perception. Subsequently we will discuss the physiology of the human visual system. Then we will discuss the spatial frequency theory and finally we discuss eye movement and visual attention.

2.1.1 Stages in visual perception

One can divide the visual perception in four different stages: image-based, surface-based, object-based and category-based (*Photons to Phenomology*, 1999). The first stage, the image-based stage, are filters that perform operations on the retinal 2D pixel like representation of the image. This includes operations like edge detection, line detection, blob detection and correlating the binocular images. The surface based stage utilizes information from the image-based stage to deduce 3D properties of the scene. Properties such as the tilt and slant of surfaces. The third stage, the object-based stage, uses the 3D information to identify and reconstruct objects. By using 3D information, occluded parts of objects can be deduced from the 3D hints. Picture a mug standing on a table. If watching the scene from a slightly elevated position the mug can occlude the edge of the table. Using the surface based information from table top, such as the texture and the edges, one can infer

it is likely that the table and its edge extend behind the mug. Once all 3D separable objects are identified, information processing can proceed to next stage, the category-based stage. In this stage all available extra information associated with an object is processed. This is called category based because the information related to an object is most likely stored in categories. So once an object is identified, relating information is retreived and a mental picture of the object becomes available.

These stages can be used as a general framework to understand and describe the computational visual process. The algorithms used in this thesis will be in the realm of the image-based stage. Almost all algorithms are inspired by the properties of cells and found on the path from the retina to the visual cortex.

2.1.2 Physiology of the human visual system

The human visual system is one of the most complex sensory systems. This system gives us detailed information about the surrounding world by detecting photons with light sensitive cells called photo receptors. The incoming stream of photons is focused on the retina. The retina is a layered tissue of nerve cells which converts the photons into electric signals. The first step is the conversion of light to electric signals by the rods and the cones. The rods react to signals in the whole visual spectrum and are more sensitive than cones. The cones come in three different variations which react to different specific more narrow frequency bands. The low, medium and high frequency cones are more sensitive to frequencies we interpret as respectively red, green and blue. This enables us to perceive colors. The distribution of the rods and cones across the retina is not uniform. The center of the retina is called the fovea. The fovea covers one to two degrees of the visual field and is exclusively covered with cones at a high density. Outwards there is a rapid decay in cone density and rods become the dominating cell type as can be seen in figure2.1.

Behind the layer of rods and cones one can find the bipolar and horizontal cells. The bipolar cells receive their input directly from rods and cones or indirectly via a horizontal cell. A horizontal cell receives input from several rods or cones. If the direct path to the bipolar cell is excitatory the indirect path will always be inhibitory. See figure 2.2.

Ganglion cells are located behind the bipolar and horizontal cells. They receive multiple inputs from several bipolar cells. The ganglion cell can be modeled as a circle surrounded by a ring. If the center circle exhibits a exitatory reaction to light the ring or annulus will react inhibitory and visa versa. The former is called an on-center cell and the latter an off-center cell.

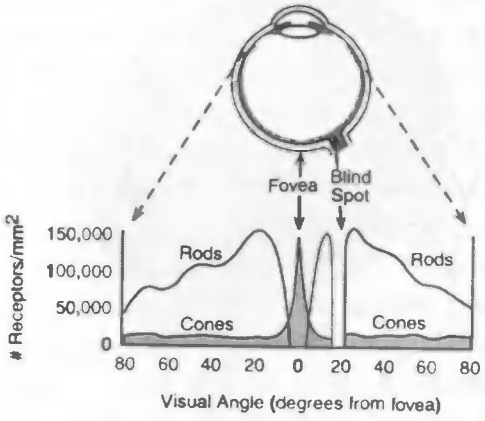


Figure 2.1: Density distribution of rods and cones on the retina. (From *Photons to Phenomology*, 1999)

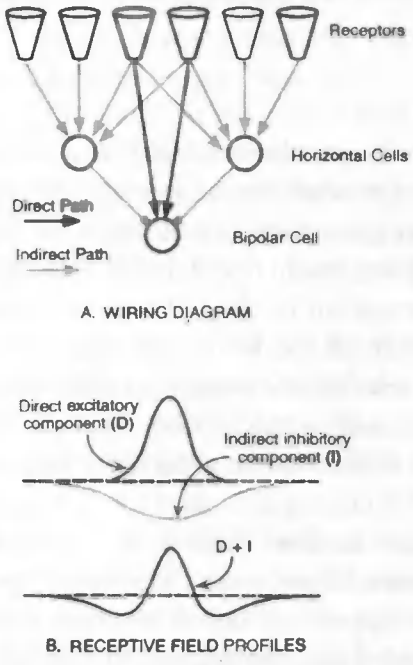


Figure 2.2: Wiring diagram of a bipolar cell. The bipolar receives exitatory signals from the receptors with which it has a direct connection. Receptors who connect indirectly, via a horizontal cell, have an inhibitory effect. This results in an on-center, off-surround bipolar cell. (From *Photons to Phenomology*, 1999)

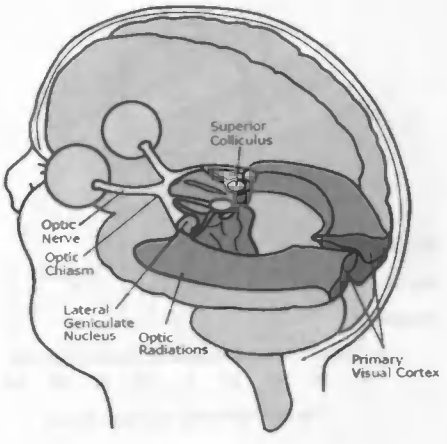


Figure 2.3: The human visual system. After the visual information is processed in the eye, the information is transported to optic chiasm by the optic nerve. From the optic chiasm the information ends up in the LGN which projects on the visual cortex via a bundle of axons called the optic radiations. (From Blind, 2004)

Although all ganglion cells can be modeled as a center surround cell there are several types which are sensitive to specific stimuli as explained later.

The axons of the ganglion cells bundle into the optic nerve and end up in the optic chiasm. Here each nerve bundle from the nasal side of the fovea crosses over. Information in the left side of the visual field is therefore processed in the right side of the brain and visa versa. From there a small nerve bundle makes its way to the superior colliculus, which primarily deals with spatial information and is said to be involved in eye movements. The larger pathway leads us to the lateral geniculate nucleus of the thalamus (see figure 2.3).

The lateral geniculate nucleas (LGN) is a layered or laminar structure formed from six 2D layers of neurons. The lower two layers are called the magnocellar layers. The upper four layers are known as the parvocellar layers. Each LGN receives input from one side of the visual field. The cells in the magnocellar layers receive their input form M ganglion cells which are more sensitive to intensity than to color and play a role in detecting motion. The opposite is true for parvocellar layers which receive their input from the P ganglion cells which are more sensitive to color.

The axons from the LGN project to the striate cortex also known as V1 or primary visual cortex. Hubel and Wiesel were the first to do single cell recordings in this area for which they used a cat. Others had tried before

but were not able to get a response from the cells. Hubel (1959) discovered by accident that the cells reacted only to lines with a certain orientation and direction. Hubel and Wiesel categorized the cells as Simple cells, complex cells, and hypercomplex cells. Simple cells got their name due to the fact that the response to complex stimuli can be predicted by their response to single spots of light. Most simple cells have an elongated receptive fields. Some have their receptive fiels split in half with a inibitory and exitatory half. These cells react to luminance edges with a certain orientation. Others hava a center-surround configuration which makes them sensitive to lines. Complex cells are the most common cells in the striate cortex. The cells receive their input from several simple cells. Complex cells do not react to stationary light spots but are sensitive to moving lines or edges with a proper orientation. Huber and Wiesel thought they had found a third type of cell found in the striate cortex is the hyper complex cell. This cell is more sensitive for lines if they are short in length. This is why they are also called end-stopped cells. It is now beleived that hypercomplex cells do not exist but are end-stopped simple or complex cells.

The striate cortex is two mm thick and has columnar structure. One such a column is called a hypercolumn (figure 2.4). Every column is split into a part for the left and right side of the vision field. These parts are again divided in sections sensitive for different orientations. Sensitivity for different orientations is well supported by single cell recordings and autoradiographic methods. A more disputed claim is the existence of sensitivity of cells to different scales. Cells located deeper in the striate cortex would be more sensitive to larger scales.

Every layer in the striate cortex, except for the first layer, also introduces lateral connections. It is obvious that a lot of signal processing is done, even before the signals are transported through the optical nerve to the brain. And that is just the beginning. Hubel and Wiesel theorized there are four different pathways all with an independent function. They suggest there is a color pathway, a form path, a binocular pathway and a motion pathway. These pathways consist of connection to and between higher level areas of the visual cortex such as V2, V3, V4 and Medial Temporal lobe (MT). Although there is considerable crosstalk between the pathways and the suggested four pathways may be a over simplification, lesions and single cell recording suggest cells react specifically to certain aspects of shape, form or color.

If algorithms discussed in this thesis have a biologically inspired part it will primarily be based on the information processing up to V1. To discuss the intricacies of research on these areas any further would be beyond the scope of this thesis. If algorithms discussed in this thesis have a biologically

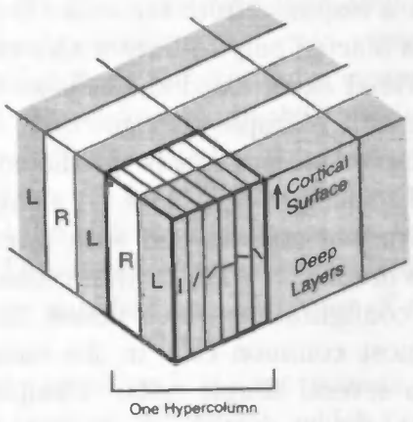


Figure 2.4: Hypercolumns in the striate cortex. Every column is divided into an area which is sensitive to the right eye and an area which is sensitive to the left eye. Each area is subdivided into areas which are sensitive to different orientations. (From *Photons to Phenomology*, 1999)

inspired part it will primarily be based on the information processing up to V1. Mostly because there no clear picture of what is happening in and between the higher visual areas.

2.1.3 Visual attention and eye movements

Eye movements

Eye movements have two functions. The first is to position targets in the fovea to maximize spatial and chromatic information. The second is the tracking of moving objects. The eye movements are influenced from several areas in the brain and not from one specific brain area as one might expect. The muscles controlling the eye movements, the extraocular muscles, are controlled by the oculomotor neurons which stem from the gaze centers in the lower part of the brain called the brain stem. The gaze centers receive input from brain areas throughout the brain such as the superior colliculus, vestibular nuclei, occipital cortex, basal ganglia, and frontal eye fields. (figure 2.5).

We can distinguish different types of eye movements, which are controlled from different areas in the brain. To position the eye at another location the eyes are moved with one fast movement. This movement is called a saccade. A saccade is a ballistic movement, once it is initiated it cannot be altered. It takes about 30ms to execute the movement itself, but when taking planning

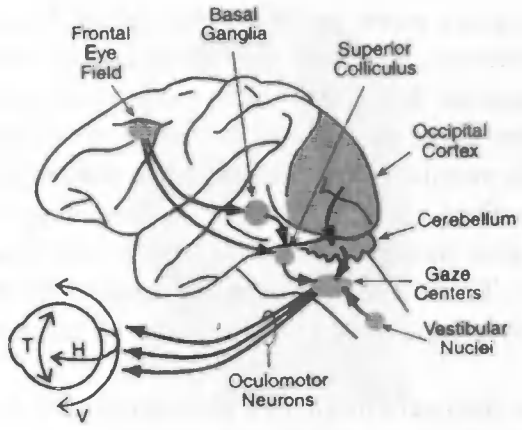


Figure 2.5: Parts of the brain related to the control of eye movements. Different areas of the brain project on the gaze centers in the brain stem, which ultimately control the horizontal (H), vertical (V) and torsional (T) eye movements. (From *Photons to Phenomology*, 1999)

into account, a saccade takes 150 to 200 ms. During the ballistic movement the information coming from the eye is suppressed. Saccadic suppression occures because the information during the saccades are masked by the images before and after the saccade. Voluntary control of the saccades stem from the frontal eye fields situated in the frontal cortex.

A second type of eye movement is the smooth pursuit movement. This movement is used to keep moving targets fixated. Smooth pursuit movement can be distinguished from saccades by their smoothness. Furthermore, the trajectory of smooth pursuits are constantly corrected to keep the fixation on the target. This is not the case for saccades, which are ballistic. Moreover, the speed of the smooth pursuit movement is much lower than the speed of a saccade. Smooth pursuit movements are controlled by the motion pathways in the visual cortex. These include the MT and MST areas. The areas project on the cerebellum and pons, which in turn pass the information to the gaze centers in the brain stem.

Another type of eye movements are the vergence movements. These movements are made to keep track of objects which move towards or away from the observer. If the object comes closer, the eyes will converge and if the object moves further away, the eyes will diverge. This movement is driven by information from the binocular disparity channels of area V2.

Yet another type of eye movements are the vestibular movements. These are the movements made to compensate for movements of the head and body.

Vestibular movements are more accurate than smooth pursuit movements. They are called vestibular, because the brain structures that control eye movement use information from the vestibular system in the inner ear. For this type of movement the oculomotor nucleas receive information from the vestibular nuclei. The vestibular nuclei, which are part of the vestibula nuclei, combine information from the hairs in the semicircular canals of the inner ear. The hairs are able to detect disturbances in the liquid present in the semicircular canals. These disturbances are caused by movements of the head and body, thereby enabling the detection of movements of the head and body.

Finally, there are the optokinetic eye movements. The optokinetic movement is a involuntary tracking movement, which occurs when a large part of the scene is moving uniformly across the retina. Its function is the same as the function of the vestibular eye movements, namely to compensate for movements of the body. The difference is the source of information. Instead of utilizing information from the vestibular system perceived translation of the scene on the retina is used. This information is obtained from the cortical motion pathway and from a subcortical pathway. The result is, that the optokinetic movements are controlled by a combination of visual and vestibular information.

During a fixation the eyes produce three other types of movements which are much smaller than the eye movements which we have just discussed. The first is the tremor. This is fast aperiodic wave-like motion, with a frequency of $\sim$ 90 Hz. The diameter of the movement is the size of about one cone in the fovea. It is thought that tremors have the function to counter the habituation effects of the photoreceptor cells. If an image is projected at exactly the same place on the retina, which can be done under experimental conditions, the subject will no longer perceive the image after several seconds of stabilization. Therefore, the tremors continuously shift the image over the retina, thereby suppressing the habituating effects. The second kind of eye movements during fixations, are the drifts. Drifts take place at the same time as the tremors. These movements are interleaved with microsaccades. During a drift the image can drift across a dozen of photo receptors. It is not really clear if drifts are more than random noise in the oculomotor system. The third kind of movements during fixations are the microsaccades. These are small jerky movements which cover a distance from a dozen to several hundreds of photo receptors. So they cannot be distinguished from real saccades by their size. Instead, a microsaccade is defined as a saccade that is made involuntary. The function of microsaccades has been a topic of debate for 30 years. It is unclear what role they have in maintaining visibility and if this role differs significantly from the role of the tremors and drifts.

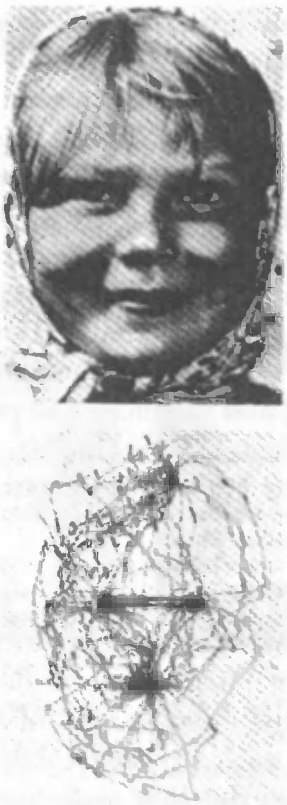


Figure 2.6: Eye movements that were recorded during visual exploration of the young girls face. (From Yarbus, 1968)

For our study saccades are the most interesting of the eye movements. The other movements only serve to keep the current object of interest fixated and are not under voluntary control. Saccades, on the other hand, are used select novel areas of interest and can be voluntary directed to a target. Therefore, saccades give us hints about the selection policy of the brain.

Eye movements have been studied for more than 100 years. For example, late nineteenth century studies of large eye movement where made by observing a reading subjects eye movements with mirrors. Later eye movements could be recorded by reflecting light of a lens which was worn by the subject. The light could be recorded on photographic film and thereby, eye movements could be studied. The Russian scientist Yarbus perfected this technique and developed a method to record eye movements with great accuracy and in a relatively unobtrusive manner. The paper by (Yarbus, 1968) is considered

a seminal paper and is widely cited. In this paper he reviewed all existing methods recording eye movements and presented his findings found using his own method. By superimposing the fixations on the images it was possible to investigate the locations of the fixations, thereby determining what the subjects found the most interesting parts of the images. He also gave different tasks to the subjects while they were viewing an image. For example, one subject was asked to remember the cloths of the people in a scene, while others were asked to estimate the wealth of the people in the scene. This resulted in very different locations of the fixations. Therefore, Yarbus concluded that it is not possible to predict the positions of the locations based on the structure of image alone, and therefore, the positions of the fixations are, at least in part, task-dependent. In other words, fixations are not only stimulus driven or only guided by bottom-up processes, but also by task-driven or top-down processes.

The sequence of fixations from a scene are somehow combined into the uniform representation we as humans seem to experience. It would seem plausible if this transsaccadic integration was done on the basis of location. The saccades would be stored in some kind of memory module which also represented the location of each fixation. Making it possible to integrate them into one representation. But this *spatiotopic fusion* hypothesis did not hold up to scrutiny. Irwin (1992) found that the fixations are not integrated based on their location, but are integrated more based on higher level object information. Irwin found that when he presented a series of letters to a subject it was not the location that was remembered between saccades, but the identity of the letters.

Instead of the idea of overlaying the fixations onto one neural memory buffer "Attention: Contemporary Theory and Analysis" (1970) as cited in *Photons to Phenomology* (1999) proposed a what he called a *schematic map*. The schematic map is a representation of the object you expect to see. So a schematic map of a face could consist of a representation of ears, eyes, mouth, etc. and information about the relative position to each other. To recognize a face, a human would test the schematic map of the face against the current scene. If a person first recognizes the mouth he/she would verify if it really is a face by searching for an ear or an eye by searching in the expected location that is encoded in the schematic map. When recognizing a face, this would give a sequence of fixation which moves from one feature of a face to another. The order of the fixations does not have to be fixed for the same object, but the locations would be typical for a specific object. Noton and Stark (1971) called this sequence of fixations a scanpath.

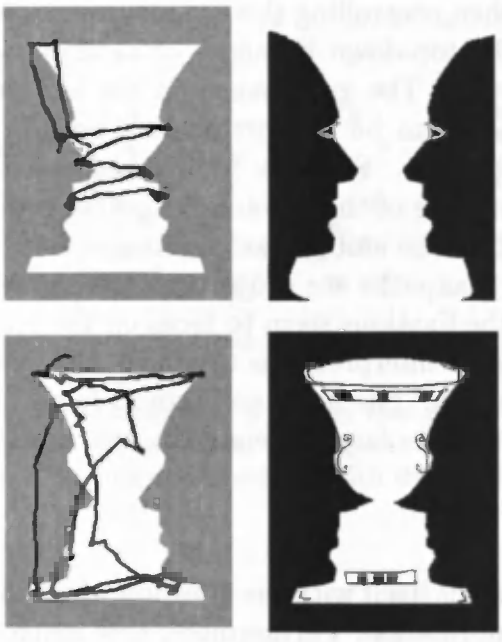


Figure 2.7: Demonstration of top down influence on the eye movements. The gray contour image can be interpreted as a bowl or as the contours of two faces facing each other. Subjects were asked to view this image after they were primed with one of the two images on the right side The resulting eye movements are plotted on the gray images. When the gray contour is interpreted as a bowl the fixations seem to focus on the top and bottom of the bowl, but if the gray contour is interpreted as a face the fixations are directed to items which are typical of a face such as eyes, nose and mouth.(From: Stark et al., 1999)

Scanpath theory

Scanpath theory was proposed by Noton and Stark (1971). They observed in experiments that eye movements have an sequential en repetitive quality to them. They defined a scanpath as:

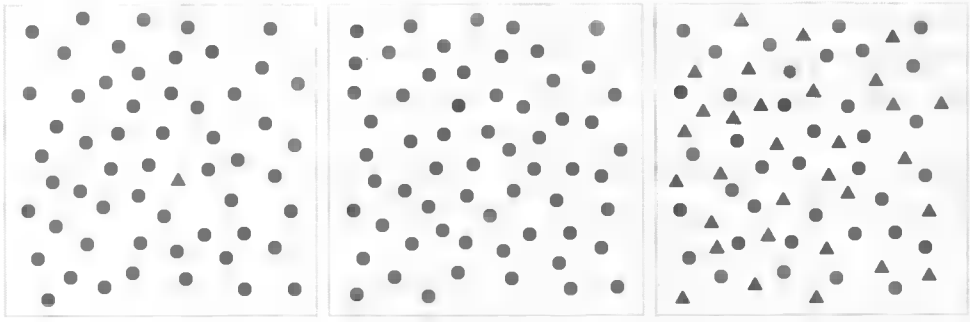
> An idiosyncratic alternation of glimpses (called fixations or foveations and rapid jumps of eye position (called saccades) to various ROIs, in the viewed scene.

Stark further states: "The scanpath theory proposes that an internal spatial-cognitive model controls perception and the active looking eye movements, EMs." (Stark et al., 1999). This means that top-down information,

the model, is used when controlling the eye movements when a such a mental model exists. The top-down influence of a picture on the scanpath is illustrated by figure 2.7. The gray image on the left side is ambiguous to the human observer. It can be interpreted as a bowl or as the contours of two faces facing each other. Subjects were asked to view this image after they were primed with one of the two images on the right side. This caused the subjects to interpret the ambiguous gray image as either the bowl or the faces. The resulting scanpaths are projected on the gray images. When interpreted as a bowl the fixations seem to focus on the top and bottom of the bowl. But if the image is interpreted as two faces, the fixations seem to focus on eyes, nose, and mouth. The fixations seem to focus on the characteristic parts of both objects. For a face these are the eye, nose, and mouth.

Visual attention

Visual attention concerns itself with the question which information we select and how we select information. Furthermore, how much information can we attend to at once, is a topic of debate. A well-known theory about spatial attention is the *spotlight theory* introduced by Posner (1978). This theory states that attention can be seen as a spotlight which illuminates part of the information available. The illuminated part represents the information which is current attention. The spotlight can only be moved by sliding the spot to another position with a certain maximum speed. This predicts that, if you shift your attention to another object, the time to do so would be proportional to the distance between the current object and the target object. This is indeed what is found in experiments (Tsal, 1983). Furthermore, if you move the spotlight from one place to another, the places which are illuminated during the movement should receive attention. Additionally, the spotlight metaphor tells us that it is not possible to split your attention between two places and this was also found in experiments (Eriksen & Yeh, 1985). Another consequence of the spotlight metaphor is that the size of the spot is fixed. Indicating that a human can only attend to a spot of a certain size. This does not correspond well to reality. Although, under some circumstances the size of the spotlight is about one degree, generally the visual angle can be adapted to a larger object or even a whole scene. This is why an adaption of the theory was proposed which is called the *zoom lens theory* . The zoom lens theory (Eriksen & St. James, 1986) introduces the possibility for the region of attention to be adapted in size. It was indeed found that the size of the spotlight can be altered by offering images with different spatial frequencies. Shulman and Wilson (1987) showed that subjects were better at perceiving gratings with a high spatial frequency when offered an image



(a) Shape popout (b) Color popout (c) Conjunction

Figure 2.8: Two examples of a display that could be offered in a pop-out search task based on shape(a) and color(b) and an example of conjunction search in which the red circle is the target.

with the same spatial frequency beforehand. The same relation was found for lower frequencies. This indicates we are indeed capable of changing the size of the zoom lens.

The former theories say something about how attention shifts as a whole. The *feature integration theory* by Treisman and Gelade (1980) tells us something what we do with the information under current consideration. The feature integration theory is a two-stage model of visual processing. The first stage is the construction of so called feature maps from low level visual features. For example a feature map could be the result of edge detection applied to an image. Such a low level property is perceived in parallel over the entire visual field. But these low level properties have to be combined some where in the visual process. According to Treismann and Gelade this happens for part of the visual field if that part is under our current attention, or to say it in other words, illuminated by the spotlight. This second stage integrates the feature maps into one saliency map. The saliency map indicates how salient or noticeable a part of an image is. Treisman investigated what kind of visual information can be processed in parallel over the entire visual field. They proposed information which triggered a pop out effect was processed in parallel. They found that color, orientation shape information could produce a pop-out effect.

One of the predictions of this theory is the difference in performance on certain search tasks. In some search tasks, the target can be located almost instantaneous and the target just seems to *pop out* whereas other tasks require a subject to consciously scan the stimuli. Treisman and Gelade

differentiated these two cases by introducing two kinds of search, namely, feature search and conjunction search. Feature search is a fast, pre-attentive, parallel process of properties which make up the feature maps. Conjunction search is slow, serial and requires overt attention. It is called conjunction because this type of search is initiated when searching for properties which are a conjunction or combination of properties used to contstruct the feature maps.

Treisman and Gelade (1980) confirmed their prediction by offering a search task to several subjects. The task consisted of finding a certain object, the target, between distractors. Examples of possible displays used for such a task are shown in figure 2.8. If the target only differs with respect to one low level property, such as color or orientation, from the distractors, the target object can be found immediately. In this case the set size, the number of distractors, has no influence on the time to find the target. If on the other hand the target object can only be found if one looked for an object with a certain combination or conjunction of low level properties, the search time increases with set size. For example, the search time is constant when looking for a triangle between different numbers of squares. The same holds true for searching for a red circle between blue distractors, but if the subject has to search for red circle between blue circles and red triangles as distractors, the search time was proportional to the number of distractors.

The feature integration theory is a well known theory of visual information processing and serves as inspiration for a well known model for visual attention, which we will refer to with saliency model, by Itti et al. (1998) which we will discuss in the next chapter.

2.1.4 Top-down and bottom-up influences on eye movements

By means of eye movements we select a part of a scene. We select a part of the available information which becomes the focus of our attention. The questions is how and why do we select that specific area and choose to ignore other areas. The control of visual attention and therefore eye movements is usually described in terms of two complementary processes. The first is a bottom-up process also called stimulus-driven or exogenous control and the second is a top-down process also known as goal-directed or endogenous control. An example of bottom-up or date-driven eye movements can be seen in the pop-out effect discussed in section 2.1.3 in which a bottom-up property (e.g. the deviating color of the item) in a display can grab attention and evoke an eye movement. Top-down influence of eye-movements is the influence of

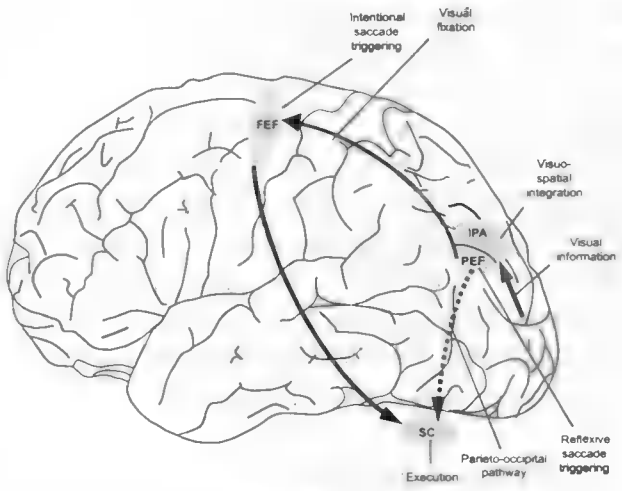


Figure 2.9: Simplified representation of the two pathways which are involved with eye movements. A parieto-tectal pathway which is involved in reflexive bottom-up movements and a pathway to the frontal eye fields (FEF) which is involved with top-down processes. Abbreviations: PEF, posterior eye field; IPA intraparietal areas; FEF; frontal eye fields; SC, superior colliculus. (Based on a figure from Pierrot-Deseillingny et al. (2004))

the task or expectations of the observer when viewing a certain scene. An example can be seen in figure 2.7. The dichotomy between bottom-up and top-down can also be found in the brain. It is thought that the parieto-tectal pathway guides stimulus-driven saccades whereas the goal-driven or top-down saccades are controlled by frontal eye fields (Pierrot-Deseillingny, Milea, & Müri, 2004). See figure 2.9 for the locations of the pathways. It interesting to know that babies are not able to make voluntary eye movement up to 2 or 3 months only the development cortical oculomotor pathways (path to frontal eye fields) makes this possible. Up to 6 months they are not able to inhibit involuntary saccades or produce anticipatory saccades. Indicating that the top-down control develops in a later stage and showing the interference between top-down and bottom-up control.

The influence of top-down and bottom-up processes on eye movements is a strongly debated topic. The question is how these processes relate to each other and which process has control over the eye movements in a certain situation. These questions are mainly investigated with the use of search tasks. On one side Theeuwes (2004) performed an experiment, which according to him, showed that if a distractor is salient enough and within the window

of attention, the size of which depends on the configuration of the display, the distractor will always grab attention. The task consisted of finding a target which has a diamant shape, between a number of circles. In some cases the display contains a distractor which is of a different color than the rest of the shapes. Although the color of the distractor is irrelevant for the search task, the reaction times in presence of the distractor become higher. This indicates that subject pays at least some amount of attention to the distractor. According to Theeuwes (2004), this shows a bottom-up stimuli will override the top-down search strategy and that top-down selection of a certain stimulus dimension (e.g. color and shape) is not possible. Meaning you cannot make a conscious decision to only look at objects with a certain color or shape.

On the other side, Bacon and Egeth (1994) maintain that the allocation of visual attention is controlled by top-down processes and argue that irrelevant distractors will only grab attention if the subject is in a so called singleton detection mode in which subjects do not search for a particular shape but look for any deviating form. The irrelevant distractor will not capture the attention if a person is in a so called *feature search mode* in which a person searches for an object with a particular feature such as shape or color. Pashler (1988) found in his experiments that if a person knows which target form to search for, the person is not distracted by the irrelevant color singleton.

If there is no task to find a certain object than both views argue that bottom-up information will grab attention. This raises the question which attributes are expected to guide bottom-up stimuli or which properties are expected to pop-out. Color, motion, orientation and size are considered to always attract attention. Less certain but also probable are luminance onset or flicker, shape (although it is clear shape does indeed guide intention it is unclear which specific properties are of importance), pictorial depth cues, line termination, curvature and closure (J. M. Wolfe & Horowitz, 2004). These are all low level properties, but Hochstien and Ahissar (2002) argue that conjunctions of low level features, and even information processed at categorical level such as recognition of your own face will also pop-out. Indicating that even attributes which are thought to be processed by top-down processes are able to grab attention in a bottom-up way.

There is now definitive view of the interactions between bottom-up and top-down. The interaction between top-down and bottom-up and how they influence fixations will remain a topic of research for quite some time.

In these sections we have discussed the several topics involved in eye movements. Besides the physiological aspects of the eye we also discussed the attentional processes involved in eye movements. Some elements of the physiological aspects will return in the models we will use to predict fixa-

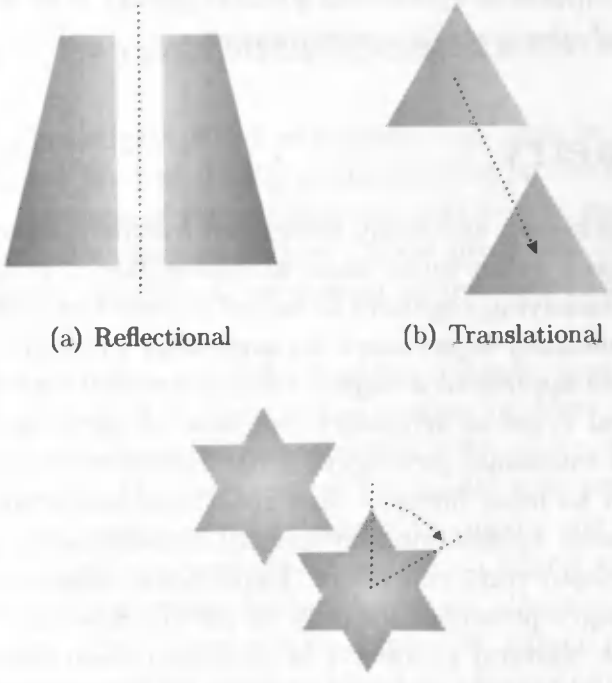
tions and the description of attentional process give an idea of the processes involved in making even a single eye movement.

2.2 Symmetry

Symmetry is a well known and easily recognized feature in images. Although the subject has been given quite some attention (for a review see Wagemans (1997)), no satisfying cognitive or neural explanation has been put forward. Symmetry means a object stays the same after 2 dimensional euclidean transformations are applied to a object, such as translations, rotations, and reflections. Several types of symmetry can now be identified: reflectional, translational, and rotational (see figure 2.10. Reflectional symmetry seems to be more salient to most humans than rotational and translational symmetry. In contrast to translational symmetry, reflectional symmetry seems to be detected without conscious effort. Reflectional symmetry can already be detected in images presented for only 50 ms (Lochner & Nodine, 1989). This suggests that bilateral symmetry is processed preattentive. Although symmetry is much less salient when subjects in experiments are not explicitly asked to detect symmetry, reflectional symmetry is seen without explicitly instructing the subjects to search for symmetry. A striking example is a subject suffering from left visual neglect, which is a condition in which the subject has no conscious access to the visual information projected onto the left hemisphere. He/She was still able to detect symmetry although the subject was not able to point out the symmetry explicitly (Wagemans, 1997).

One factor influencing the detectability of mirror symmetry is the axis in which the is image is reflected. Mirror symmetry is detected more readily when the axis of symmetry is vertical. Furthermore, symmetry with a vertical axis is more notable than symmetry with a diagonal axis. Deviations from main axis are less conspicuous to the human observer. However it is to early to conclude that these preferences for some orientations is hardwired in the neural tissue. Other experiments by Wenderoth (1994) as cited by Wagemans (1997), suggest that the frequency of the orientations within a trial can to a large extend influence the saliency of the orientations by modulating a subjects scanning strategy. If the same oblique orientation is offered many times the sensitivity for this orientation increases.

Reflectional symmetry is easier to discover if the axis of symmetry is located at the center of the gaze, and deviation away from the center especially influences the symmetry perception of high frequency images like dot groups. For closed form patterns, such depicted in figure 2.10, with a low spatial frequency central presentation of the stimulus is less important



(a) Reflectional (b) Translational

(c) Rotational

Figure 2.10: Types of symmetry illustrated using a random group of dots. (a) Illustrates reflectional symmetry. The dots are mirrored in the vertical axis. (b) Example of translational symmetry. The connected dots illustrate the translation applied to every dot. (c) Example of rotational symmetry.

(Barlow and Reeves (1979) as cited by Wagemans (1997)). This suggests detection of symmetry may be done differently for low and high spatial frequency images. High spatial frequency images may require local comparisons of its constituents, e.g. dots, to detect the symmetry.

Symmetry does not appear do be detected in a completely bottom-up fashion. Not all parts of a pattern need a symmetric counter part for the pattern to be considered symmetric. For example, the parts of the image closest to the mirror axis contribute the most to the perception of symmetry. However symmetry still can be perceived if symmetry features close to the mirror axis are not present (Wenderoth, 1995). Furthermore small disturbances of symmetry often go unnoticed. For example a face is considered symmetric, but on closer inspection no face is really symmetric. This suggests that the visual system emphasizes symmetric properties. But studies

in animals suggest that lack of *symmetricness* can be linked to genetic deficiencies (Møller, 1993). Making it advantageous to be able to detect small perturbations of symmetry to enhance the chance of finding a good mate. Surprisingly this apparent contrasting sensitivity is also found in psychological trials. Human subjects detected symmetry in patterns which consisted of only 30% symmetric dot pairs. On the other hand small deviations from perfect symmetries were detected in a comparison test between perfect and imperfect symmetries. Concluding that symmetry detection is both robust and sensitive at the same time.

A last example which seems to contradict local processing of symmetry is the lack of effect of properties of the smallest parts of an image containing symmetry. If an image is made of short line segments instead of a dot pattern such as in figure 2.10 the spatial grouping of the lines is more important for the perception of symmetry than the orientation of the individual line segment. Indicating large scale blobs seem to be processed earlier than the individual properties the line segments.

Although sensitivity to symmetry does not appear to be a completely bottom up process, reflectional symmetry seems to be processed fast and preattentive. Furthermore, experiments have shown that symmetry influences the fixation patterns of humans. It was found that when subjects watched a symmetric image they confined their fixations to one side of the symmetry axis (Lochner & Nodine, 1989). Possibly to take advantage of the redundancy in the information present in the image. These experiments used very simple shapes and dot patterns. To find out if the location of fixations were also influenced by symmetry in more complex images, Lochner and Nodine (1989) conducted experiment with works of art. The experiments showed a tendency for subjects to fixate near the axis of symmetry.

Summarizing, symmetry seems a cue for guiding eye movements. Symmetry is hardly used as a predictor for human fixation. Only in the article by Privitera and Stark (2000) a algorithm based on symmetry is used, but did not receive much attention. Therefore, we think it is interesting to see if we can predict fixations by using symmetry as predictor. The algorithms we use to calculate symmetry are based on the *generalized symmetry transform* by Reisfeld et al. (1995) which, contrary to other computational models, does not require preprocessing, such as object segmentation, of the source image before applying the symmetry measure. Making it a suitable low level and bottom up algorithm to detect *regions of interest*.

2.3 Related research: comparing eye movements to computational models

In this thesis we investigate which algorithms can best be used to predict human eye fixations. We test this by comparing the predictions by the algorithms with the fixations of human subjects obtained in experiments. Our goal is to find out which properties give a good prediction of the saliency of a given image region. Therefore, we present an outline of studies that made a similar comparison.

Because eye movements are defined as the tell-tale signs of visual attention, most related research is done in the context of overt visual attention. One way to better understand visual attention is to construct computational models. By comparing the outcomes of the model with experimental data from humans we can assess the correctness of the model and this can confirm or refute the assumptions about visual system on which model was based. Several models aim to simulate bottom-up visual attention. Such a model predicts which part of an image will be fixated by human observers. Many of these models are based on findings about the functioning of early visual processing as outlined in section 2.1.2. One model of visual attention is the model of saliency based visual attention for rapid scene analysis or the *saliency model* as we will refer to it.

Itti, Koch, and Niebur (1998) based the saliency model on the *feature integration theory* by Treisman and Gelade (1980) (see section 2.1.3 for a discussion). The assumption that several low level visual properties are combined into one saliency map is at the base this model. A saliency map is a representation of the saliency found in a image at certain location and calculated with an algorithm such as the saliency model. An example can be seen in figure 2.11. The low level properties are based on color, orientation, and intensity information. All of which are features for which the human visual system is also sensitive. The exact manner in which these features are processed and combined is explained in section 3.1. Itti et al. (1998) did not compare their model directly to human subjects, but found that their model showed a similar performance as human subjects in pop-out and conjunction search tasks. As explained in section 2.1.3 these tasks entail finding a certain element in a display (an example displays can be seen in figure 2.8. The saliency model always predicts that attentios is immediately shifted to the element which differs from the distractors (other elements in a display) with respect to one of the low level propertie such as color. The effect is not influenced by the number of distractors present in the displays. Thereby, reproducing the pop-out effect. If the element which has to be found differs

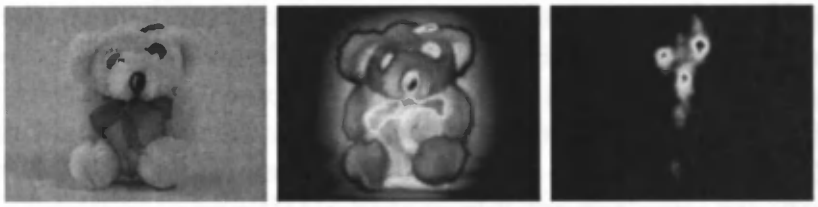


Figure 2.11: (a) Input image (b) Saliency map (c) Fixation map. The saliency map is generated by the saliency model with the image of a bear as input. The fixation map which represents all fixation made in response to the same image.

from the distractors by a conjunction of low level features, e.g. color and orientation, the model produced a search time which was linearly dependent on the number of elements present in the display just as observed with humans performing these tasks.

A more direct comparison of the saliency model to human behavior was done by Parkhurst, Law, and Niebur (2002). Parkhurst et al. (2002) investigated to what extend saliency guides the overt visual attention of human subjects. To this end, they setup an experiment in which they recorded the eye movements of human participants while the participants were viewing images. They used a so called *freeview* experiment in which the participants watched images without a specific task. Every image was presented for five seconds. The images used were divided into four groups: fractals, home interiors, natural landscapes, and buildings and city scenes. The same images were processed by the saliency model (Itti et al., 1998) and the outcomes were compared with the data obtained from the experiments. They found that the saliency at locations of the first fixation made by a participant in response to an image were significantly higher than the saliency at random locations. This shows that local stimulus properties, such as color, orientation and intensity, indeed guide the overt visual attention of humans. Furthermore, they found that the saliency of the first fixation on an image is significantly higher than the rest of the fixations. The saliency of the later fixations gradually drops, but later fixations did not drop to levels of chance and the saliency remained constant after an initial decline. Moreover, they found that the correlation of locations of fixations with saliency was weakest for the interior photos and strongest for the fractals. They proposed two possible explanations for this fact. First, a possible influence of top-down attentional biases. For example, they noticed subjects had the tendency to fixate objects on table tops even if the objects were not particularly salient. They conjectured

that the top-down directed tendency to search table tops is a good strategy to find interesting objects. Second, it is possible that the images of fractals contain fewer areas with high saliency compared to the other types of images. This will cause this area to pop-out and therefore attract more attention. Confirmation for the second possibility was found in the observation that the saliency maps generated from the fractals contained fewer but more prominent local maxima of saliency compared to the other types of images. Parkhurst et al. also compared the performance of the different feature channels of the saliency model. Overall, the channels which used intensity and color information performed better than the channel based on orientation information, but the rankings of the different channels varied considerably for each image group. For example the orientation channel performed better on the buildings and city scenes than the channel based on color information. According to Parkhurst et al. (2002) this shows the importance to use multiple properties while evaluating saliency, because sensitivity to only one property can perform poorly on certain sets of images.

A similar approach was taken by Ouerhani, Wartburg, Hügli, and Müri (2004). They also conducted a freeview experiment in which they asked to human participants to watch several images for five seconds without a specific task. Again, the saliency model was used as the computational model with which the human fixations were compared. However, they introduced a different method to compare the saliency maps with the human data, as will be explained in detail in section 5.2, Ouerhani et al. constructed so called fixation maps from the human fixation data. These *fixation maps* can be directly compared with the computational models by means of correlation. Thereby providing a metric which indicates the similarity between human and the computational model. They found a positive correlation between the model and human subjects, but also found that the variability among the human subjects was quite high. Meaning that there were considerable differences in correlation between the different subjects and the saliency model. They came to the conclusion that their results tended to agree with the view that visual attention is influenced by bottom-up stimuli, but considered it a preliminary conclusion due to the small number of subjects and images used in the experiment.

Another paper which compared computational models to human fixations to computational models was written by Privitera and Stark (2000). In their paper, ten algorithms, as described in section 3.4, were compared to human fixations. The human fixations were gathered in a freeview trial in which each picture was presented for three seconds. They offered 15 images with, among others, terrain photographs, paintings, and landscapes. They used yet another method to compare human fixations with the algorithms. A clustering

method was applied to the saliency maps to obtain a number of local maxima equal to the average number of fixations made in response to an image. Subsequently, the locations of the local maxima and fixations were compared with each other to determine the similarity between the human fixations and the algorithms. This method enabled them to compare the sequence predicted by the algorithms with the sequence of the fixations, but this did not yield any significant results. When they compared the positions of the local maxima and the human fixations, they found that different algorithms performed better on different images but overall they concluded found that a measure based on wavelets gave the best performance. Another observation was that a measure based on symmetry performed well on general images and a measure contrast performed well on Mars terrain images. A qualitative comparison with a questionnaire. In this questionnaire the participants were asked to judge to what extend saliency maps corresponded visually with the fixation maps. The results from the questionnaire picked the algorithms based on orientation and edge detection as the best performers.

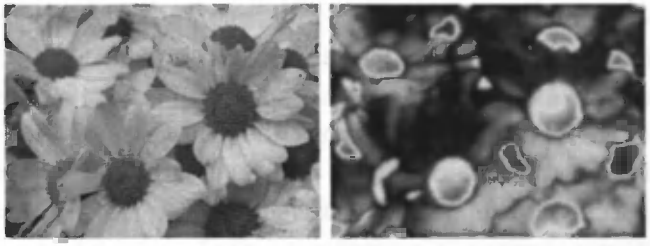
An alternative model was proposed by Le Meur, Le Callet, and Thoreau (2006). Le Meur et al. argued that the saliency model containa arbitrary steps which cannot be justified when taking the human visual system into consideration and that their model is more biologically plausible. They compared the performance of their model with the saliency model. Although they did find a tendency of their model to perform better than the saliency model the results were not significant. They also concluded that the saliency model tended to perform better on images with few and small interest points.

The research summarized in these sections is used as basis for our research and experiments. Several methods of evaluating the experiments and the setup of the experiments will serve to answer our research questions.

Chapter 3

Predictors of human fixations

Our goal is to investigate to which extend human fixations can be predicted. We are comparing data from humans with prediction of fixations made by several algorithms which we are going to describe in this chapter. Each algorithm takes a image as input and produces a saliency map. A saliency map assigns a value to an area in an image. This value indicates how likely it is for that area to be fixated or how interesting this area is to the human observer. An example can be seen in figure 3.1. First, we discuss the saliency model which is a model of visual attention by Itti, Koch, and Niebur (1998). Second, we explain the SIFT model, a well know algorithm from the field of computer vision. Third, we will go over algorithms based on symmetry. Finally, we discuss several algorithms from a article by Privitera and Stark (2000).



(a) (b)

Figure 3.1: (a) Input image. (b) Saliency map. The saliency map is generated with the saliency model. The saliency model is one of the algorithms we use to predict human fixations.

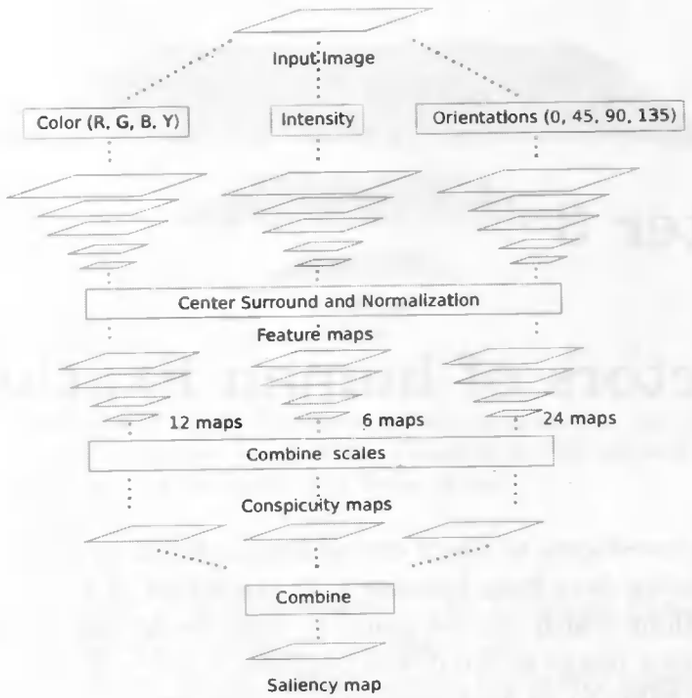


Figure 3.2: Model of Saliency-based visual attention by Itti, Koch, and Niebur (1998). (From Itti et al., 1998)

3.1 Salieny model

The model proposed by Itti et al. (1998) is a model of saliency-based visual attention for rapid scene analysis. It is a biological plausibel model which given a inpute image, models the locations of human fixations. To make the model saliency-based, a saliency map has to be extracted from the current scene. The saliency map is a 2D topographic representation in which every point is a value between zero and one representing the saliency of a region in the picture. This value indicates how likely it is for a human observer to direct its attention to a certain part of the image. Based on this information we can make a prediction of the eye movements.

3.1.1 Model description

The model takes a color image as input. The saliency model extracts which are considered to be biologically plausible. Itti et al. use features based on: intensity, color and orientation. Therefore, the image is split into three

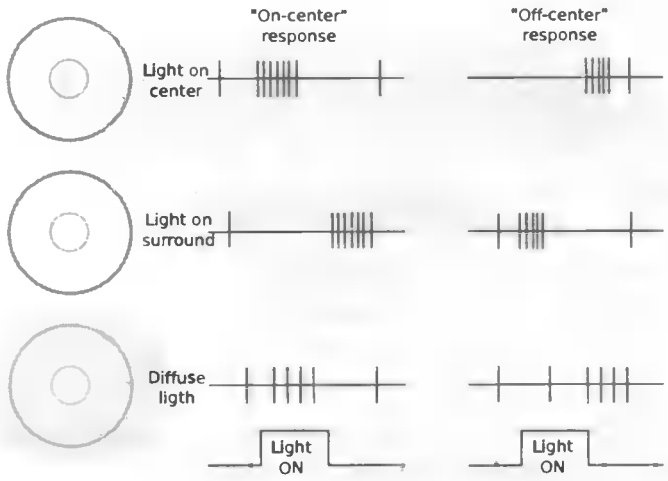


Figure 3.3: Reaction of a center-surround receptive field illuminated at different areas. The frequency of the spike trains indicate the response of the receptive fields. A higher frequency means a greater response. The response of the two configurations of the cell are shown: the on-center and off-center configuration. In the on-center configuration the cell will respond excitory to light on the center and inhibitory to light on the annulus. The off-center configuration will react inhibitory to light on the center and excitory to light on the annulus.

different modalities which reflect these features (See figure 3.2).

For every modality a local operation has to be constructed to extract a feature from an image region. For example, if the modality is orientation, we have to determine if a image patch has a unique orientation compared to its surroundings, otherwise it would not stand out with respect to its orientation and therefore is not considered salient. Therefore, features are extracted by applying center-surround operations. The center-surround operation reacts the same as a center-surround receptive field. An explanation of a center-surround receptive field is given in figure 3.4. A center-surround operation yields a large value if there is a large difference between the center and the surround. The center-surround operation is implemented as a subtraction of a coarse scale from a fine scale images from a Gaussian pyramid.

The Gaussian dyadic pyramid is constructed for each modality. Gaussian because each step the image is low pass filtered with a Gaussian kernel. Dyadic in the sense that in each step the image is subsampled by a factor of two. This process is repeated eight times. First on the original image and subsequently on the resulting image of the previous step. Resulting in nine

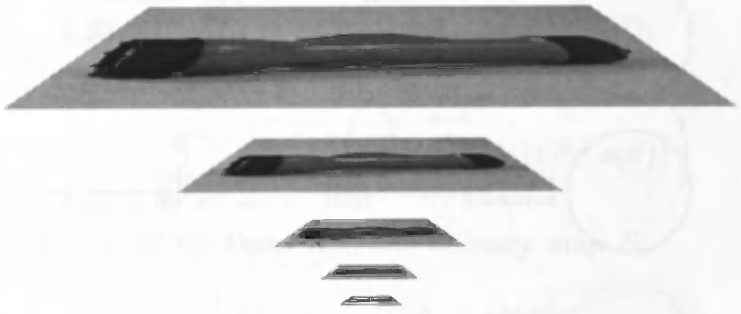


Figure 3.4: A dyadic Gaussian pyramid is formed by convolution of the original image with a Gaussian kernel, followed by downsampling the image by a factor of two.

spatial scales ranging over eight octaves. This Gaussian pyramid will form the basis of the feature extraction (See figure 3.4).

To compare features across different scales, the center-surround operation is performed with several combinations of images obtained from the gaussian pyramid. The scale of the center pixel is denoted by $c \in \{2, 3, 4\}$. The surround pixel is $s = c + \delta$ with $\delta \in \{3, 4\}$. The coarse scale in upsampled to the finer scale and then substracted from the finer scale, thereby implementing an on-center off-surround process. This results in a total of six feature maps. The construction of these feature maps from a Gaussian pyramid is denoted by $\ominus$ from now on.

To construct the saliency map related to intensity, an intensity map $I = (r + g + b)/3$ is extracted from the red, green, and blue channels of the rgb encoded picture. r b g are the 8 bit values decoding red, green and blue. From the intensity map a Gaussian pyramid is constructed which is followed by the center-surround operation on the different scales. This results in six feature maps $\mathcal{I}(c, s)$ as follows:

$$\mathcal{I}(c, s) = |I(c) \ominus I(s)|. \tag{3.1}$$

Furthermore four color channels are created: a red channel $R = r - (g + b)/2$, a green channel $G = g - (r + b)/2$, a blue channel $B = b - (r + g)/2$, and finally a yellow channel $Y = (r + g)/2 - |r - g|/2 - b$. Resulting in four Gaussian pyramids. Again the center-surround calculation is performed. Two color opponency maps, $\mathcal{RG}(c, s)$ and $\mathcal{BY}(c, s)$, defined respectively as:

$$\mathcal{RG}(c, s) = |(R(c) - G(c)) \ominus (G(s) - R(s))| \tag{3.2}$$

and

$$\mathcal{BY}(c,s) = |(B(c) - Y(c)) \ominus (Y(s) - B(s))|, \tag{3.3}$$

are formed using the color combinations red/green and blue/yellow. Giving a total of 12 feature maps.

To construct the orientation modality, four orientation channels are extracted. A grayscale image is convolved four times with a Gabor filter with 4 different orientations $\theta \in \{0°, 45°, 90°, 135°\}$. Four Gaussian pyramids are created from these maps. Subsequently, 24 orientation feature maps,

$$\mathcal{O}(c,s,\theta) = |O(c,\theta) \ominus O(s,\theta)|, \tag{3.4}$$

were created from the pyramids.

The three modalities result in a total of 42 feature maps which have to be integrated into one saliency map. The difficulty in combining these maps lies in the fact that they have been extracted in different ways which results in different ranges for maps of different modalities. If a simple summation was used to combine all these maps, a salient point in a map with a relative low range of values can easily be obscured by high valued noise in another feature map. This way, it is possible that a clear local maximum or salient point in one of the constituent maps is not represented in the overall saliency map. Therefore, a normalization, $\mathcal{N}(.)$, is performed on all the feature maps. This process consists of two parts. First, all maps are normalized to the same range. Second, the maps are multiplied by $(M - \bar{m})^2$ with M being the global maximum of the map and $\bar{m}$ the average of all other local maxima where a local maximum is defined as a value that is higher than the values of its neighbors in a radius of 8 points. The global maximum is the local maximum with the highest value. By applying the normalization given by:

$$\mathcal{N}(f_{map}) = (M - \bar{m})^2 \cdot f_{map}, \tag{3.5}$$

maps with a strong global maximum compared to the other local maxima are promoted. On the other hand, maps with a relative homogeneous distribution of maxima are demoted. This is desired because these maps do not have clear salient points.

The next step is to combine the feature maps into three *conspicuity maps*. An intensity map $\bar{\mathcal{I}}$, a color map $\bar{\mathcal{C}}$, and a orientation map $\bar{\mathcal{O}}$. The maps are created by scaling each map to scale four, followed by a summation of the different maps, which in the following formula is denoted by $\oplus$:

$$\bar{\mathcal{I}} = \bigoplus_{c=4}^{2} \bigoplus_{a=c+3}^{c+4} \mathcal{N}(\mathcal{I}(c,s)) \tag{3.6}$$

$$\bar{C} = \bigoplus_{c=4}^{2} \bigoplus_{a=c+3}^{c+4} [\mathcal{N}(\mathcal{RG}(c,s)) + \mathcal{N}(\mathcal{BY}(c,s))] \tag{3.7}$$

$$\bar{O} = \sum_{\theta \in \{0°,45°,90°,135°\}} \mathcal{N}\left(\bigoplus_{c=4}^{2} \bigoplus_{a=c+3}^{c+4} \mathcal{N}(\mathcal{O}(c,s,\theta)) \right). \tag{3.8}$$

The last step is to combine these into one saliency map $\mathcal{S}$:

$$\mathcal{S} = \frac{1}{3} \left(\mathcal{N}(\bar{I}) + \mathcal{N}(\bar{C}) + \mathcal{N}(\bar{O}) \right). \tag{3.9}$$

We will use this saliency map to compare the outcome of this model to the human data obtained from the experiment.

3.1.2 Biological plausibility

Itti et al. emphasize the biological plausibility of their model. Much of the information discussed in chapter 2 served as bases to construct the saliency model. They base their model on the feature-integration theory (Treisman & Gelade, 1980). Which entails that several modalities are combined to determine the overall saliency of a part of an image. The features or modalities to be integrated in the model by Itti et al. are extracted with a center-surround process using intensity, color-opponency, and orientation information. The intensity modality is supported by the fact that ganglion cells in the retina react center-surround to light intensity (Kuffler, 1953). Furthermore center-surround color-opponency can be found in the human visual cortex (Engel, Zhang, & Wandell, 1997). This *double opponent* system involves cells found in the striate cortex of which the center reacts exitory to one color, blue for example, and inhibitory to another, e.g. yellow. Contrary to the surround which has the opposite reaction to both colors (Livingstone & Hubel, 1984). Sensitivity to orientation is found in the hypercolumns in the striate cortex (Hubel, 1974) (See section 2.1.2).

Calculating the center-surround features between different combinations of spatial scales suggests a integration of information over these scales. Sensitivity to different spatial scales is reported to be found in the hypercolumns. Deeper tissue in a hypercolumn are more sensitive to more coarse spatial scales (Silverman, Grosof, De Valois, & Elfar, 1988). Futhermore cells located low in the visual hierarchy, such as the V1 area, react to a visual angle of 0.1-0.5 degrees. Cells higher in the hierarchy react to a much bigger area. Cells in the IT or inferior temporal cortex, a high level area, react to 25 degrees or more. These areas are all interconnected (Essen & Maunsell,

1983). It is therefore likely that information from different scales is somehow integrated.

3.2 SIFT

SIFT or Scale Invariant Feature Transform by Lowe (2004) is developed for artificial object recognition. The algorithm represents an image by a set of interest points, so called keypoints, that are extracted from an image and stored in a database. The keypoints can be compared to the keypoints extracted from the image which has to be identified. The selection of keypoints is robust to affine transformation such as rotation, scaling and shearing. For example, if the SIFT algorithm is offered an image and a rotated version of that image the algorithm will extract the same keypoints. This makes the algorithm very well suited for object recognition. For this thesis it is interesting to see if a well performing algorithm from the field of computer vision selects the same areas as humans do with fixations. The SIFT algorithm will not be used to match images, but we will use the locations of the keypoints extracted by the SIFT algorithm as salient points. Therefore, we are only interested in the extraction of the keypoints and not the matching process.

The SIFT algorithm consists of four parts. One, scale-space extrema detection. This is the process of searching for locations which are potential robust features. Robust means not sensitive to rotation and unique within its neighborhood. A difference-of-Gaussian function is used to find the location of the keypoints in different scales. Two, keypoint localization. Third, a selection made from the interesting points found in step one based on various metrics of stability. For example, keypoints will not be located on an edge, because an edge can produce a series of identical keypoints. Three, orientation assignment. Every keypoint gives rise to one or multiple dominant orientations of the local gradient in the vicinity of the keypoint. For every dominant orientation a keypoint with its assigned orientation is added for consideration. This orientation is used to obtain a description of a keypoint which is invariant to rotation of the image. Fourth, a keypoint descriptor invariant to change in 3D perspective is constructed from the locations and orientation of the keypoints.

Only the first two steps are interesting for selecting salient locations in an image. Step three and four are used to make the descriptor more robust for the matching process for which SIFT is intended. Subsequently only the first two steps are discussed in detail. We will use notation adopted from Lowe (2004).

3.2.1 Difference-of-gaussian pyramid

To construct a scale space representation the function $L(x, y, \sigma)$ is repeatedly used to convolve a Gaussian function with a image, $I(x, y)$. $L(x, y, \sigma)$ is defined as:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y). \tag{3.10}$$

The Gaussian function, $G(x, y, \sigma)$, is defined as:

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}. \tag{3.11}$$

After one convolution σ is increased with a factor k resulting in several parts of the scale-space representation of the image. Each octave in scale-space is divided in a number of intervals s, therefore $k = 2^{1/s}$. To find stable keypoints the local extrema in scale space are investigated. To construct a scale space representation of the image, a difference-of-gaussian function is used to contstruct a gaussian-pyramid. The difference-of-gaussian is found by subtracting two adjoining scale space images (see figure 3.5). Therefore, the difference-of-gaussian, $D(x, y, \sigma)$, is given by:

$$\begin{aligned} D(x, y, \sigma) &= ((G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \\ &= L(x, y, \sigma) - L(x, y, \sigma). \end{aligned} \tag{3.12}$$

After one octave is processed the image is downsampled by a factor two and the whole proces for finding extrema is repeated for another octave.

3.2.2 Selection of keypoints

To cover one octave of scale-space $s + 3$ images are generated. It is now possible to determine the extrema in scale space. A point is an extreme if it is the maximum or minimum of its 26 neighbors in the difference-of-gaussian pyramid. The selected extrema will now be examined to see if they are stable.

To determine if a candidate keypoint is stable, a 3D quadratic function is fitted to the scale-space function $D(x, y, \sigma)$, with a potential keypoint as origin. $D(x, y, \sigma)$ is defined as:

$$D(x) = D + \frac{\partial D^T}{\partial \mathbf{x}}\mathbf{x} + \frac{1}{2}\mathbf{x}^\mathbf{T}\frac{\partial^2 D}{\partial \mathbf{x}^2}\mathbf{x}. \tag{3.13}$$

The interpolated extremum, $\hat{\mathbf{x}}$, is determined by taking the derative of $D(\mathbf{x})$ and setting it to zero, giving:

$$\hat{\mathbf{x}} = -\frac{\partial D^{-1}}{\partial \mathbf{x}^2}\frac{\partial D}{\partial \mathbf{x}}. \tag{3.14}$$

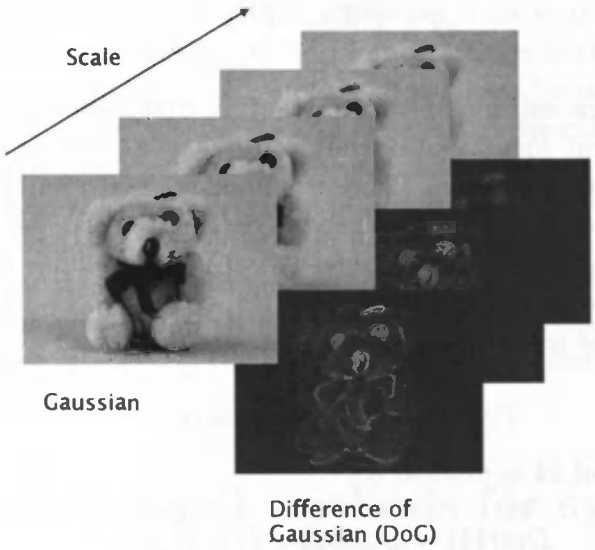


Figure 3.5: Difference-of-gaussian images are constructed by subtracting two adjoining images from scale space. The scale space images are constructed by convolving the input image with a gaussian function with a increasing sigma.

If the offset, $\hat{\mathbf{x}}$, is larger than 0.5 in any dimension then the extremum is closer to another point. In this case the calculation is redone with this new point as origin. Finally the offset is added to the sample point under consideration. Furthermore, the value of the function $D(\hat{\mathbf{x}})$ can be used as an indication of the stability of the extremum. The extremum is rejected if $|D(\hat{\mathbf{x}})| < 0.03$, where $D(\hat{\mathbf{x}})$ is given by:

$$D(\hat{\mathbf{x}}) = D + \frac{1}{2}\frac{\partial D^T}{\partial \mathbf{x}}. \tag{3.15}$$

This way, keypoints with low contrast are rejected. It is also desirable to disregard points situated on edges. Keypoints located on an edge are not local maxima in all directions and therefore less suitable as keypoint.

To determine if a keypoint is poorly defined, the principal curvature has to be calculated using a 2x2 Hessian matrix. The principle curvature are the minimum and maximum curvature of a point on a plain. So if both values are positive, than one can speak of a maximum at that location. By calculating the eigenvalues of the Hessian matrix, $\mathbf{H}$, we can estimate the principal curvatures of D as they are proportional to the eigenvalues. $\mathbf{H}$ is

given by:

$$\mathbf{H} = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix}. \tag{3.16}$$

The derivatives are estimated by subtracting neighboring points from the value of the current location. By determining the ratio of the eigenvalues we can decide if the keypoint is well defined in multiple directions. It is not necessary to determine the eigenvalues themselves. Only the ratio is needed. If α is the largest eigenvalue and β is the smaller one, the sum of the eigenvalues can be determined with the trace of $\mathbf{H}$. Furthermore, the determinant can be used to determine their product. The trace of $\mathbf{H}$ is given by:

$$Tr(\mathbf{H}) = D_{xx} + D_{yy} = \alpha + \beta. \tag{3.17}$$

The determinant of $\mathbf{H}$ is defined as:

$$Det(\mathbf{H}) = D_{xx}D_{yy} - (D_{xy})^2 = \alpha\beta. \tag{3.18}$$

If the curvatures have different signs the extremum is discarded. Now we can define r as the ratio between α and β so $\alpha = r\beta$. Therefor r is defined as:

$$\frac{Tr(\mathbf{H})^2}{Det(\mathbf{H})} = \frac{(\alpha+\beta)^2}{\alpha\beta} = \frac{(r\beta+\beta)^2}{r\beta^2} = \frac{(r+1)^2}{r}. \tag{3.19}$$

Because r is always larger than one, $(r+1)^2/r$ has a minimal value if α and β are equal. Furthermore, $(r+1)^2/r$ will increase if r increases. Therefore, given r, in our case $r = 10$, we can use the the following equation to determine if a keypoint is selected.

$$\frac{Tr(\mathbf{H})^2}{Det(\mathbf{H})} < \frac{(r+1)^2}{r} \tag{3.20}$$

If this is the case the location is kept as an interesting keypoint. At this point the keypoints can be used to construct a saliency map which can compared to human human data. The keypoint is treated similar as a fixation obtained from human data. To construct a saliency map from the set of keypoints we plot a Gaussian function for every fixation. The center of the Gaussian is the location of the keypoint and the variance is chosen to model the area which is fixated by the fovea. The number of keypoints we used was the average number of fixations made by the subjects given a certain picture. Which is usually about fifteen fixations.

If we compare the SIFT and the saliency model we can see a couple of similarities and differences. SIFT and the saliency model both use a scale space representation constructed with a Gaussian pyramid. While the saliency

saliency model uses a dyadic pyramid which uses, by definition, only whole octaves, the SIFT model uses multiple scales per octave to construct a more detailed scale space representation. Another difference is that SIFT, as just discussed, produces a number of locations which can be considered salient, while the saliency model produces a saliency map directly. Furthermore, SIFT only considers contrasts in intensity while the saliency model also looks at contrast based on color and orientation information. A similarity between the algorithms is that they both use difference-of-gaussian operation on the pyramid to extract points with high local contrast, but SIFT only subtracts neighboring scales, while the saliency model uses different combinations of scales. Concluding, SIFT and the saliency model are related, but there are obvious differences.

3.3 Computational methods for symmetry

As mentioned in chapter 2, symmetry is considered a cue for guiding eye movements. This section will be devoted to the description of the regions of interest predictors which try to detect symmetry in images. Several symmetry algorithms where used to build a saliency map. First we will discuss two algorithms by Reisfeld, Wolfson, and Yeshurun (1995). Second we will discuss a variation on these algorithm by Heidemann (2004). Finally we will discuss the phase symmetry algorithm by Kovesi (1997).

3.3.1 Isotropic symmetry

The first algorithm we will discuss is proposed in an article by Reisfeld et al. (1995). Like the previous algorithms it is a context free, local algorithm, meaning the algorithm attempts to define a measure of symmetry without knowing anything about the content of the image. The algorithm, isostropic symmetry, detects presence the symmetry around a given pixel.

Symmetry is defined with the use of gradients in an image. The symmetry measure is constructed as follows. A intensity gradient of a point p_k is defined as:

$$\nabla p_k = \left(\frac{\partial}{\partial x} p_i, \frac{\partial}{\partial y} p_i \right). \tag{3.21}$$

For each point, p_k a vector, $v_k = (r_k, \theta_k)$, of the local gradient is defined, with $r_k = \log(1 + \|\nabla p_k\|)$ and $\theta_k = \arctan(\frac{\partial}{\partial y} p_k / \frac{\partial}{\partial x} p_k)$. This vector represents the direction and magnitude of the gradient.

Now, let p be a point for which the symmetry value has to be determined. Given two points p_i and p_j, which form a point symmetric pair centered at

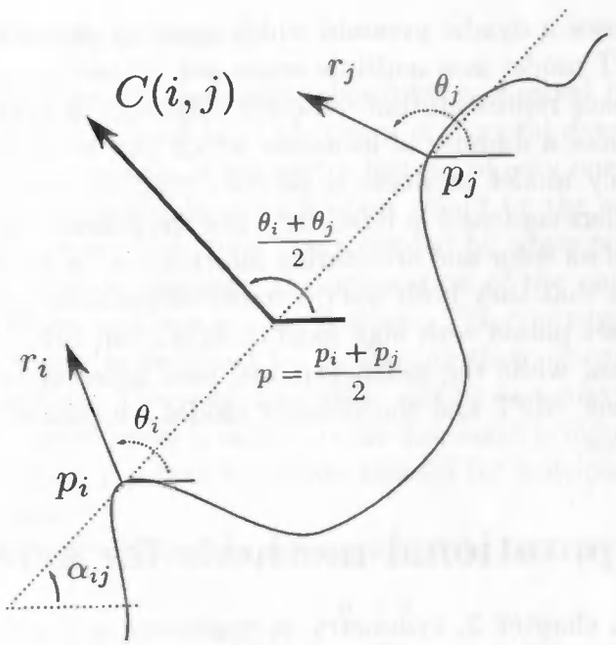


Figure 3.6: The contribution, $C(i,j)$, to the symmetry value of point p is determined by considering the orientations of the gradients, θ_i and θ_j, at point p_j and p_j. p_i and p_j form a point-symmetric pair of points with center p.

p, and line l defined by these points, α_{ij} is the angle between l and the horizontal (See figure 3.6). Furthermore, a point symmetric set of points around p, $\Gamma(p)$, within a certain radius R, is defined as:

$$\Gamma(p) = \left\{ (i,j) \left| \frac{p_i + p_j}{2} = p \ \wedge \ \|p_i - p_j\| \leq 2R \right. \right\}. \tag{3.22}$$

A distance weight function, $D_\sigma(i,j)$, is used to enforce the circular shape of the neighborhood around p. $D_\sigma(i,j)$ is defined as:

$$D_\sigma(i,j) = \frac{1}{\sqrt{2\pi\sigma}} e^{\left(-\frac{\|p_i - p_j\|}{2\sigma}\right)}. \tag{3.23}$$

The locality of this symmetry transform can be adjusted by varying the σ of the distance weight function. For a large value of σ a larger region around p will be taken in consideration.

To determine the contribution of the pair of points a phase weight function $P(i,j)$ is defined as:

$$P(i,j) = [1 - \cos(\theta_i + \theta_j - 2\alpha_{ij})] \times [1 - cos(\theta_i - \theta_j]. \tag{3.24}$$

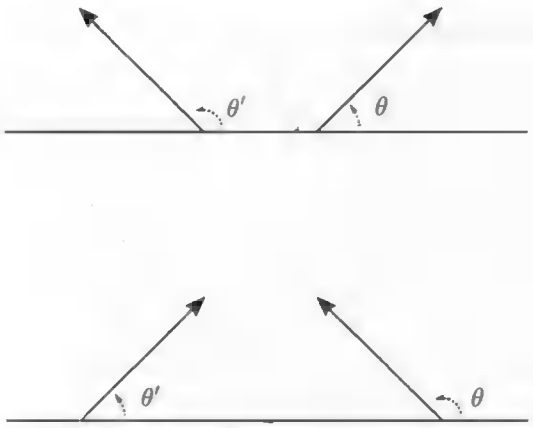


Figure 3.7: Gradients pointing to or away from each other correspond to a light object on a dark background or a dark object on a light background. Both cases result in a equal symmetry value.

The phase weight function determines the contribution of the pair of point symmetric points by looking at the orientation of the corresponding gradient. The phase weight function consists of two terms. The first, $1 - \cos(\theta_i + \theta_j - 2\alpha_{ij})$, has its maximum value when $(\theta_i - \alpha_{ij}) + (\theta_j - \alpha_{ij}) = \pi$. This is the case when the gradients are pointing toward or away from each other as can be seen in figure 3.7. The first term will also reach its maximum value when $\theta_i - \alpha_{ij} = \theta_j - \alpha_{ij} = \pi/2$. This occurs when p_i and p_j are on a straight edge. We are not interested in finding edges of an object, but we are interested in detecting symmetric objects. Therefore, we introduce the second term, $1 - \cos(\theta_i - \theta_j)$, to compensate for this case.

Now we can express the contribution of the points p_i and p_j as:

$$C(i,j) = D_\sigma(i,j)P(i,j)r_i r_j. \tag{3.25}$$

$r_i r_j$, the gradient weight function, is introduced to emphasize the symmetries which are found on edges. r_i and r_j are the logarithm of the magnitude of the gradient at both points. The isotropic symmetry of point p can no be defined as:

$$M(p)_\sigma = \sum_{(i,j)\in\Gamma(p)} C(i,j). \tag{3.26}$$

The saliency map of an image can be constructed by determining the isotropic symmetry of ever pixel in the image.

3.3.2 Radial symmetry

A variant of the isotropic symmetry algorithm is dubbed radial symmetry (Reisfeld et al., 1995). Reisfeld et al. claim it is useful detect points which are crossed by multiple axis of symmetry instead of only one dominant symmetry direction. An example of such point would be the center of a black circle on a white circle. The circle is surrounded by edges pointing outward. This is done by first determining the orientation of the pair of pixels with the highest isotropic symmetry contribution. The contribution, $C(i,j)$, of the other pixel pairs is weighted by comparing their orientation with this dominant orientation. The weighting value will be at a maximum if the orientation of the contribution is at 90° to the dominant symmetry orientation. This ensures that only points which are crossed by multiple symmetry axis obtain a high value.

To this end we first define the orientation of the contribution of p_i and p_j as:

$$\sigma(i,j) = \frac{\theta_i + \theta_j}{2}. \tag{3.27}$$

The *symmetry direction*, $\phi(p)$, of point p is the direction of the line through the pair of points which have the largest contribution to M_σ. Which gives us a measure of the symmetry for p as:

$$S_\sigma = [M_\sigma(p), \phi(p)]. \tag{3.28}$$

The radial symmetry value for point p is now defined by:

$$R_\sigma(p) = \sum_{(i,j)\in\Gamma(p)} C(i,j)\sin^2[\varphi(i,j) - \phi(p)]. \tag{3.29}$$

The weighing of the contribution with the factor $\sin^2[\varphi(i,j) - \phi(p)]$ will be at a maximum if $\varphi(i,j) - \phi(p) = \pi/2$. The saliency map of an image can be constructed by determining the isotropic symmetry of ever pixel in the image.

3.3.3 Color symmetry

In a article by Heidemann (2004) an extension to the aforementioned algorithm is proposed. The extension is proposed to remedy the problem that arises when looking at gray scale images. If only intensities are used and the edges are between two colors of equal intensity. Edges, and therefore the symmetries around these edges, will not be detected. An example would be a square on a uniform background of a different color, but with the same

intensity value. A solution to this problem is to use color information in the calculation of the symmetry values. The image is split into three different channels and not only are the gradients within the channels compared, but also the gradients between the channels. This ensures that gradients which are not present in the intensity channel will be detected in the different color channels. The algorithm which does not incorporate color is stated in the following form:

$$M(p) = \sum_{(i,j)\in\Gamma(p)} P_{Gray}(i,j) \cdot G(i,j) \cdot D_\sigma(i,j). \tag{3.30}$$

$D_\sigma(3.23)$ is the distance weight function and $G(i,j)$ is the gradient weight function given by $r_i r_j$. With the phase weight function P_{Gray} defined as:

$$P_{Gray} = \underbrace{[1-\cos(\gamma_i+\gamma_j)]}_{P^+_{Gray}(i,j)} \cdot \underbrace{[1-\cos(\gamma_i-\gamma_j)]}_{P^-_{Gray}(i,j)}. \tag{3.31}$$

With γ_i and γ_j defined as:

$$\begin{aligned} \gamma_i &= \theta_i - \alpha_{ij} \\ \gamma_j &= \theta_j - \alpha_{ij}. \end{aligned}$$

The term $P^+_{Gray}(i,j)$ of the phase weight function obtains it maximum value if the orientations of the point gradient point away from each other and therefore differ by π. The $P^+_{Gray}(i,j)$ reaches its greatest value when the gradients are pointing towards or away from each other.

To avoid having to use a Gaussian weighting function as distance weight function Heidemann (2004) only considers points within a certain distance. But we will consider all points within a certain radius, R, and additionally weigh them with a distance weight function. Therefore, the set of points, $\Gamma^*(p)$, is defined as:

$$\Gamma^*(p) = \left\{ (i,j) \;\middle|\; \frac{p_i+p_j}{2} = p \;\wedge\; \|p_i - p_j\| \le 2R \right\}. \tag{3.32}$$

To incorporate the color information in the algorithm Heidemann (2004) extends M_{Gray} with an extra summation over all color pairs. The set of combination of colors is denoted by $k,l \in 0,1,2$. Where $0,1,2$ represents the color channels red, green, and blue. Heidemann (2004) only considers points which have an gradient magnitude above a certain threshold, but we will consider all points and weigh them with a distance weight function. The set of color channel combinations is given by:

$$\Lambda(p) = \{(k,l) \mid k,l \in 0,1,2\}. \tag{3.33}$$

Which results in the following expression for calculation of the color symmetry value of point p:

$$M'_{Col}(p) = \sum_{(i,j)\in\Gamma^*(p)} \sum_{(k,l)\in\Lambda(p,i,j)} P'_{Col}(i,j,k,l) \cdot G_{Col}(i,j,k,l) \cdot D_\sigma(i,j). \quad (3.34)$$

The incorporation of the color channels gives the following formulation for the gradient weight function:

$$G_{Col}(i,j,k,l) = \log(1 + GM_k(p_i)) \cdot \log(1 + GM_l(p_j)). \quad (3.35)$$

Where $GM_i(p)$ denotes the gradient magnitude of point p in color channel i.

There is another problem with the original formulation of Reisfeld et al. (1995). Namely, consider a symmetric object on a non-uniform background. If one side of the object has a greater intensity than the background and the inverse is true for the other side, the original algorithm will not detect a symmetry. This is because a good response from the phase weight function necessitates the gradients to point either to or away from each other. To remedy this problem Heidemann changed the phase weight function to become a π-periodic function instead of the original period of 2π. The new phase weight function is denoted as follows:

$$P_{Col}(i,j,k,l) = \underbrace{[\cos^2(\gamma_{ik} + \gamma_{jl})]}_{P^+_{Gray}(i,j,k,l)} \cdot \underbrace{[\cos^2(\gamma_{ik} \cdot \cos^2(\gamma_{jl})]}_{P^-_{Gray}(i,j,k,l)}. \quad (3.36)$$

With γ_{ik} and γ_{jl} defined as:

$$\begin{aligned} \gamma_{ik} &\doteq \theta_{ik} - \alpha_{ij} \\ \gamma_{jl} &= \theta_{jl} - \alpha_{ij}. \end{aligned}$$

with γ_{ik} and γ_{jl} defined as the orientation of the gradient at point i in color channel k and the orientation of the gradient at point j in color channel l respectively. $P^-_{Col}(i,j)$ expresses how parallel the gradients are relative to the line between p_i and p_j denoted by $\overline{p_i p_j}$. If either of the gradients is perpendicular to $\overline{p_i p_j}$ the measure will be zero. $P^+_{Col}(i,j)$ will have its maximum value when there is a perfect mirror symmetry with $\overline{p_i p_j}$ as mirror axis. For every pixel the radial symmetry value will be collected into the saliency map, $\mathcal{S}$S.

Compared to isotropic symmetry, color symmetry will be more sensitive to objects of which the intensity does not differ significantly from the background provided the color of the object is different. As humans generally also perceive color this should provide an advantage in predicting human fixations. Furthermore, the sensitive for symmetric objects on non-uniform backgrounds makes the color symmetry algorithm more versatile.

3.3.4 Phase symmetry

One objection that is made by Kovesi (1997) with respect to the symmetry measures by Reisfeld et al. is the dependency of the measure on contrast. A circle with a large intensity compared to its background will be considered more symmetric than a circle with low contrast. This emphasizes parts of the image with a local maximum with respect to symmetry but there is no absolute measure of symmetry. The phase symmetry algorithm, as discussed in this section, computes symmetry from local phase information. Symmetry can be identified by looking at the phase of different frequency components.

On a point of symmetry all frequency components are either at a minimum or at a maximum. Consider the 1D example is shown in figure 3.8. This illustrates the detection of symmetry present in a square wave. Symmetry can be located by looking for locations where all the frequency components are either at a minimum or maximum. For a square wave the symmetry axis is located in the center of the up and down peaks of the wave and indeed, all frequency components are either at a minimum or a maximum at the center of the peak. This enables the phase symmetry algorithm to detect mirror symmetry present in images.

A wavelet transform is used to extract local frequency and phase information. Because phase information is utilized to determine symmetry or asymmetry the wavelet has to have a linear phase response. Meaning the filter cannot alter the phase information. Therefore, we use wavelets based on complex valued Log-Gabor functions. Convolutions with Log-Gabor functions do not alter the phase of the signal on which the convolution is applied. The wavelet pair consists of a even and odd pair which can be seen in figure 3.9. Now that we have our filters, the signal can be analyzed by convolving it with a quadrature pair of wavelets for each scale n. Each pair consists of an even-symmetric and odd-symmetric wavelets denoted by M_n^e and M_n^e given by:

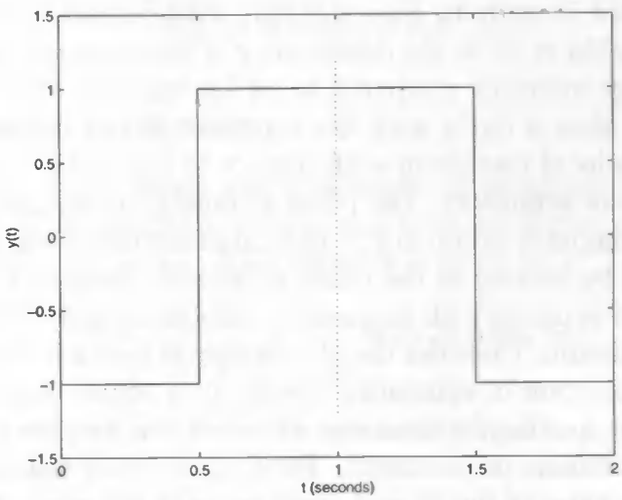
$$[e_n, o_n] = [I * M_n^e,\ I * M_n^o]. \tag{3.37}$$

For every pixel in the image there now exists an array of responses of the convolution with the even and odd function for every scale. To aide extraction of the symmetry information we consider the values e_n and o_n the real and imaginary part of a complex valued frequency component. The amplitude at scale n is given by:

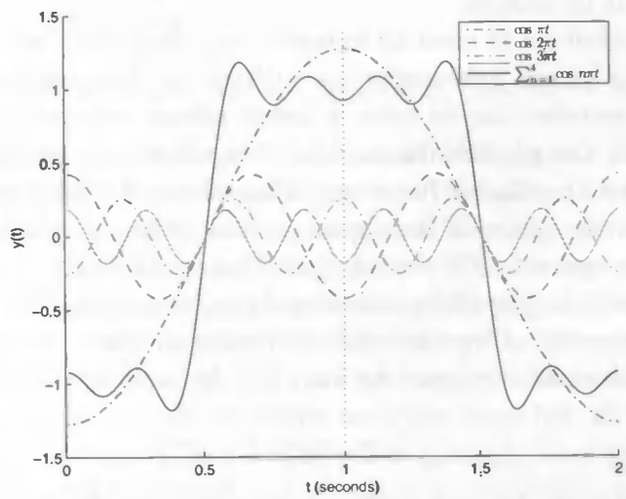
$$A_n = \sqrt{e_n^2 + o_n^2} \tag{3.38}$$

and the phase is denoted as:

$$\Phi_n = \text{atan2}(e_n, o_n). \tag{3.39}$$

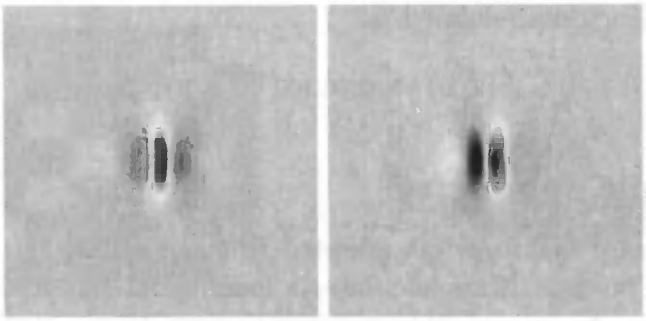


(a) Squarewave



(b) Decomposition of squarewave

Figure 3.8: (a) Squarewave (b) Fourier decomposition of a square wave. The Fourier decomposition consists of the first four components which are plotted together with their summation. The symmetry axis is located at the point were all frequency components are either at a minimum or a maximum, which is for $t = 1$.



(a) Even filter (b) Odd filter

Figure 3.9: (a) Even filter. (b) Odd filter. Even and odd filter used by the phase symmetry algorithm.

The amount of symmetry can now be determined. The even-symmetric filter will give a great response to a symmetric point in the image. The converse is true for the odd-symmetric filter which will give a high response on a asymmetric point in the image. Therefore to construct a measure of symmetry at a certain point, the response from the odd-symmetric filter is subtracted from the even-symmetric filter. Which in turn is normalized with the total response of both filters. A next step is to combine the information of several scales by summating the symmetry values from different scales, resulting in the following formula for the symmetry value:

$$\begin{aligned} V(x,y) &= \frac{\sum_n [A_n(x,y)[\,|\cos(\Phi_n(x,y))| - |\sin(\Phi_n(x,y))|\,] - T]}{\sum_n (x,y) + \varepsilon} \\ &= \frac{\sum_n [[\,|e_n(x,y)| - |o_n(x,y)|\,] - T]}{\sum_n A_n(x,y) + \varepsilon}. \end{aligned} \tag{3.40}$$

The factor T is a noise compensation factor. This factor is estimated by looking at the response of the filter at the finest scale. The reasoning behind this is that the most fine scale is so small it will have a limited response to structure. Therefore, responses will most likely be due to noise. Therefore, the mean reaction of the filter at the smallest scale will be used as estimation for the noise in the image. On every scale this factor will be estimated by looking at the size of the scale compared to the smallest scale.

We now have determined the amount of symmetry for one orientation. To determine the symmetry present in several orientations we have to repeat the calculation with filters with several different orientations, $O = \{0°, 30°, 60°, 90°, 120°, 150°\}$. The symmetry value of a certain pixel is the maximum of the symmetry value calculated for every orientation. Which

gives us the saliency map defined as:

$$\mathcal{S}(x,y) = \underset{o \in O}{\operatorname{argmax}}(V_o(x,y)) \tag{3.41}$$

3.3.5 Simple symmetry

The last symmetry measure was adopted from an article by Privitera and Stark (2000). This algorithm calculates a local symmetry measure, $S(x,y)$, for every pixel in the image. This measure is given by:

$$S(x,y) = \sum_{(i_1,j_1),\in\Gamma(x,y)} s((i_1,j_1),(i_2,j_2)). \tag{3.42}$$

$\Gamma(x,y)$ is an area with a radius of 7 surrounding pixel x,y. The area, $\Gamma(x,y)$, is defined as follows:

$$\begin{aligned}\Gamma(x,y) = (x-r,y),\ldots,(x,y),\ldots(x+r,y),\\ (x,y-r),\ldots,(x,y+r).\end{aligned} \tag{3.43}$$

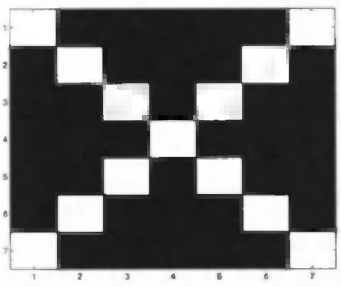
$s((i_1,j_1),(i_2,j_2))$ is given by:

$$s((i_1,j_1),(i_2,j_2)) = G_\sigma(d((i_1,j_1),(i_2,j_2))|cos(\theta_1-\theta_2)|. \tag{3.44}$$

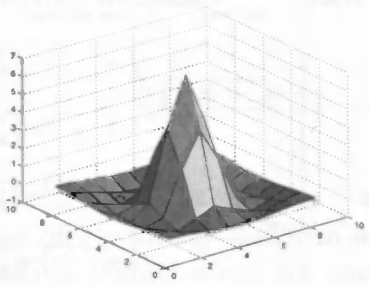
G_σ is a Gaussian with a standard deviation of 3 pixels. Its argument $d(\cdot)$ calculates the distance between two points. Together they form a distance weight function. θ_1 and θ_2 are the angles of the intensity gradients at the two points (i_1,j_1) and (i_2,j_2). The angles are found using a convolution of a Sobel kernel with the intensity only version of the input image. If the angles are oriented in the same direction or differ by π, the cosine term will reach its maximum. The contribution of the cosine term is influenced by the distance weight function. If the pair of points are close together, the contribution of the points will be bigger. The contribution of every pixel pair is summed to give the symmetry magnitude S is equail to the saliency map $\mathcal{S}$. In chapter 4 we will compare this symmetry algorithm to the symmetry algorithms, discussed earlier.

3.4 Further image processing algorithms

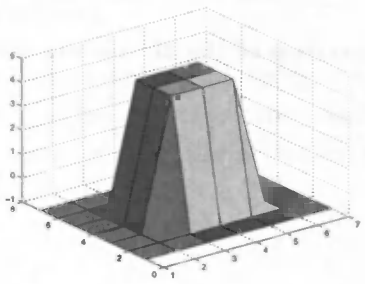
Together with the aforementioned symmetry measure simple symmetry, several other algorithms were adopted from Privitera and Stark (2000). As discussed in section 2.3 they compared these algorithms to human data and compared them to each other to look for similarities. Every algorithm takes



(a) Xlike



(b) LoG



(c) Center surround

Figure 3.10: (a) Xlike kernel (b) Laplacian of the Gaussian (LoG) kernel (c) Center surround kernel. Kernels used in the xlike, LoG and center surround algorithms

an image as input and produces a saliency map. After the explanation of the algorithms we will discuss the adaption of the algorithms to produce a saliency map based on multiple scales. For most algorithms we will adopt the same notation as the article if appropriate.

3.4.1 Xlike

This algorithm convolves an image with a 7×7 kernel. The kernel is x-shaped with positive values on the diagonals and negative values anywhere with total sum of zero. The kernel will give a large response to high intensity edges oriented in the direction of the diagonals. In contrast to uniform areas, which will yield a low response. The kernel used can seen in figure 3.10

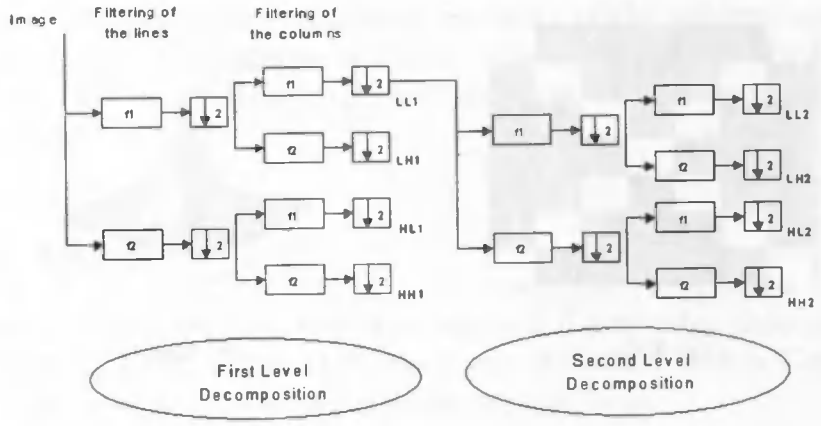


Figure 3.11: Illustration of the 2D discrete wavelet transform up to two levels.

3.4.2 Wavelet

This algorithm is based on a discrete wavelet transform. We used a pair of so called conjugate quadrature filters (CQF) based on the Daubechies wavelet family (Daubechies, 1992). The wavelet chosen for this algorithm is the Cohen-Daubechies-Feauveau 9/7 tap wavelet. This specific wavelet is also used for lossy compressions in the JPEG-2000 standaard. From the Cohen-Daubechies-Feauveau 9/7 a low-pass filter and a highpass filter are constructed.

The order of processing is outlined in figure 3.11. First, the rows of the source image are separately filtered with the low pass filter and with the high pass filter. Second, the rows of the two resulting matrices are both down sampled by a factor of two. Subsequently, the columns of both matrices are high and low pass filtered. After applying the filters, the columns are also down sampled by a factor of two. Applying the filters in this manner results in a total of four matrices which represent four different frequency bands. One represents the horizontal lows and vertical lows (ll), another the horizontal highs and the vertical lows (hl), a third represents the horizontal lows and the vertical highs (lh), and finally one represents the horizontal highs and vertical highs (hh). The whole procedure is then repeated with the ll image as input which represents a low passed version of the original image. By repeating this procedure 3 times we can obtain three matrices hh_1, hh_2, hh_3. These matrices are scaled to the original size of the image and

combined into a saliency map with the following formula:

$$\mathcal{S} = \sum_{i=1}^{N} (hh_i). \tag{3.45}$$

Repeating the procedure three times gives us: $n = 3$. The saliency map gives summary of the high frequency components of 3 octaves. Uniform areas at each scale are filtered out and only areas which contain a certain amount of detail in the image are preserved and are therefore considered salient.

3.4.3 Center surround

This algorithm mimics a center surround receptive field by convolution of the image with 7×7 mask. The values of the mask have a total sum of 1. The mask has a positive value of 4 in the 9 center positions and a negative value in the periphery (See 3.10).

3.4.4 Orientation

The image is separately convolved with four Gabor functions with the orientations: 0°, 45°, 90°, and 135°. Every pixel, x, y, gives four scalar values obtained from the convolutions. Every scalar is associated with one of four unit vectors. Each of the unit vectors has the same orientation as one of the Gabor filters. Subsequently, the unit vectors summed with the corresponding values from the convolution as weights. This results in the orientation vector $\bar{\mathrm{o}}(x, y)$. A center-surround orientation difference measure can now be defined as:

$$\mathcal{O}(x, y) = (1 - \bar{\mathrm{o}}(x, y) \cdot \bar{\mathrm{m}}(x, y))\|\bar{\mathrm{o}}(x, y)\|\|\bar{\mathrm{m}}(x, y)\|. \tag{3.46}$$

$\bar{\mathrm{m}}(x, y)$ is the average orientation vector of the 7×7 neighborhood surrounding the point (x, y). The first term, $1 - \bar{\mathrm{o}}(x, y) \cdot \bar{\mathrm{m}}(x, y)$, will have a small value if the orientation vector, $\bar{\mathrm{o}}(x, y)$, and average orientation, $\bar{\mathrm{m}}(x, y)$, are pointing in the same direction. By multiplying with the magnitude of the average orientation vector, the function $\mathcal{O}(x, y)$ will be larger if the orientation vectors in the neighborhood are large and smaller if the orientation vectors in the neighborhood are small. Thereby, the second term acts as a low-pass filter. The orientation difference measure is now considered as saliency map, $\mathcal{S} = \mathcal{O}$. One of the modalities of the saliency model described in section 3.1 is also based on orientation. The basic idea of finding areas which deviate from their surroundings with respect to gradient orientation is

the same. Contrary to the orientation modality of the saliency model, this algorithm does not use multiple scales to implement a center-surround operation with a difference-of-gaussian, but calculates the average orientation in a area around the pixel.

3.4.5 Edges

For this algorithm the first step is to apply a Canny edge detector to the image (Canny, 1986). After that, the image is filtered with a Gaussian of $\sigma = 3$. This gives us a measure for edges per unit area.

3.4.6 Entropy

Entropy in the context of information theory is a measure of the amount of data measure in bits. Shannon (Shannon, 1948) defined information entropy as follows. Given a discrete random variable X with outcomes $x_1 \ldots x_i$ entropy is defined as follows:

$$H(x) = \sum_{i=1}^{n} p(x_i) \log_2 \left(\frac{1}{p(x_i)} \right) = -\sum_{n=1}^{n} p(x_i) \log_2 p(x_i). \tag{3.47}$$

$p(x_i) = Pr(X = x_i)$ is the probability of the i^{th} outcome of X.

This tells us how much information a certain outcome conveys. If a random variable has only one outcome, observing the outcomes will tell nothing new. If there are many different outcomes a outcome will remove a considerable amount of uncertainty about the outcome thus giving a considerable amount of information. In the former case the entropy is low and in the latter case the entropy is high.

To use entropy as a saliency measure (Privitera & Stark, 2000) considers a 7×7 patch of pixels to calculate a probability distribution of intensity values. In the case of a digital image this is a discrete distribution or histogram. This measure is given by:

$$\mathcal{S}(x, y) = -\sum_{i \in G} f_i \log f_i. \tag{3.48}$$

G is the local set of intensity values and f_i the frequency of the i^{th} intensity value. In this way patches with many different intensity values will yield a high entropy value and are considered salient.

3.4.7 Michaelson contrast

Michaelson contrast is measure for contrast which compares the intensity of a patch of pixels to the intensity in the whole images and is defined as:

$$\mathcal{S}(x,y) = \|(\mathcal{L}_m - L_M)/(\mathcal{L}_m + L_M)\|. \tag{3.49}$$

$\mathcal{L}_m$ is the average luminance within a 7×7 neighborhood around (x, y) and L_M is the mean luminance of the whole image.

3.4.8 Discrete cosine transform

The discrete cosine transform is related to the Fourier transform. The cosine transform expresses a signal, or in this case an image, in terms of cosines with different periods and amplitudes. The transform for in two dimensions for a square matrix ($N \times N$ is given by:

$$t(i,j) = c(i,j) \sum_{n=0}^{N-1} \sum_{n=0}^{N-1} s(m,n) \cos \frac{\pi(2m+1)i}{2N} \cos \frac{\pi(2n+1)j}{2N}. \tag{3.50}$$

The inverse transform is given by:

$$s(m,n) = \sum_{n=0}^{N-1} \sum_{n=0}^{N-1} c(i,j)s(m,n) \cos \frac{\pi(2m+1)i}{2N} \cos \frac{\pi(2n+1)j}{2N}. \tag{3.51}$$

With $c(0,j) = 1/N$, $c(i,0) = 1/N$ and $c(i,0) = 2/N$ if $i \neq 0 \wedge j \neq 0$. We used $N = 8$, to this end image was divided blocks with a size of 8×8. Subsequently, every block was transformed with a DCT. The 64 coefficients, resulting from the transform, each represent the presence of a cosine with a certain frequency. By removing the sixteen coefficients which represent the lowest frequencies, we only retain the high frequency parts of a block. After applying the inverse transform, on the blocks the image is effectively high pass filtered and this results in the saliency map, $\mathcal{S}$.

3.4.9 Laplacian of the Gaussian

The image is convolved with a Laplacian of the Gaussian. The Laplacian of the Gaussian is given by the following formula:

$$LoG(x,y) = -\frac{1}{\pi\sigma^4}\left[1 - \frac{x^2+y^2}{2\sigma^2}\right] e^{-\frac{x^2+y^2}{2\sigma^2}}. \tag{3.52}$$

The 9×9 kernel was constructed with $\sigma = 1.4$ (see figure 3.10). The saliency map is defined as:

$$S = I * LoG \tag{3.53}$$

The convolution of the image with a LoG is very similar to the the difference-of-gaussian operation. Both can be regarded as a center-surround operation.

3.5 Constructing the saliency maps.

The model of visual attention by Itti et al. (1998) takes several scales of the input image into account. The image is repeatedly down sampled to enable the model to calculate the features, for example orientation, on different scales. The outcomes of the different scales are then combined to form the eventual saliency map.. Before the images are combined by summing the images of the different the levels, a normalization method (3.5) is applied to the image. The normalization procedure emphasizes maxima which stand out in their surroundings. This is to prevent a saliency map from one level with a uniform but high value to dominate the overall saliency map.

The original versions of the algorithms adopted from the article by Privitera and Stark (2000), described in section 3.4, and simple symmetry, described in section 4.5, did not incorporate multiple scales. Privitera and Stark applied every algorithm, with the exception of $\mathcal{W}$, on the image at its original size. To be able to make a fair comparison to the saliency maps obtained from the model by Itti et al. (1998), each algorithm performs its operations on five octaves, which are subsequently combined into a single saliency map after being normalized with the method from Itti et al..

Chapter 4

Analysis of symmetry

The algorithms based on the detection of symmetry play a prominent role in this thesis, therefore, we shall devote this chapter to discuss and compare the properties of the five symmetry algorithms discussed in chapter 3. We will give some examples of the output of the algorithms and we will discuss similarities and differences between the algorithms. The first algorithm is the *isotropic symmetry* algorithm as described in section 3.3.1. The second is the *radial symmetry* algorithm as discussed in section 3.3.2. The third is the *color symmetry* algorithm as described 3.3.3. The fourth is the *phase symmetry* method as described in section 3.3.4. And finally the algorithm *simple symmetry* as discussed in section 3.3.5.

4.1 Isotropic symmetry

The isotropic symmetry algorithm tries to determine if a given location or pixel is situated which is centered at symmetric gradients. Given a certain location, pairs of gradients, which form point symmetric pairs around the pixel under consideration, are compared to each other. A gradient is the first derivative of the intensity at a certain location. The gradient has a magnitude and an orientation for every location. To determine the symmetry of a certain pixels the gradients associated with pairs point-symmetric points around that pixel are investigated. Only pairs of gradients within a certain radius, r, contribute to the symmetry value of the pixel under consideration. A pair of gradients will have a positive contribution to the symmetry value of that pixel if the gradients point towards or away from each other. This will occur with a light symmetric object on a dark background or a dark symmetric object on a light background. This is illustrated by figure 4.1. We expect strong responses on axis of mirror symmetry. Which includes the

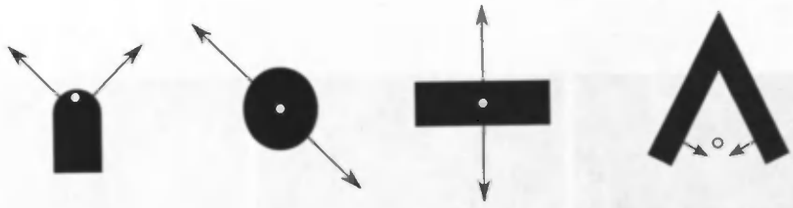


Figure 4.1: Illustrations of reactions to symmetry by isotropic symmetry algorithm. The arrows are the directions of the gradients. The white circle is the location which receives the contribution of the pair of gradients.

center of circles, corners and also the center lines of rectangles (see figure 4.4). The response of the algorithm to a center of a circle is stronger than the response to the point on a center line of a rectangle, because the center of the circle is crossed by multiple axis of symmetry while the point of a rectangle is crossed by one or two axis of symmetry.

The magnitude of the gradients is also taken into account. A gradient will give the greatest contribution if it is located on a edge between a black and a white area. The isotropic symmetry algorithm will not show any response in uniform areas since the gradient magnitudes are zero in that case. Furthermore, as explained in section 3.3.1, the algorithm does not respond if the gradients of the points under consideration are located on the same edge. If this is the case, both gradient angles will be the same and the term $1 - \cos(\theta_i - \theta_j)$ of the phase weight function (3.24) and subsequently the phase weight function will be zero.

An example of the response of the algorithm on simple shapes can seen in figure 4.2. We have used a two star shapes, odd and even, and a circle to demonstrate the reponse of the algorithms discussed in the chapter. The isotropic symmetry algorithm highlights all symmetry present in the stars correctly and also highlights the center of the circle.

Figure 4.3 gives a example of the response of the isotropic symmetry algorithm on a photographic image. At level 1 the response is mainly to details in the image, such as the symmetry axis between the threads of the bolt. At higher levels such as level 3, the center line of the bolt is visible as the radius now covers the width of the bolt. Moreover, the center of the wings, which roughly have a circular shape, are also highlighted. Furthermore, at level three you can see a response to the symmetry axis between the wings. The response is even more pronounced at level 4. Which illustrates a possible drawback of the symmetry measure as saliency predictor. Namely, that the axis of symmetry between the wings are located in a uniform area, but human

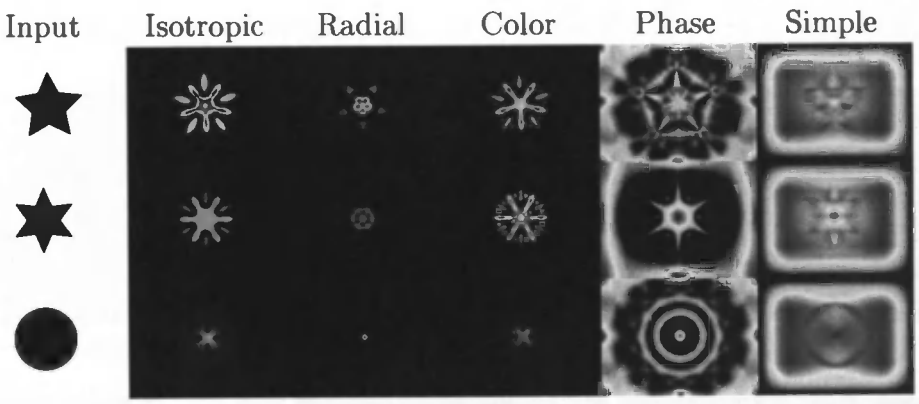
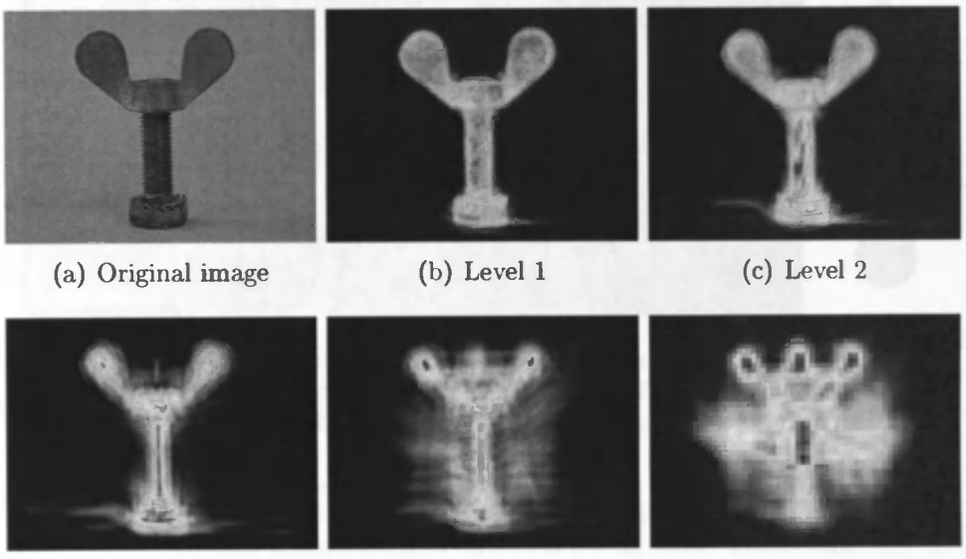


Figure 4.2: Examples of responses of the symmetry algorithms to simple shapes. The three input images, an odd and even star shape (128×96) and a circle (128×96), were offered to isotropic symmetry($r = 24, \sigma = 32$), radial symmetry($r = 24, \sigma = 32$), color symmetry($r = 24, \sigma = 32$), phase symmetry, and simple symmetry($r = 24, \sigma = 32$).

subjects do not tend to fixate on uniform areas (Krieger et al., 2000).

4.2 Radial Symmetry

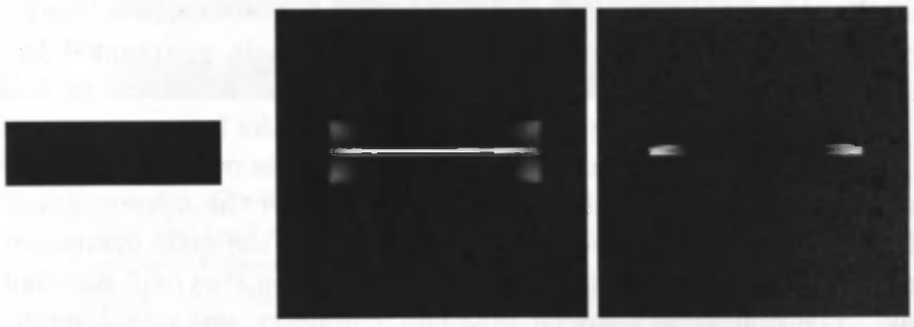
As outlined in section 3.3.2, radial symmetry is an extension to the isotropic symmetry measure. Its purpose is to detect points surrounded by edges. For every point a main symmetry direction is determined. This is the angle formed by the pair of gradients which yields the largest contribution to the symmetry value. Subsequently, the angle of the other gradient pairs are compared to this orientation. The contribution of the other gradient pairs will be large if they are oriented perpendicular to the main symmetry direction, and if they are oriented in the same direction they will not contribute at all. The difference between isotropic symmetry and radial symmetry is illustrated with figure 4.4. The consequence is that radial symmetry will not detect long symmetry axis as present in the length axis of the rectangle. This property can be seen in figure 4.2. Namely, the algorithm does not highlight the symmetry present in the lobes of the stars but does accentuate the center of the stars. Furthermore, the center of the circle is clearly highlighted as expected from radial symmetry.



(a) Original image (b) Level 1 (c) Level 2

(d) Level 3 (e) Level 4 (f) Level 5

Figure 4.3: The isotropic symmetry algorithm, $r = 24$, is applied to an image from a highly symmetric wing nut and bolt. The image (1024×768) is repeatedly down sampled by a factor of two, resulting in the smaller images levels two to five.



(a) input image (b) isotropic symmetry (c) radial symmetry

Figure 4.4: Comparison of the response to a black rectangle (85×25). The radius used was $r = 24$.

4.3 Color symmetry

Symmetry from color is, as explained in section 3.3.3, a adoption of isotropic symmetry by Heidemann (2004). Instead of comparing only gradients present in the intensity channel of an image, color symmetry compares the gradients present in all color channels with each other. Consider a filled square on a uniform background. The square is of a different color than the background, but has the same intensity value. This will cause the square to be undetectable to the isotropic symmetry algorithm as it depends on gradients present in the intensity channel, but the square will detected by the color symmetry algorithm. Thereby, adding a property shared with the human visual system: sensitivity to color.

Another difference with isotropic symmetry is illustrated by an example adopted from the article Heidemann (2004). The phase weight function determines the contribution of a pair of gradients by comparing their orientation. The phase weight function(3.24). of the isotropic symmetry algorithm is defined as $1 - \cos(\theta_i - \theta_j)$ The contribution is greatest when the gradients are pointing to or away from each other, as illustrated in figure 3.7. Consider a symmetric object on a background which is lighter than the object on one side and darker than the object on the other. If there is a equal difference in intensity between the object and the light background and the object and the dark background, the gradients at the side of the object will not point toward or away from each other but will differ by π in orientation, as illustrated in figure 4.5. Therefore, the object will not be considered symmetric by the isotropic symmetry algorithm. To remedy this situation, Heidemann (2004) introduced a new phase weight function (3.3.3) with a periodicity of π. This has the consequence that the two situations in figure 4.5 are considered to be equal. Figure 4.6 illustrates this. The color symmetry algorithm detects the symmetry axis present in the center of the bars. The isotropic symmetry algorithm will only detect the symmetry of the bar located between to darker bars, but the color symmetry algorithm will also detect the symmetry of the bars flanked by a darker and lighter bar. Therefore, the color symmetry algorithm will detect symmetry on a non uniform background where isotropic symmetry algorithm does not. All together the color symmetry algorithm will generally detect the similar features as the isotropic symmetry algorithm with the exception of the situations described above. This can be seen in figure 4.2. Like isotropic symmetry, color symmetry detects all symmetry present in the stars and also the center of the circle.

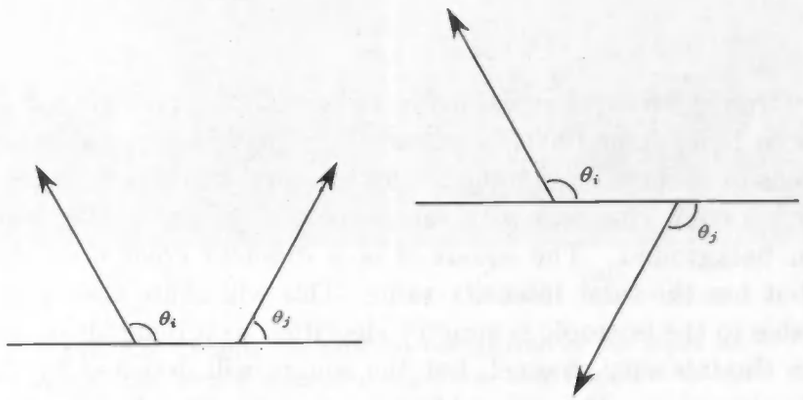
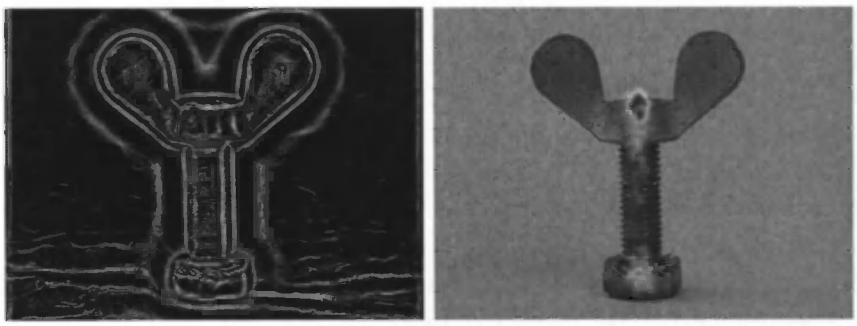


Figure 4.5: Examples of the two pairs of gradients. The first(a) corresponds to a dark object on a light background. The second (b) is a situation with an object on a non uniform background where one side of the background is darker than the object and one side is lighter than the object.

Figure 4.6: Comparison of response between *isotropic symmetry*(b) and *color symmetry*(c). The input image(a) consists of 5 bars with different shades of gray. The bars have a width of 5. The aim is to detect the symmetry axis in the through the center of the bars. The radius used was $r = 5$. The *color symmetry* algorithm detects all symmetry present but the *isotropic symmetry* algorithm only detects the symmetry on the bar which is flanked by bars with a darker shade of gray, contrary to the *color symmetry* which detects all symmetries.

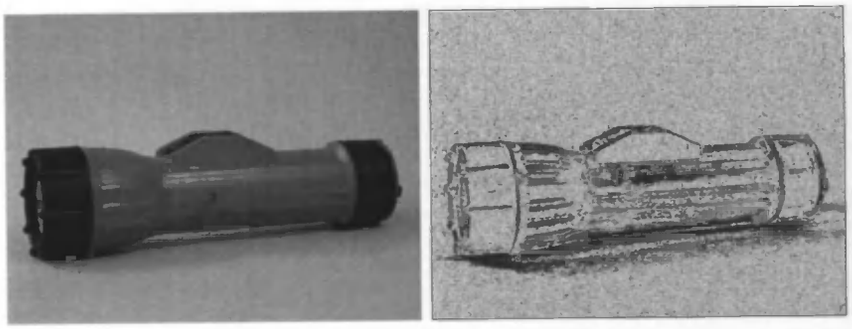


(a) Phase symmetry saliency map. (b) Human fixations.

Figure 4.7: This figure illustrates that the symmetry from phase measure gives a strong response to almost uniform areas. The shadows on the ground have very poor contrast, but give very prominent features in the saliency map. When humans view the this image, they do not appear to be interested in the background. This can be seen in the plot of the human fixations on top of the target image.

4.4 Phase symmetry

According to Kovesi (1997), the algorithms based on the algorithm by Reisfeld et al. (1995) essentially combine symmetry and contrast detectors. Contrary to the algorithms from Reisfeld et al. (1995) and Heidemann (2004), the phase symmetry does not incorporate the strength of the intensity gradients. Therefore, according to Kovesi, his algorithm is a more pure symmetry detector. However, if you use the phase symmetry algorithm on an image, the algorithm will give a strong reaction to parts of an image which are not very salient to the human eye. As shown in figure 4.7 the phase symmetry algorithm gives a very strong response to parts of the background. To a human observer these parts seem to be very uniform and therefore not very interesting regions of the image. As can be seen in the same figure which plots all fixations for this object. In figure 4.2 we show the response of the phase symmetry algorithm to the stars and the circle. Although the symmetric axis of the stars and the center of the circle are highlighted, there are several artifacts due to the use of the frequency spectrum. Especially the rings around the edge of the circle and the response on the borders of the image. Parts of the image are highlighted which are really not symmetric, this is a clear drawback of the algorithm.



(a) (b)

Figure 4.8: Despite the uniform background of the input image(a) the *simple symmetry* algorithm still responds to the structure present in the image(b).

4.5 Simple symmetry

The simple symmetry measure is adopted from the article by Privitera and Stark (2000). The algorithm works is a similar way as the in similar way as the algorithms from Reisfeld et al. and Heidemann the phase weight 3.44 function is given by $|\cos(\theta_i - \theta_j)|$. Which means the function will give the biggest response if the gradient angles are the same, which is the case for edges, or if the angles differ by π, which is the case for the center of a circle. Contrary to the algorithms from Reisfeld et al. and Heidemann, simple symmetry does not take the magnitude of the gradients into account. Therefore, the algorithm will achieve a maximal value on a uniform part of an images. This is because the orientation of the gradient on a uniform background will always be zero and therefore the angles will be equal and the algorithm will give a large response. While this is a problem for artificial images with perfect uniform areas, as we are not interested in uniform areas but in symmetry, it will be less of a problem for real life pictures as uniform areas will always contain noise. The noise will vary the angles of the gradients thereby giving less of a response to uniform areas than to structured areas such as lines. This is illustrated by figure 4.8. The figure shows an object on a relatively uniform background to which the algorithm does respond, but the object can still be distinguished which seems a decent response. However in we look at figure 4.2 the algorithm does highlight the center of the circle but this is hardly noticeable given the surrounding responses. Which makes this algorithm not very suited as detector of symmetry.

4.6 Concluding

The isotropic symmetry, radial symmetry and color symmetry algorithm detect the symmetries in both the simple shapes (see figure 4.2) and in the photo of the wingnut. Although the color symmetry method adds sensitivity to color to the isotropic symmetry algorithm, both algorithms produce a similar response. The radial symmetry algorithm, which does not react to center lines of elonguated objects, (see figure 4.4) gives a different response than the isotropic and color symmetry algorithms. All three methods respond to image properties they are designed to respond to. The phase symmetry does detect symmetries, but produces artifacts at locations which can not be considered symmetric points. The simple symmetry measure produces the worst responses of all symmetry based algorithms. Simple symmetry does not seem to give any information about the symmetries present in an image. We conclude that isotropic symmetry, radial symmetry and color symmetry algorithms are the best detectors of symmetry.

Chapter 5

Experiments

The experiment collected data of human viewing to be able to compare the predictions of the algorithms to human data. Using a eye tracker we obtained the locations of fixations by humans on a set of images. In this chapter we will discuss the experiments and the methods used to analyse the results from the experiments.

5.1 Experimental setup

To register the eye movements of the human subjects, we used the EyeLink I from RS Research. Pictures were presented on a 19 inch display at a resolution of 1024x768. With an effective diameter of 450mm (360x270mm). The subjects viewed the image from a distance of circa 70cm. To track the eye movements of the participants an eyetracker helmet is mounted on the head of the participant (See figure 5.1). The helmet is fitted with three infrared cameras which work at a frequency of 250Hz. Two cameras are aimed at the eyes of the participants and serve to track the position of the pupils of the participant. The third camera watches four infrared lights attached to the corners of the display. A calibration session determines the position of these markers relative to the participants. The calibration consists of looking at a series of predetermined positions on the screen. This makes it possible to determine the locations of the fixations and to compensate for movements of the participants. The eyetracker records the eye movements and produces data with time and locations of the recorded fixations. The data was collected in two experiments. Both experiments used the same equipment and setup.

5.1.1 The freeview experiment I

Forty three students of the University of Groningen with normal or corrected-to-normal vision participated in the experiment. The participants paid attention to the presented images without a specific task. Hence the name *freeview experiment*. To encourage the subjects to keep investigating the images, they were told beforehand that they would have to answer several questions about the images after the experiment.

The images used are divided into three categories: mugs, objects and nature, each containing 10 pictures. The mugs category consists of white mugs without any print. The images from the objects category were common household items such as: pliers, flashlight, stapler, etc. The images from mugs and objects category were shot against an uniform background. The nature pictures consisted of different kinds of landscapes and detail pictures of natural objects like flowers or leafs. A table of all figures is can be seen in appendix A.

For every image, a mirrored version was included to compensate for possible lateral effects. Furthermore, every picture was presented twice. Every picture was presented for five seconds. The pictures were shown in random order to each subject. This way one can assume no influence of order and memory on the results. In total every subject was presented $3 \times 10 \times 2 \times 2 = 120$ pictures.

Every presentation of a picture by the eye tracker is preceded with a so called drift correction for a quick recalibration of the eyetracker. The eye tracker presents a dot in the center of the screen. The subject is asked to focus on the dot and press a key when the subject is certain his or her gaze is on the dot. When the drift correction is accepted by the eye tracker the next picture was presented.

5.1.2 The freeview experiment II

The subjects consisted of 31 students from the University of Groningen with normal or corrected-to-normal vision. For the second experiment we selected nature pictures with objects with symmetric features. The image set mainly consists of images of flowers, but also some leafs, and a butterfly were used. This gave a group of 19 images we called symmetry. Every image was presented once to every subject for a duration of five seconds. Again, every presentation of an image was preceded by a drift correction. A table of all figures is can be seen in appendix A.

Figure 5.1: Helmet of the EyeLink eye tracker used in the experiments.

5.2 Methods for analysis

The fixations on an image obtained from the human are compared to the saliency maps obtained from several computational algorithms. A metric was constructed to give us an idea to what extend a saliency map is capable of predicting the fixations made by the human subjects. We adopted two methods: a first method from Ouerhani et al. (2004) which is based on correlation and a second method from Parkhurst et al. (2002) which looks a the saliency value at the location of a fixation. Our aim is to compare the performance of the algorithms with each other. The comparison is done within one category of images.

Correlation measure

Given a certain image, the first step for the method by Ouerhani et al. (2004) is to transform the fixations of one trial, which consists of one presentation of a image to a participant, a so called fixation set, into a fixation map. This is done by adding a Gaussian function for every fixation. The Gaussian is centered at the position of a fixation. The standard deviation used for the Gaussian is determined by the size of the region which can be observed by the fovea. This is the size of the fovea in the visual field which is about 2°. For our setup this results in a fixation radius of about 34 pixels. To ensure the

values at the periphery of the fixation are nearly zero the standard deviation of the Gaussian was set to $\sigma = 12$. The fixation map, M, for are set of fixations, F, can be expressed as:

$$M = \sum_{i \in F} N(\mu_i, 12). \tag{5.1}$$

Where u_i is the location of the i^{th} fixation. Subsequently, the saliency map and the fixation map are normalized by making their total value sum to 1. Thereby, making a direct comparison between both maps possible. This comparison is done by calculation of the correlation coefficient between the maps according to the following formula:

$$\rho = \frac{\sum_x \sum_y [M_h(x,y) - \mu_h) \cdot (M_c(x,y) - \mu_c)]}{\sqrt{\sum_x \sum_y (M_h(x,y) - \mu_h)^2 \cdot \sum_x (M_c(x,y) - \mu_c)^2}}$$

Where μ_h and μ_c are the mean values of respectively the saliency map, M_h, from the fixation map, M_c. The resulting correlation value, ρ, represents the similarity between the saliency map and the fixation map. This process is repeated for every trial of every subject viewing an image of the same image category (e.g. nature). The resulting values are averaged which gives us a measure of the performance of an algorithm on a specific category.

Besides the comparison between the fixation methods we also made three other comparisons. We call these three measures: *random*, *inter subject* and *intra subject*. Random is the correlation with randomly generated human fixation maps. This is the average of correlations of all fixation sets with random sets of fixations. This results in a correlation of zero and gives us the bottom line. A method which performs equal to random has no predictive powers. The self measure indicates to what extend the subjects predict each others fixations on the same image. This is achieved by taking all the fixation maps from one image and comparing them by correlating all fixation maps with each other. This is done for every picture in one category (e.g. mugs). All the results of the correlations are averaged and this results in the value of the inter subject measure. We expect a subject to be a reasonably good predictor of another subject assuming that the visual information processing proceeds in a similar manner for every subject and visual attention is for a significant part data-driven. If each person would process visual information completely different or the processing would not depend on the contents of the image this measure would be equal to random, but we do not expect this to be the case. Finally, we calculated the intra subject measure. This measure should give us an idea to what extend a

subject predicts him or herself. Every subject has viewed the images four times except for the symmetry images which are viewed only once. Given an image, the fixation sets from one subject are correlated with each other. Per category, except for the symmetry images, these values are averaged to give us the intra subject measure. We would expect this to be the top-line of the predictions, because the data-driven or bottom-up component should not change between trials and the top-down component, which is by definition influenced by past experiences of the subject, should vary less within subjects than between subjects.

Fixation saliency measure

The second method, based on a method from Parkhurst et al. (2002), takes a different approach. The idea is to look the saliency maps generated by the algorithms and see if they predict higher than average saliency at the locations of the fixations made by the participants in the experiment. For a given picture a saliency map is generated with a algorithm. A measure for the quality of the algorithm is constructed by extracting the values of the saliency map at the locations of the human fixations. This is done by considering an area centered at the location of the fixation with the size of one fixation. The area has a radius, r, of about 34 pixels. Subsequently, the values in the area are weighted with a 2 dimensional Gaussian, N, with $\sigma = 34/3$ and average μ located at the position of the fixation. The Gaussian matrix sums to one and returns a weighted average of the patch under consideration. We will call this value the saliency value. The function s to extract the saliency value of a fixation f with saliency map M can be can be expressed as:

$$s(f_x, f_y) = \sum_{i=0}^{2r} \sum_{j=0}^{2r} [N(i-r, j-r, \sigma) M(f_x - r + i, f_y - r + j)] \qquad (5.2)$$

The resulting saliency value of the fixation is normalized by dividing it by the average saliency, $\bar{s}$ obtained by randomly sampling the saliency map a 1000 times in the same manner with which the saliency value of the fixations are extracted. This finally results in the following expression for the fixation saliency value:

$$s_f = \frac{s(f_x, f_y)}{\bar{s}}. \qquad (5.3)$$

If s_f is bigger than 1, it means that according to the saliency map the location of the fixation is more salient than the saliency of random fixations. To determine the performance of a algorithm within a certain image category the procedure is repeated for every fixation of every trial of every subject

for every image within one image category. The value is averaged over the number of fixations resulting in a measure of average predicted saliency for one category of images. This give a indication to which extend an algorithm is able to predict human fixation. about human fixations. on a given category. The base-line is the average saliency obtained by randomly sampling the saliency map.

Using the correlation and the fixation saliency method we hope to gain insight in the performance of the different algorithms used to construct the different saliency maps.

Chapter 6

Results

In this chapter we will present the result of the analysis done on the data obtained from the two experiments discussed in section 5.1. First, we will present the results of the analysis with the two measures: correlation and fixation saliency as discussed in section 5.2. We will use these results to discuss the ability of the different methods to predict the human fixations maps. Second, we will make a separate comparison with the methods from the article by Privitera and Stark (2000). Third, we will have a look at some properties which are interesting to humans and look if and how these relate to properties that the algorithms respond to.

6.1 Comparing the fixation predictors

In this section we will use the methods discussed in section 5.2 to test the performance of the different methods. Furthermore, we will make a separate comparison with the methods found in the article by Privitera and Stark (2000) where we will only use a single scale to construct the saliency map.

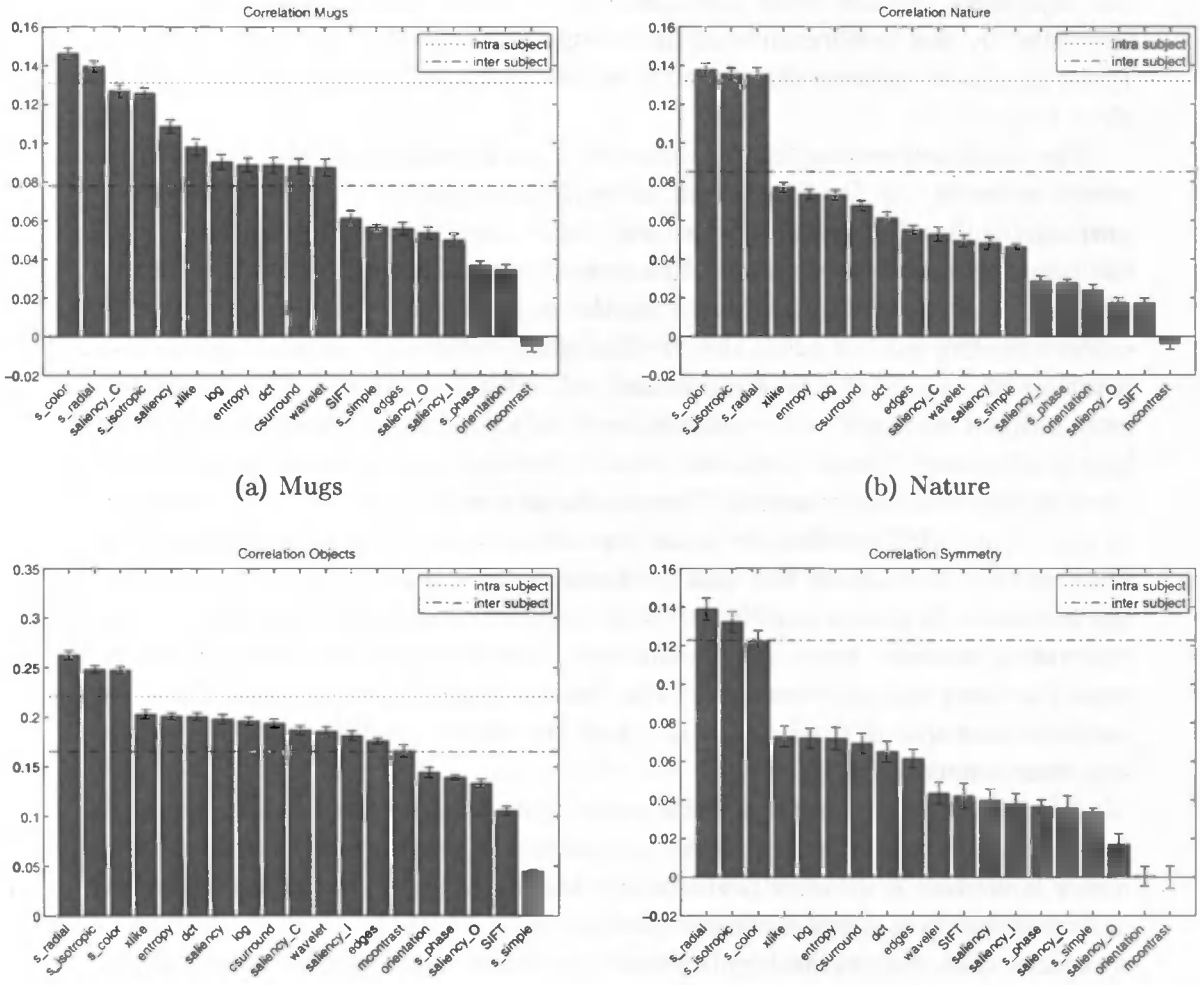
6.1.1 Correlation between experiment and prediction

In this section we used the correlation measure to compare the predictions of the methods with the human data. With this measure, we can compare the different prediction methods with each other. We compared the methods within the same image category, to see if the performance of the methods depend on the category. For example the images of the mugs contain very little color compared to the object and nature, symmetry categories. While the mugs and object categories consist of images of a single object on a uniform background, the nature and symmetry categories contain images

with much more structure. The nature and symmetry categories contain images of objects on texture rich background or images of landscapes which both contain structure than a photo of a object. All algorithms as discussed in chapter 3 were used in this comparison. The results of the correlation measure are presented in figure 6.1.

In the mugs category the color symmetry method is the best performing measure. Closely followed by the radial symmetry method. Both perform significantly better than the intra subject score. The next position is occupied by the color channel of the saliency model with in fourth position the isotropic symmetry method. The phase symmetry method performs poorly with a seventeenth place. The saliency model at the fifth place performs significantly worse than the four best methods, but still performs better than the inter subject measure. As mentioned earlier, the color channel of the saliency model exhibits a good performance, but the channels based on orientation and intensity perform far worse with a fifteenth and sixteenth place respectively. The first to perform worse than the inter subject measure is the SIFT algorithm at the twelfth place. The algorithms from the article by Privitera and Stark (2000) can be divided into two groups. One group that performs better than the inter subject score and a second group that performs worse than the inter subject score. From the first group the xlike method on the fifth position gives the best performance followed by the Laplacian of Gaussian (LoG), entropy, discrete cosine transform (DCT), the center surround (csurround), and wavelet method. The latter five methods show no significant difference in performance. The second group performs significantly worse than the intra subject score and starts at the thirteenth position with the simple symmetry method followed by the edges, orientation, and, as worst performer, the Michaelson contrast (mcontrast) method.

For the nature category the first thing that attracts attention is the performance of the three symmetry methods isotropic, radial, color symmetry. All these symmetry methods have no significantly different score and all methods score better than the intra subject score. The fourth place occupied by the xlike method which performs far worse than the top three and also performs worse than the inter subject method. Just as in the mugs category the color channel is the best performer of the saliency model, but with a tenth place no where near the best scoring methods. Again the orientation and intensity channel perform worse and this results in a twelfth place for the saliency model. The SIFT algorithm, with its second to last position, performs very poorly. The methods from the article by Privitera and Stark (2000) all perform worse than the inter subject score. The best performing algorithm is again the xlike method on the third place followed by the entropy, DCT, edges, wavelet, and simple symmetry showing a slight



(a) Mugs

(b) Nature

(c) Objects

(d) Symmetry

Figure 6.1: Results of the analysis by correlation. See section 5.2 for a detailed explanation. For every image category a bar plot was constructed. The average correlation was plotted for every method in descending order. Furthermore, horizontal lines are plotted for the inter subject and intra subject measures. The random measure omitted as it was zero or close to zero as expected. The error bars give 95 percent confidence interval. For the flower group the intra measure is not calculated, because the symmetry group were only viewed once by each subject. A description of the methods color symmetry (s_color), radial symmetry (s_radial), isotropic symmetry (s_isotropic), simple symmetry(s_simple), symmetry from phase, (s_phase), saliency model consisting of: the combined channels (saliency), contrast channel (saliency_C), intensity channel (saliency_I), and orientation channel (saliency_O), Laplacian of Gaussian (LoG), xlike, entropy, discrete cosine transform (DCT), center surround (csurround), wavelet, SIFT, edges, orientation and, Michaelson contrast (mcontrast), can be found in chapter 3.

but significant decline from the third to the tenth place. This trend is interrupted by the performance of the orientation and the mcontrast method. These methods perform significantly worse with a sixteenth and nineteenth place respectively.

The best performing method for the object category is the radial symmetry measure. As for the nature category the symmetry measures: radial, isotropic, and color symmetry are the three best performing methods, but the phase symmetry occupies a fourteenth place. In contrast to the nature category the vast majority of the methods performs better than the inter subject measures. Not until the twelfth place, which is occupied by the orientation channel of the saliency model, all methods perform better than the inter subject measure. The color channel of the saliency model is again the best performing channel, but this time is does not perform significantly better than the intensity channel. This results in a seventh place for the saliency model. The SIFT method is again second to last of the 19 methods. The xlike method was again the best performing method for the methods from the article by Privitera and Stark (2000) followed by the entropy, DCT, LoG, csurround, wavelet, edges and mcontrast. All these methods perform better than the inter subject measure. The first to perform worse than the inter subject measure, is the orientation and the worst performer is the simple symmetry method.

For the symmetry category the radial symmetry method is the best performer. Again the three symmetry measures radial symmetry, isotropic symmetry, and color symmetry perform almost the same, but this time the radial and isotropic symmetry methods perform just above the inter subject measure and the color symmetry method performs equal to the inter subject measure. The phase symmetry method takes the fourteenth position. From the different channels of the saliency model, the channel based on the intensity channel gives the best performance. Followed by the color channel which does not perform significantly worse. The channel based on orientation information is yet again the worst performer. The three channels combined can be found at the twelfth place. The SIFT method occupies a mediocre eleventh place. Based on their performance we can divide the methods from Privitera and Stark (2000) into two groups. The first group which consist of the xlike, LoG, entropy, csurround, DCT, and edges method which all have comparable performance. The xlike method is again the best performer. The second group consists of the wavelet and the simple symmetry methods. The third group is comprised of the orientation and the mcontrast method which give the worst performance.

Overall the performance of the three symmetry methods radial, isotropic and color is very good compared to the other methods. They give the best

performance in all categories. As discussed in chapter 4 these measures are variations of one algorithm and therefore respond to very similar properties. Between these three symmetry models the radial symmetry model has the overall best performance.

The symmetry methods all perform on par with the intra subject measure but the other methods perform a lot worse. Contrastingly, the phase symmetry and simple symmetry methods perform consistently bad in all categories. The performance of the saliency model is in all cases worse than the best performing symmetry algorithms. The saliency model performs reasonably well on the mugs and objects category, but mediocre on the nature and symmetry category. Regarding the separate channels of the saliency model, the channel based on color information gives the best performance. For all categories, except the mugs category, the channel based on orientation information gives the worst performance. The channel based on intensity information performs on par with the color channel in the objects and symmetry category, but worse in the mugs and nature category. The SIFT method performs bad in all categories. Looking at the methods from Privitera and Stark (2000) the xlike method is the strongest performer in all cases, although entropy, LoG, DCT, csurround show a similar performance in all categories. These methods in this group generate saliency map that are visually very similar. This is shown in figure 6.2. Especially the Laplacian of Gaussian and the center surround method produce similar saliency maps this is due to the almost identical shape of the kernels used. Which can be seen in 3.10. The symmetry methods isotropic, radial and color symmetry also have a similar performance. This is expected because these methods are variations on the same algorithms, but to what extend do they produce the same saliency maps?

To quantify the similarity between the methods, we looked at the correlation between the saliency maps generated by the methods. As seen in table 6.1 the symmetry measures have very high correlations with each other and the correlation between the center surround and Laplacian of Gaussian is even higher. Furthermore, there are high correlations between the methods in the group with xlike, LoG, DCT and, csurround. With the exception of the correlation between the xlike and DCT. The high correlation explains the very similar scores. Especially the symmetry methods, which according to the correlations, are almost interchangeable as saliency predictor.

6.1.2 Fixation saliency

A second comparison was made with the fixation saliency method. Again the method was calculated within each image group. As explained in section 5.2,

(a) Mugs

	s_isotropic	s_color	s_radial	xlike	LoG	csurround	DCT
s_isotropic	–						
s_color	0.87	–					
s_radial	0.75	0.81	–				
xlike	0.61	0.47	0.37	–			
LoG	0.68	0.48	0.33	0.78	–		
csurround	0.65	0.46	0.30	0.78	0.99	–	
DCT	0.51	0.37	0.21	0.70	0.82	0.84	–

(b) Nature

	s_isotropic	s_color	s_radial	xlike	LoG	csurround	DCT
s_isotropic	–						
s_color	0.95	–					
s_radial	0.95	0.92	–				
xlike	0.53	0.49	0.50	–			
LoG	0.56	0.50	0.50	0.63	–		
csurround	0.53	0.47	0.47	0.62	0.98	–	
DCT	0.37	0.34	0.35	0.53	0.63	0.64	–

(c) Objects

	s_isotropic	s_color	s_radial	xlike	LoG	csurround	DCT
s_isotropic	–						
s_color	0.94	–					
s_radial	0.87	0.87	–				
xlike	0.62	0.59	0.60	–			
LoG	0.69	0.64	0.59	0.78	–		
csurround	0.69	0.63	0.57	0.77	0.99	–	
DCT	0.63	0.60	0.56	0.72	0.83	0.83	–

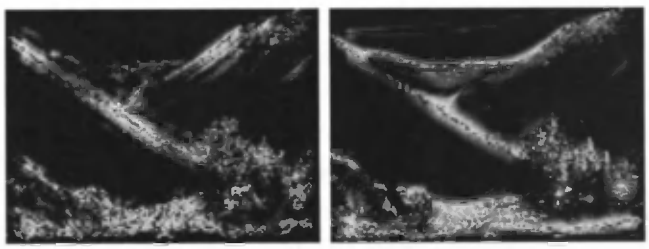
(d) Symmetry

	s_isotropic	s_color	s_radial	xlike	LoG	csurround	DCT
s_isotropic	–						
s_color	0.92	–					
s_radial	0.93	0.89	–				
xlike	0.49	0.43	0.47	–			
LoG	0.56	0.47	0.50	0.67	–		
csurround	0.54	0.44	0.47	0.65	0.99	–	
DCT	0.35	0.29	0.32	0.55	0.67	0.68	–

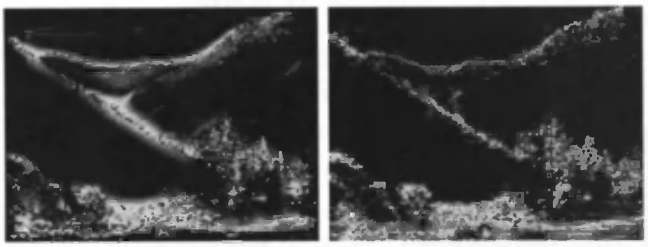
(e) All

	s_isotropic	s_color	s_radial	xlike	LoG	csurround	DCT
s_isotropic	–						
s_color	0.92	–					
s_radial	0.89	0.87	–				
xlike	0.55	0.48	0.48	–			
LoG	0.61	0.51	0.48	0.71	–		
csurround	0.59	0.49	0.46	0.70	0.99	–	
DCT	0.44	0.38	0.35	0.61	0.73	0.74	–

Table 6.1: Results of the correlations between several methods. A pairwise comparison between the methods was performed by calculating the correlation between the saliency maps produced by the two methods for a certain image. This was done for all images in the mugs, nature, objects, and symmetry categories and the results were averaged. The last table is an average for all categories. All results can be found in appendix B.



(a) xlike (b) csurround

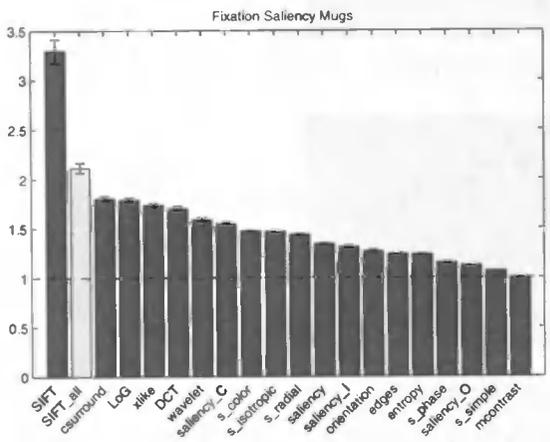


(c) LoG (d) DCT

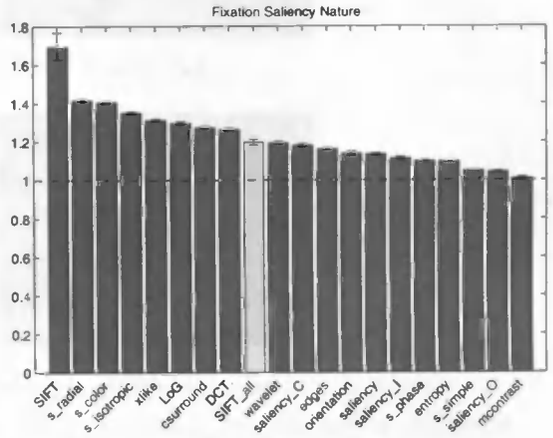
Figure 6.2: The saliency maps generated by the xlike method (a), the csurround method (b), Laplacian of Gaussian method (c), and DCT method (d) are visually very similar. The source image can be seen in figure 6.9.

for every fixation made within one of the image categories a value is computed with the fixation saliency method. The averaged of these values gives us a measure of performance for a certain group. The results are presented as the magenta bars in figure 6.3.

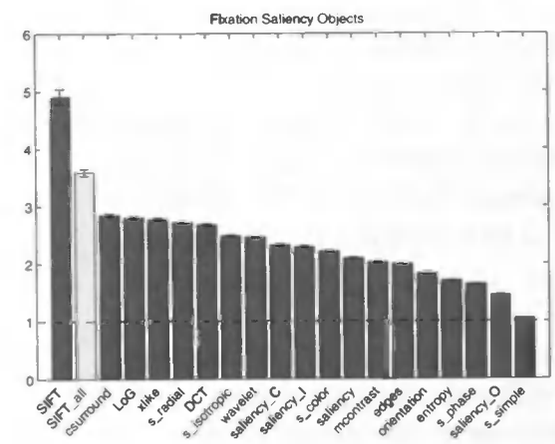
For the mugs category the SIFT method is the best performer by a large margin. The runner up is the center surround method and in the third position the LoG method, which does not differ significantly from the center surround method. This is expected because of the similarity between both methods. Both are methods from the article by Privitera and Stark (2000). The symmetry methods: color symmetry, isotropic symmetry, and radial symmetry are positioned at the eighth, ninth, and tenth place respectively. The color symmetry method is the best performer of the symmetry methods, but compared to the other methods the performance is average. The symmetry from phase algorithm at the sixteenth position gives a poor performance. Looking at the saliency model we see that from the separate channels the color channel, at a seventh gives the best performance. The intensity channel can be found at a eleventh place and the worst performer, the orientation channel, can be found at the seventeenth position. Overall



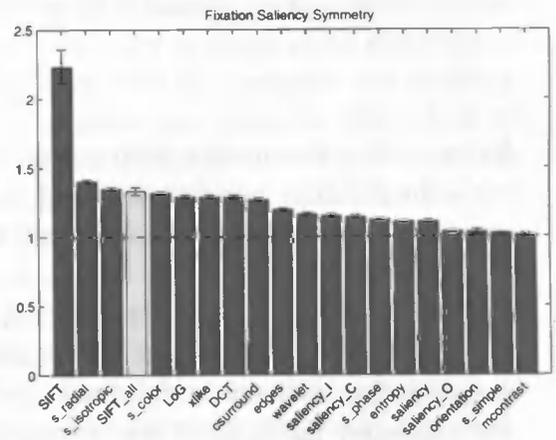
(a) Mugs



(b) Nature



(c) Objects



(d) Symmetry

Figure 6.3: Results of the analysis with fixation saliency. The horizontal line gives the line which indicates random performance. The yellow bar (SIFT_all) is a variant of the SIFT algorithm (See text).

the saliency model is located at the eleventh position.

When we look at the nature category we again see that the SIFT method is the best performer. The runner up is the radial symmetry method followed by the color symmetry and isotropic symmetry. The symmetry from phase method performs poorly with a fifteenth position. From the methods from Privitera and Stark (2000) the xlike method at the fourth place is the best performer. Followed by the LoG and the center surround method. From the saliency model the color channel, at the tenth position, is the best performing channel. The orientation channel is yet again the worst performer of the channels. The combined channels take a thirteenth position.

For the objects category the SIFT method is also the best performer. The second position is for the center surround method, followed by the LoG and the xlike method. The best symmetry measure is the radial symmetry method at the fifth position. Isotropic symmetry can be found at the seventh position and color symmetry at the eleventh position. In contrast to the other categories the symmetry measures are not grouped together. The phase symmetry algorithm can be found at a seventeenth position. The combined performance of the saliency model is positioned at a twelfth place. The color channel has the best performance, but does not perform significantly better that the intensity channel. Again, the orientation channel is the worst performer.

For the symmetry category the SIFT method is the best performer. The second position is occupied by the radial symmetry method followed by the isotropic symmetry and color symmetry method. The LoG method, at the fifth position, followed by xlike, DCT, and center surround method. These are the best performers of the methods from Privitera and Stark (2000). The intensity channel from the saliency model gives the best performance of the different channels of the saliency model, but the color channel does not perform significantly worse. The orientation channel is again the worst performer. Overall the saliency model occupies the fifteenth place.

What stands out is the score of the SIFT algorithm. In every category, this algorithm is the best performer by a considerable margin. If we look at the symmetry methods: radial symmetry, isotropic symmetry, and color symmetry perform better on the nature and symmetry categories and perform worse on the object and symmetry categories. The saliency model is in all categories below average. From the separate channels the color channel is the best saliency predictor, although the intensity channel performs equally on the objects and symmetry categories. The orientation channel is the worst performing channel in all categories.

Overall the best performing method from Privitera and Stark (2000) is the xlike method. Followed by the center surround and LoG methods. These

	# keypoints
Mugs	47
Nature	270
Objects	84
Symmetry	248

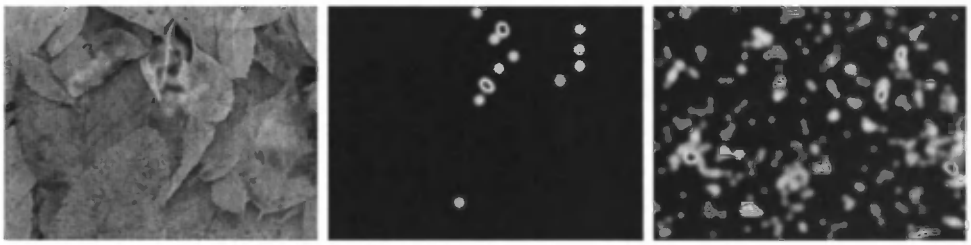
Table 6.2: # keypoints gives the average number of keypoints available per image in a image category.

methods perform better than the symmetry methods on the mugs and the objects category.

A closer look at SIFT

We will now take a closer look at the high scores of the SIFT algorithm. To construct a saliency map from the SIFT method we can use a different number of keypoints. As described in section 3.2 we used the average number of fixations a subject made given a certain image. As can be seen in figure 6.4 this results in a saliency map with very few points. In the saliency map generated with all available keypoints the saliency is much more distributed over the entire image. When we use all keypoints to construct the saliency map the values of the fixation saliency measure are closer to the values of the other methods. The results are plotted as the yellow bars (SIFT_all) in figure 6.3. Compared to the results in the SIFT method the result for the mugs category is still the biggest value but it is a lot closer to the runner, up the center surround method. If we look at the nature group the new score will put SIFT at a ninth place. For the objects category it is still the best performer and for the symmetry category SIFT will occupy a third position. In table 6.2 we give the average number of available keypoints per image in a certain category which gives an indication how much keypoints are used. The fact that the average number of keypoints from the mugs and objects group is considerably lower than the average number of keypoints for the nature and symmetry category is the could be the cause that the SIFT_all methods are still at the first position.

To find out what the influence was from different numbers of keypoints we generated saliency maps from the nature images with the SIFT method with different numbers of keypoints and compared them with the human data using the fixation saliency method. We choose the nature category, because the SIFT method produces the largest number of keypoints with images from this category. This gives us the largest range of keypoints to



(a) (b) (c)

Figure 6.4: Human fixation map (a) of the image used generate a saliency map generated with the SIFT algorithm. One with the 14 best keypoints (b) and one with the 333 best keypoints (c).

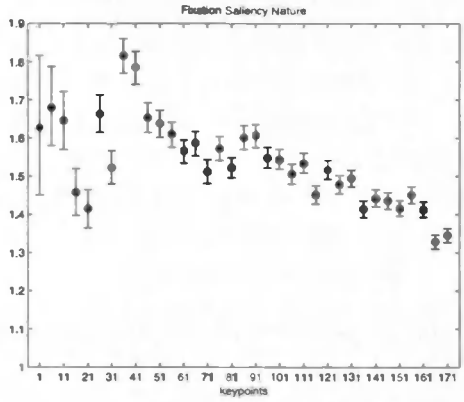


Figure 6.5: Fixation Saliency measure calculated for the SIFT method with different numbers of keypoints. The keypoints were ordered by their quality according to the SIFT method and each time a different number, plotted on the x axis, of the best keypoints was selected.

compare. The keypoints were ordered by their saliency according to the SIFT method and each time a different number of the best keypoints was selected. The results can be seen in figure 6.5. Although the measure shows great variation for low numbers of keypoints the general trend seems to be reduced scores for the saliency extraction measure with higher number of keypoints. This could be due to one of two causes. First, the fixation saliency measure is sensitive to the total amount of saliency present in a saliency map and the normalization method used does not compensate for this. Second, the first SIFT keypoints give a better prediction of the human fixations.

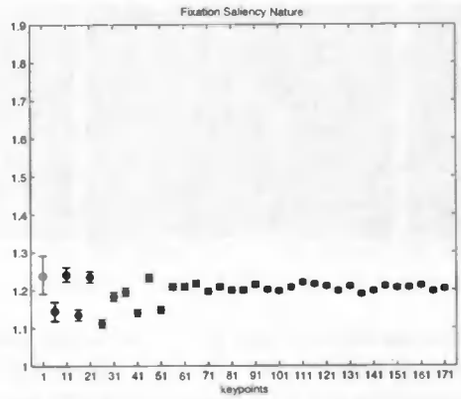


Figure 6.6: Fixation Saliency measure calculated for SIFT method with different numbers of keypoints. Each time a number, plotted on the x axis, of unique keypoints was selected at random from the keypoints available for a certain image.

To find out which of the possible explanations is correct we repeated the evaluation, but now by not using the best keypoints. Instead we selected unique random keypoints from all available keypoints. To compensate for random selection effects, we calculated the fixation saliency ten times for each number of keypoints and averaged the results. The results are shown in 6.6. For all number of keypoints the scores are about the same. Indicating that the decrease in saliency score in the figure 6.5 is due to the lower average quality of the selected keypoints and not only based on the number of keypoints used. This shows that best keypoints according to the SIFT algorithm are good predictors of fixations. Therefore, we can conclude that the fixation saliency measure has a bias for conservative estimates. A few good keypoints, or a more specific prediction, will give a better score than a larger number of points with an average lower quality. This is caused by the normalization of the fixation saliency value with the average fixation saliency obtained from a set of random fixations (see section 5.2). If SIFT predicts a few very salient parts it will perform better than if SIFT also predicts keypoints at less salient parts. This will also hold for other methods. Although this is not necessarily a problem it introduces a extra choice, the number of keypoints to use, which makes the SIFT algorithm harder to compare to the other methods.

Method	Summated rankings
Radial symmetry	26
Isotropic symmetry	33
Color symmetry	36
Xlike	36
LoG	43
Center Surround	52

Table 6.3: Summated rankings of the best scoring methods on all categories and both evaluation measures.

6.1.3 And the winner is ...

To determine the overall performance of the methods we summated the rankings of the methods on all categories and both evaluating methods. The method with the lowest score will be considered the best overall performer. Although this is a very crude method it gives some indication of the overall performance. We will present the five best performing methods. The lowest score, and thereby best performance, is for the radial symmetry method. The radial symmetry performs good when we evaluate the results with the correlation method and performs decent when we use the fixation saliency method. The second performing method is the isotropic symmetry method. This method performs good when we use the correlation method for evaluation and somewhat worse when we use the fixation saliency method. The third position is split between the color symmetry method and the xlike method. The fourth place is for the Laplacian of Gaussian. Not present in the top five is the saliency model which never performs above average. Finally we should mention the SIFT method which performs very bad when we use the correlation evaluation method, but is the best performer if we evaluate the performance with the fixation saliency method. This shows that the outcomes as discussed here should be taken purely as an indication of performance and not as a definitive ranking.

6.1.4 Comparison with Privitera and Stark

The algorithms described in section 3.4 are all from the article by Privitera and Stark (2000). We would like to compare our findings to their results. We are not able to provide a direct quantitative comparison with the article by Privitera and Stark (2000), since the metric used by Privitera and Stark (2000) to compare the artificial and human ROIs is not comparable

to our evaluation method. However, we can say something about the qualitative conclusions reached by Privitera and Stark (2000). Contrary to our approach, see section 3.5, Privitera and Stark did not use multiple scales to construct their saliency maps. To be able to make a more accurate comparison, we performed a new evaluation where we only used the first level of the saliency maps generated by the algorithms. Subsequently, we calculated a new comparison with the human fixations with the correlation and fixation saliency method. The results are shown as the magenta bars in figure 6.7 and figure 6.8.

Privitera and Stark found a high performance of the wavelet algorithm and overall poor performance of the DCT algorithm. They furthermore noticed that the Michaelson contrast algorithm performed well on their images of terrains, such as a martian landscape. We do not see these findings in our metrics. The overall best performer in the comparison by correlation is the the entropy measure and not the wavelet algorithm. The DCT algorithm is not the best performing algorithm but certainly not the worst.

The results obtained from the fixation saliency method do not give a clear picture of the individual difference in performance of the methods. Only the results from the objects group, and to lesser extend the symmetry group, show a ranking in the results. The only overall observation is the poor performance of the Michaelson contrast method in the mugs, nature and symmetry group, and the poor performance of the simple symmetry and orientation measures in all groups. As the Michaelson contrast method is based on intensity, the better performance of the method on the objects is due to the fact that the dark or light areas were always located on the object. This results in a positive correlation with the fixations, because people tend to only look at the objects. For the images from the nature and symmetry group contrast rich areas were not solely located on prominent objects, but also on the background or other details. For the mugs the high contrast areas were primarily the shadows at the base of the mugs which are not very interesting areas to fixate. The differences between our findings and the findings by Privitera and Stark (2000) can be attributed to the different metric that was used to compare the human and artificial ROIs. Another factor is probably the simple fact that we used different images than Privitera and Stark (2000).

Comparing single with multiple scales

We can also compare the single scale results with the multi scale performance calculated earlier (see figure 6.1 and figure 6.3) which are plotted again in the figure 6.7 and figure 6.8 as the yellow bars. If we look at the correlation

measure we see that all methods perform worse when only a single scale is used. An exception is the score of the entropy on the symmetry group which shows no significant difference in both cases. The performance for the mugs and object categories benefit the most from the use of multiple scales. With the fixation saliency measure the scores when we used a single scale are also generally lower. Again the benefits of using multiple scales are more pronounced for the mugs and object category. This suggests that the usage of multiple scales coheres better with the human visual system than using a single scale especially if an image contains very few small details, e.g. busy background textures, like the images from the nature and symmetry categories.

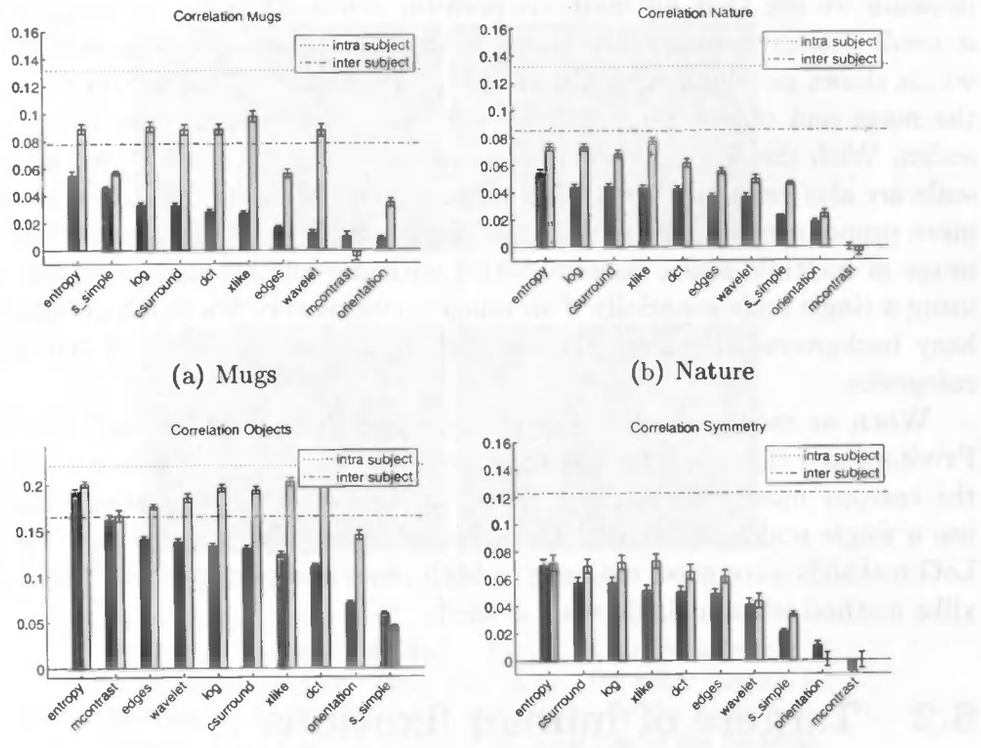
When we used multiple scales the best performer from the methods from Privitera and Stark (2000) was the xlike method. This role is taken over by the entropy method which is the best performer in all categories when we use a single scale to construct the saliency maps. The center surround and LoG methods have good rankings in both cases and perform better than the xlike method when a single scale is used.

6.2 Targets of human fixations

We will now look at the fixations to say something about the type of points humans find interesting. To be able to qualitatively discuss the results of the experiment we plotted a Gaussian function ($\sigma = 12$) for all fixations made by the participants in response to a certain picture. The Gaussian represents the area that is approximately covered by the fovea. An example can be seen in figure 6.9. This gives us a general idea about the interesting fixation points for human participants.

A first observation is that the subjects do not focus on uniform areas present in the pictures. The fixation pattern over the pliers as seen in 6.9 is exemplary of the way subjects fixated on images of the objects and mugs group. Nearly all fixations were on the object and none were on the background. Which shows humans do not use a random sampling strategy to process a visual scene but seek out interesting points.

Striking for the mugs group that almost no attention was given to the edge opposite to the ear of the mug. Commonly the ear of the mug was fixated. Furthermore, spots of dirt on the mugs are also guaranteed to attract attention. The spots of dirt are characterized by dark spots so we could conclude that this is due to high contrast difference. Indeed, highlights with a large intensity are also fixated. Two examples of fixations on a mug can be seen in figure 6.10.



(a) Mugs (b) Nature

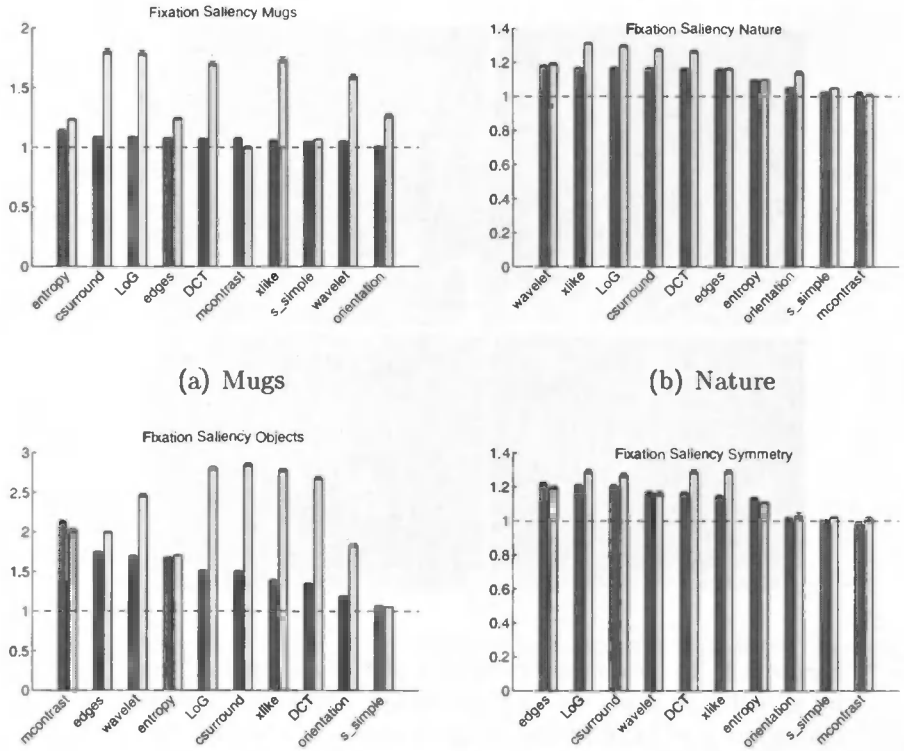
(c) Objects (d) Symmetry

Figure 6.7: Results of the analysis by correlation done on the saliency maps constructed from a single scale with the algorithms from the article by Privitera and Stark (2000) (magenta bars) together with the multi scale results (yellow bars).

What can be seen in pictures of symmetry in the nature and symmetry groups was the preference of the subjects to scan the center of flowers. As shown in figure 6.11 the subject extensively focused on the center of the flowers. Even the centers of the painted flowers on the bucket from the object category receive a lot of attention.

Color can also be decisive factor to draw attention. A photo which is almost entirely filled with the same flowers shows a concentration of fixations on the flower which deviates in color from the other flowers. As can be seen 6.12 the flower which is slightly more orange receives more attention than for example the flower above and to the right. Although this flower is slightly bigger the flower with the deviating color is fixated more frequently.

Just like color orientation is also a basic feature which attracts attention.



(a) Mugs (b) Nature

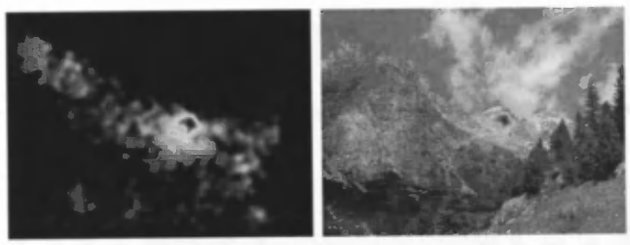
(c) Objects (d) Symmetry

Figure 6.8: Results of the analysis by fixation saliency done on the saliency maps constructed from a single scale with the algorithms from the article by Privitera and Stark (2000) (magenta bars) together with the multi scale results (yellow bars).

Although the methods based on orientation perform very poorly, we do see interest of humans for parts of the image where lines of different orientations cross each other. What seems to attract attention are what we will call *structural centers*. These are locations were different structures, such as the handles of the pliers in figure 6.9, come together. Other examples can been seen in figure 6.13 subjects attracted to places where nerves of the leaf come together. Furthermore, a concentration of fixations can also be found at the point were the stems of the leaves join. Moreover, the location were the stems of the flowers converge is also interesting for the human subjects.

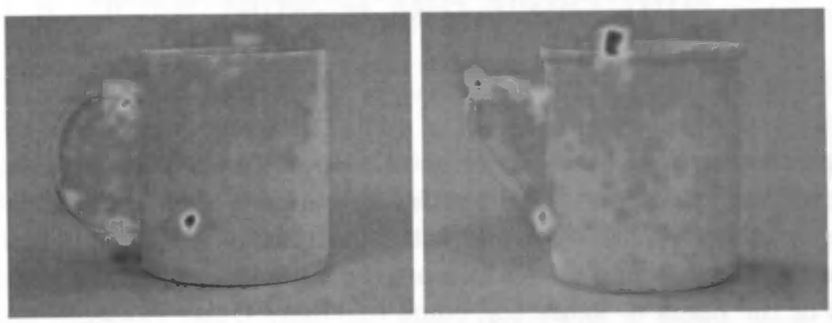
Although we offered some properties which were of interest to human observers, it is hard to find obvious properties and hard rules for the kind of locations which will be fixated.

(a)



(b)

Figure 6.9: Two examples (a,b) of a plot of all fixations made by all subjects when viewing the image. For every fixation a Gaussian, $\sigma = 12$, centered at the location of the fixation, is plotted. To give a better idea of the locations of the fixations they are plotted over the image.



(a) (b)

Figure 6.10: Two examples of fixations in response to a mug. The ears and the top are fixated and a spot of dirt on the front of the mug (a) an a highlight at the top of the mug (b) catch the attention of the observers.

Figure 6.11: Examples of the tendency of subjects to focus the center of flowers.

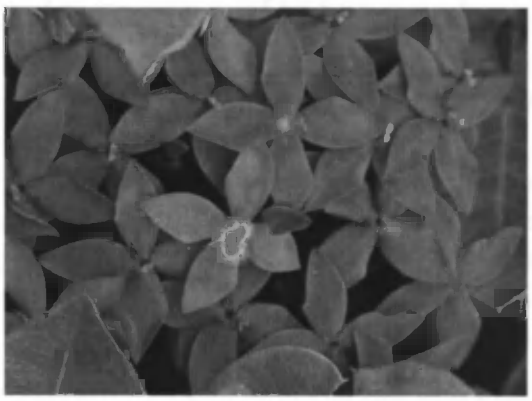
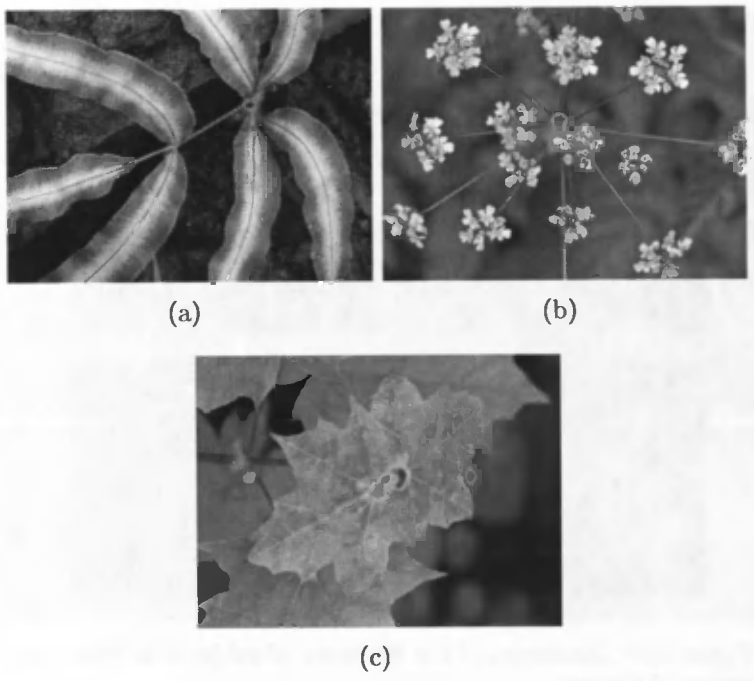


Figure 6.12: Example of the influence of color. The flower with the different shade of red receives the most attention.



(a) (b)

(c)

Figure 6.13: Examples of the tendency of subjects on *structural centers*. Fixations are concentrated at the location were the stems meet (a,b), and extra attention is given to locations on the leaf were the nerves meet.

6.3 A closer look at the algorithms

We will have a closer look at the properties of the methods as described in chapter 3 and how they relate to the interests of human subjects. In table 6.4 we present some characteristics of all used methods, but in this section we will only discuss the saliency model and to the symmetry methods. The saliency model receives extra attention because it is the best known model to construct saliency maps and the symmetry methods because they have our specific interest.

One of strong points of the isotropic, radial and color symmetry methods is that they respond to the center of the flowers. This tendency coheres well with the fixations pattern of the human subjects. As illustrated with figure 6.14 the saliency model gives a more fragmented saliency map with no clear center, whereas the color symmetry algorithms produce a saliency map where the saliency is located at the center of the flower and the saliency decreases towards the periphery of the flower. This is caused by the fact that

Method	Description	
	Good	Bad
SIFT	• –	• Does not pick up center of flowers. • To few keypoints taken into account.
the saliency	• Reacts to the ear of the mug. • Reacts to the top edge of the mug. • Reacts to highlights and dark spots.	• –
saliencyI	• Reacts to highlights and dark spots.	• Strong reaction to shadow with the mugs.
saliencyC	• Reacts to ear of the mug. • Reacts to the top edge of the mug. • Reacts to highlights.	• Takes color into account.
saliencyO	• –	• Produces a halo effect. • Does not react to fine structure. • Does not respond to center of flowers.
s_isotropic	• Reacts to highlights and dark spots. • Reacts to ear of the mugs. • Reponds to center of flowers.	• Reacts center axis of elongated objects. People only tend to focus at the extremities of such objects. • Produces reactions in uniform areas. • Reacts to shadows of the mugs. • Reacts with undefined blur if images consist of a uniform but busy images such (see nature images).
s_hcolor	• See s_isotropic. • Takes color into account.	• See s_isotropic. • Reacts to the center of mugs.
s_rradial	• Does not respond to shadows of the mugs. • Only reponds extremities of elongated objects and not to the central axis.	• Produces reactions in uniform areas. • Reacts to the center of the ears of the mugs.
s_phase	• –	• Produces halo and spurious reactions. • Produces reactions in uniform areas. • Does not react to large structures.
s_privsym	• –	• Reacts to uniform areas. • Does not react to large structures.
xlike	• Responds to highlights and spots of the mugs. • Responds to shadow of top edge of the mugs. • Reacts to the extremities of object_02 and object_03.	• Does not react to large structures. • Sensitive to orientation of structure so same structure with different orientation will give different response.
wavelet	• Does not react to right edge of the mug.	• Does not react to horizontal or vertical structures. • Does not react to large structures.
csurround	• Responds to highlights and spots of the mugs.	• Reacts to shadows at the bottom of the mugs.
orientation	• –	• Halos around edges.
edges	• Reacts to highlight of mugs.	• –
entropy	• –	• Reacts with undefined blur if images consist of a uniform but busy images such (see nature images).
mcontrast	• –	• Only reacts to parts of the image which is lighter or darker than the average of the whole picture. • Only reacts strongly to the shadows in the mugs.
DCT	• Reacts to highlights and spots.	• –
LoG	• see csurround.	• see csurround.

Table 6.4: Listing of the methods and notable properties of the saliency maps produced by the methods compared to the human fixations.

(a) (b) (c)

Figure 6.14: Shown are the human fixation maps (a), the saliency maps produced by the saliency model (b), and the color symmetry method.

color symmetry saliency map can give a better representation of the overall structure of the image than the saliency model. As described in section 3.5 the saliency map is constructed from five scales. In the fifth scale the dimension of the image is reduced by a factor 16. This enables the symmetry algorithms to detect large symmetries present in the images. Figure 6.15 shows the feature maps generated for three images at the coarsest level. The saliency coincides with the centers of the flowers and the symmetry axis between the handles of the pliers. This illustrates that the color symmetry algorithm reacts better to large structures in the images than the saliency model. Although this property will not always cohere with human tendencies, as can be seen in the example of the pliers in figure 6.15, it does work really well for the flowers.

As explained in chapter 4 the radial symmetry method differs from the color and isotropic symmetry methods in such a way that it does not respond to center of elongated objects such as rectangle. Radial symmetry only reacts to the extremities of these objects. The same tendency can also be found in the human data. We will illustrate this tendency with figure 6.16. For example the grip of the pliers is very pronounced in the saliency map generated

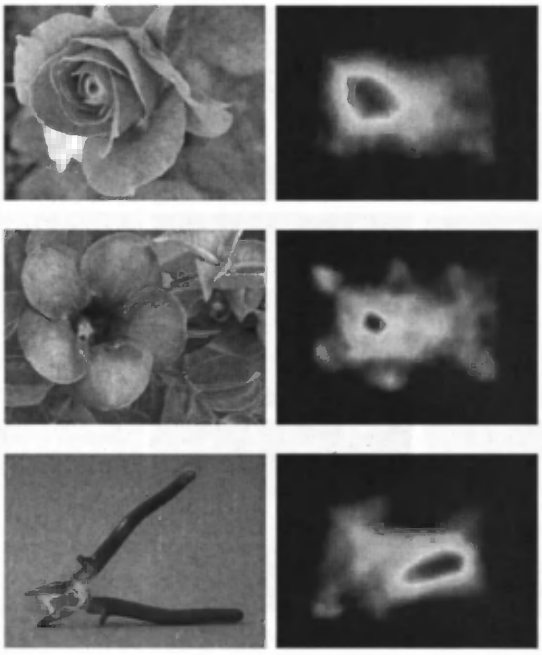
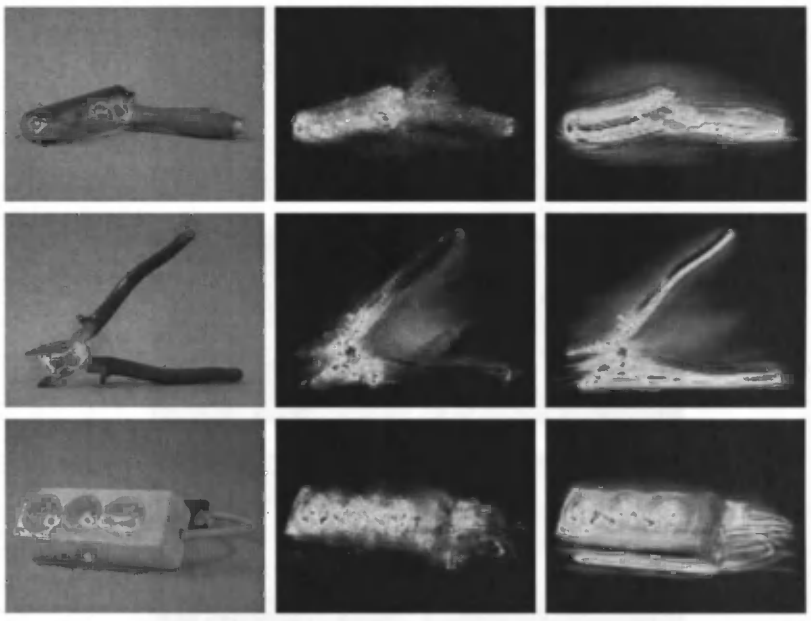


Figure 6.15: Figure 6.15 shows the human fixation maps and the feature maps generated by the color symmetry measure at the coarsest at the coarsest level. At that level the size of the image is down scaled by a factor 16. This illustrates the capability of the color symmetry measure to highlight large structure such as the center of the flowers and the symmetry axis between the pliers.

by the isotropic symmetry method. The center of the elongated shape of the grip is clearly defined. This is not case for the saliency map generated by the radial symmetry method. Only the tips of the grips are clearly defined but not the whole grip. This causes the fixation map to agree more with the saliency map generated by the radial symmetry than with the saliency map generated by the isotropic symmetry method. A comparable case is the grip of the roller and the roller itself. While the isotropic symmetry method responds to the whole roller and grip, the radial symmetry responds more to the extremities these parts. Again, the latter method looks more like the fixation map. The final example is the picture of the multiple socket. The isotropic symmetry method accentuates the sides of the socket and the shadow of the socket. The radial symmetry method does not respond to the areas. This again results in a greater resemblance of saliency map from the radial symmetry measure with the human data. This property of the radial



(a) (b) (c)

Figure 6.16: Shown are the human fixation maps (a), the saliency maps produced by the radial symmetry method (b), and the color symmetry method (c).

symmetry is maybe the cause of its overall better performance compared to the isotropic and color symmetry methods.

The performance of the saliency model on the mugs is reasonably good in both comparison measures. This is mainly due to the performance of the color channel from the saliency model. The saliency map form the color channel always highlights the inside of the ear, the ears itself and the top of the mug. For an example see figure 6.17. This results in a reasonably good performance on the mugs. Overall the color channel from the saliency model is the best performing feature map of the saliency model. With both the correlation measure and the fixation saliency measure the color channel is the best performer or at least ties with the intensity channel. The orientation channel is the overall worst performer.

6.4 Concluding

The results clearly show the possibility of predicting human fixations with bottom-up features. Practically all methods with both evaluation measures

(a) (b) (c)

Figure 6.17: Human fixation map (a) of an image from the mugs category followed by a saliency map generated by the saliency model (b) and the separate color feature map from the saliency model (c).

are able to predict human fixations to some extend. The two evaluation measures do show some differences in the rankings of the different methods. For example, we consider SIFT to be a good predictor of saliency although this is not shown with the correlation measure. On the other hand, Symmetry performs consistently good with both evaluation measures, but the performance of an algorithm does depend on the type of evaluation method used to assess the performance.

Overall, we consider radial symmetry to be the best performing method closely followed by isotropic and color symmetry. Radial symmetry seems to perform better than the other algorithms based on symmetry, because it does not detect the centers of elongated objects. Symmetry as fixation predictor was not thoroughly investigated before and it is striking that the methods isotropic, color, and radial symmetry perform better than the well known saliency model as a predictor of fixations. This makes algorithms based on symmetry and especially radial symmetry very interesting algorithms to predict fixations.

It is not easy to determine which features which are interesting to humans, but we have identified some points of interest of human observers, e.g. spots, highlights, and also line crossings. We also gave a illustration of the dominance of the bottom-up influence of color on the eye movements. Furthermore, symmetry is able to detect the center of flowers which is also a location which is frequently fixated by humans. Moreover, the comparison with the article by Privitera and Stark (2000) shows that the usage of multiple scales produces better performance of the methods.

Concluding, we see that it is possible to predict human fixation with bottom-up features and that the performance of several algorithms based on symmetry and radial symmetry in particular give the best performance.

Chapter 7

Discussion

In this thesis we investigated which bottom-up feature is the best suited feature for predicting human eye movements. We used several algorithms to process an image to determine the interesting regions in an image. One of the algorithms we used is the saliency model. This algorithm is a biologically plausible model of visual attention by Itti et al. (1998). Another algorithm was the SIFT algorithm (Lowe, 2004) which is well know algorithm used in computer vision research. We gave special attention to algorithms which detect symmetry (Reisfeld et al., 1995; Heidemann, 2004; Kovesi, 1997). The concept of symmetry has received quite some attention (Wagemans, 1997) with respect to how people perceive symmetry, but was not thoroughly tested as predictor of human fixations and was never compared to the performance of the saliency model. We also included ten algorithms from an article by Privitera and Stark (2000) in our comparison.

To determine what parts of the image were found interesting by humans we setup two experiments. The experiments offered the same images as processed by the algorithms to the participants. The eye movements in response to these images were recorded and compared to the predictions by the algorithms. To compare the predictions of the algorithms to human eye movement data we used two methods: the correlation measure and the fixation saliency measure, which are described in chapter 5. The results of this comparison can tell us something about the human visual system. If a certain algorithm consistently performs better than all the other algorithms, than maybe the algorithm corresponds to a similar process present in the human visual system. Another aim was to find a algorithm which could possibly be used as an algorithm which selects interesting information in an artificial system.

Similar to the results in the articles by Ouerhani et al. (2004) and Parkhurst et al. (2002), the saliency model showed a performance above

chance level in predicting fixations. Not only the saliency model performed above chance, but almost all algorithms we used, were able to predict fixations to some extend. The most successful algorithms are the symmetry algorithms from Reisfeld et al. (1995), isotropic and radial symmetry, and from Heidemann (2004), color symmetry. The saliency measure performed worse with both evaluation measures. The SIFT algorithm showed very mixed results. The algorithm performed very bad when evaluated with the correlation measure and very good with the fixation saliency measure. The methods: xlike, Laplacian of the Gaussian and Center surround from the article by Privitera and Stark (2000) performed pretty well. Better than the saliency model but worse than the isotropic, color and radial symmetry methods.

Although the results indicate that the mentioned symmetry algorithms perform better than all other algorithms, it is not possible to give an absolute indication of the performance of radial symmetry algorithm or any other algorithm. The problem is that we do not know to what extend it is possible to predict fixations on the basis of bottom-up information. It is known that several brain areas play a role in the control of eye movements (Pierrot-Deseillingny et al., 2004). Some areas are involved in reflexive control (bottom-up), while other areas are involved in intentional saccade control (top-down). Unfortunately, we do not know the precise workings of the bottom-up and top-down processes. The amount of influence of both processes also remains a much debated issue. Some research suggests that bottom-up processes can always override top-down influences (Theeuwes, 2004), while others are of the opinion that top-down process can override bottom-up influences on eye movements (Bacon & Egeth, 1994). This makes it very hard to determine a top-line or best possible performance for our measures, which are considered bottom-up measures and therefore, it is hard to judge the absolute performance of the algorithms.

We did include two possible top-line predictions with the correlation measure (see section 5.2). One was the inter-subject score which determines to what extend the participants predict each other fixations. Which should give an indication of the influence of bottom-up processing because the bottom-up processing in the visual system is assumed to be equal or nearly equal for all participants, but the best symmetry methods scored consistently better than this measure. The second was the intra-subject measure which indicates the extend to which subjects predict their own fixation pattern. This measure should give us a top-line, because both the bottom-up process and top-down process should be reasonably similar for each time the participant viewed the same image. But the best symmetry measures scored on par with this measure, which would mean symmetry measure would predict fixations

better than the participant would predict him or herself. This seems very unlikely, because the best predictor of a system is the system itself. But an algorithm can outperform the intra-subject measure if participants in the experiment did not repeatedly look at the same location, but did look at similar features such as symmetry. This means we did not succeeded to establish a suitable top-line because the intra-subject measure does not give the best prediction possible, but we can also conclude that symmetry is a good predictor of fixations.

We cannot get an absolute measure of the performance of the algorithms, but can we learn something about the human visual system? The most biological plausible algorithm we used was the saliency model. We expected the saliency model to be one of the best performers, because the model is based on a well-known theory of visual attention, the feature integration theory (Treisman & Gelade, 1980), which is based on findings of the physiological structure of the human visual system. However, the model performed consistently worse than the radial, isotropic and color symmetry methods for example. A possible explanation could be that the saliency model starts to process images at a coarser scale than the other algorithms. The saliency model first scales the images down before extracting the information. The other algorithms start processing the image at its original size. This could lead to a loss of information which would have been useful for the prediction of the fixations. Another reason of the relatively bad performance could be that the different modalities such as color and intensity are combined in a erroneous way. If we look at the results we see one channel, the color channel, always performs better than the overall performance of the model. Maybe if another way to combine the different channels was used, a more cumulative performance of the combined channels would be observed. Another possibility could be that features such as color, orientation and intensity which are known to pop-out in search tasks with simple shapes are not adequate to predict fixations with more complex images. Maybe that the modeling of only the very first stages of the visual system leads to this poor performance. Possibly the inclusion of later stages in the model could improve the performance. Alternatively, a channel based on one of the other algorithms could be included. A channel based on symmetry would be an interesting addition to the saliency model.

If we look at the symmetry algorithms: radial, isotropic and color symmetry, we can ask the question if humans really process symmetry or if symmetry detection is a byproduct of some other visual process. There is no established neural substrate which is responsible for detecting symmetry. If we do detect symmetry, can we assume the human visual system detects symmetry, on all scales? The scales we use in our algorithms are chosen

quite arbitrarily. The algorithms we use to predict human fixations all use multiple scales of the input image. The saliency model uses different combinations of scales of an image. The scales range from the second octave up to the seventh octave. The SIFT model uses three octaves, and the other algorithms, including the symmetry algorithms, use five octaves. An article by Dakin and Herbert (1998) suggests that symmetry is only detected in a region the size of a fixation or just a little bigger. Contrary to this finding, we already cover an area of this size in the first scale. None the less, when we compared the performance of the algorithms from Privitera and Stark with a single scale to the same algorithm with multiple scales, we saw an improved performance when we use multiple scales. Which could indicate that the use of multiple scales is in better agreement with the human visual system than using a single scale. Furthermore, as we have seen in chapter 6, the predictions of the symmetry on higher scales can give us good predictions of fixations. For example, the prediction of fixations on the centers of flowers which are in accordance with the fixations in the data from the participants in the experiment. But we could ask the question if the fixations are really the consequence the detections of symmetry of the flower by the participants. Maybe the detection of the center of the flower is the outcome of some top-down directed object segmentation which separates the flower from the background and finds the center of that object. Our research suggests that symmetry detection plays a role in visual attention, but more behavioral and neurophysiological research is needed to confirm this finding.

A weakness of the symmetry algorithm is the fact that the algorithm predicts fixations of points which are surrounded by symmetric gradients, but are themselves located in uniform regions. We found that uniform regions, such as the uniform background of the object images (See appendix A), are unlikely to be fixated. This makes the predictions of these points unlikely. Finding solution to this problem, which could improve the performance of a symmetry algorithm, would be an interesting topic of feature research.

We are also interested in finding new algorithms for predicting fixations. One of the observations made in chapter 6 is that participants in the experiment tend to focus on *structural centers* or line crossings. People focus on areas of the image were several prominent lines cross (See figure 6.13). This may be an interesting feature to use as a basis for a region of interest predictor. Although intersections do not pop-out (J. Wolfe & DiMase, 2003) and can therefore by definition not considered to be processed pre-attentively (Treisman & Gelade, 1980) it still may be an adequate fixation predictor, because symmetry detection is also not considered a pre-attentive process.

Concluding, a few methods based on detection of symmetry: isotropic, color, radial symmetry are the best predictors of human fixations. Symmetry

performs surprisingly well compared to, for example, the more biological plausible saliency model.

To learn more about symmetry processing in humans, we suggest two further lines of research. One, If we want to understand the effect of the image on the fixations we should construct specific images with small controllable modifications, instead of showing different types of images which differ in so many ways that it is hard to determine the influence of symmetry. Two, we will have to find out how the human visual system detects symmetry or which process responds to symmetrical patterns. These two objectives could for example be tested in one experiment with an experiment which uses both an eye tracker and fMRI to register the responses of the participants.

We could also consider if an algorithm, which performed well as predictor of fixations, could also be used in an artificial system. We tested the performance of the SIFT algorithm, which is a well known algorithm from the field of computer vision. As mentioned earlier the performance of the SIFT method produces very mixed results for the two evaluation measures, but we can conclude SIFT does predict locations which are likely to be fixated. It would be interesting to see if we tried the reverse and used symmetry as an algorithm to extract descriptors of an image for object recognition which is what SIFT is intended to do. (See section 3.2). Symmetry is expected to be robust to changes in rotation, translation, and scaling, which is an important property for computer vision models.

If we want to use one of the algorithms in an autonomous robot setting, than the robot could use such an algorithm as a filter for visual information. A symmetry algorithm, radial symmetry for example, could pre-process visual information to extract interesting locations. This could lead to a reduction in visual data which has to be processed in later steps. This would require the algorithms to process images in real-time. The symmetry algorithms are unfortunately to slow on commodity hardware to achieve practical resolutions and frame rates, but the algorithm can be easily adapted to be executed in parallel which could enable it to run real-time on specialized hardware.

This gives us several possibilities to research the influence of symmetry and other algorithms on fixations, obtain a better understanding of the human visual system, and possibly find an application in the field of autonomous systems.

Appendix A

Pictures

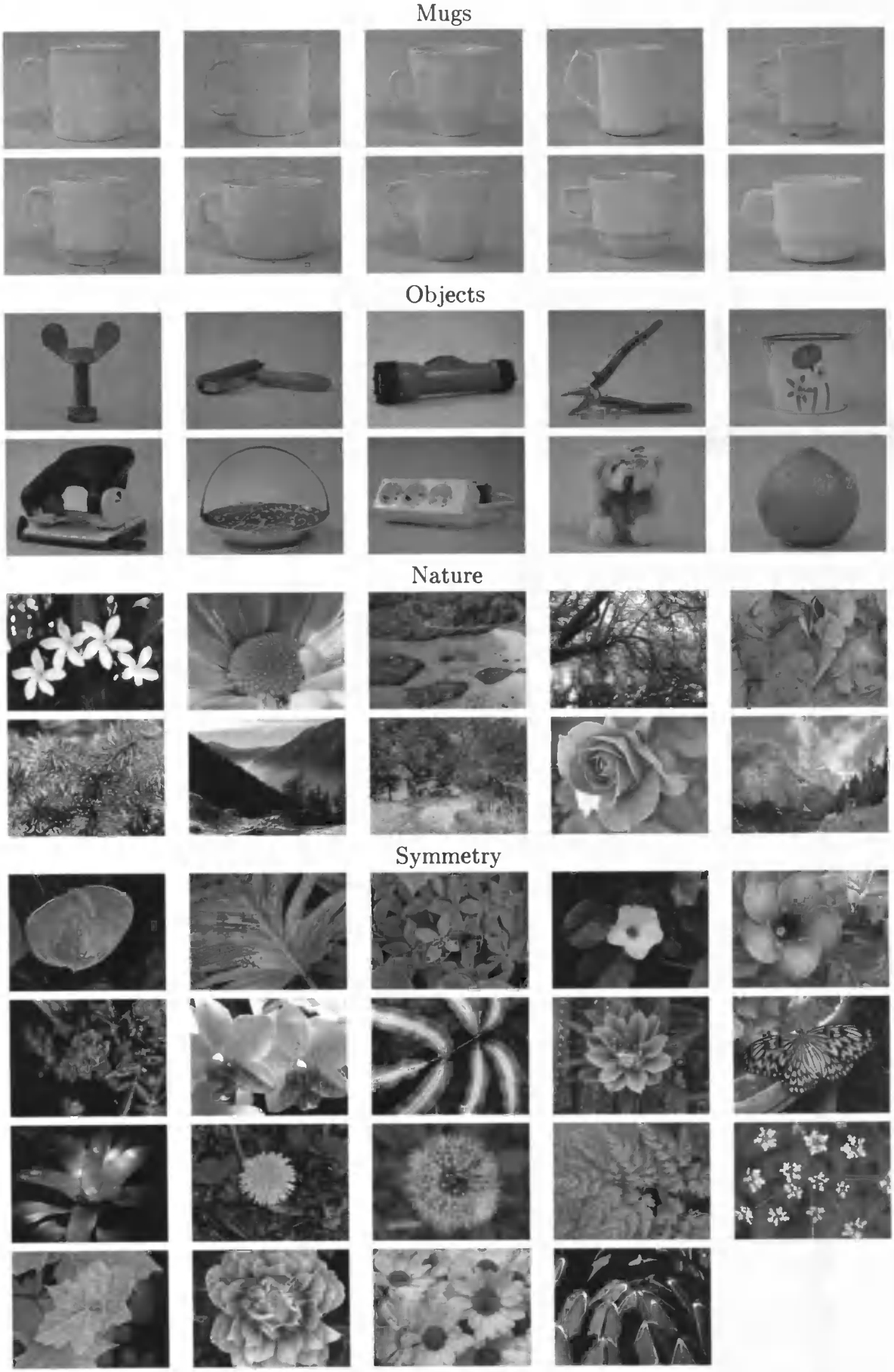


Figure A.1: All photos used in the experiments. The photos are divide into four categories: mugs, nature, objects and symmetry.

Appendix B

Correlation between methods

	s_isotropic	s_color	s_radial	xlike	log	csurround	dct	wavelet	saliency	saliency_I	saliency_C	saliency_O	sift	orientation	edges	entropy	mcontrast	s_phase	s_simple
s_isotropic	–																		
s_color	0.87	–																	
s_radial	0.75	0.81	–																
xlike	0.61	0.47	0.37	–															
LoG	0.68	0.48	0.33	0.78	–														
csurround	0.65	0.46	0.30	0.78	0.99	–													
DCT	0.51	0.37	0.21	0.70	0.82	0.84	–												
wavelet	0.38	0.33	0.33	0.59	0.41	0.40	0.40	–											
saliency	0.77	0.67	0.51	0.60	0.70	0.69	0.59	0.38	–										
saliency_I	0.67	0.45	0.29	0.53	0.70	0.70	0.56	0.29	0.80	–									
saliency_C	0.53	0.53	0.42	0.45	0.48	0.47	0.42	0.26	0.77	0.33	–								
saliency_O	0.52	0.57	0.47	0.34	0.33	0.31	0.27	0.33	0.64	0.49	0.22	–							
sift	0.19	0.16	0.22	0.15	0.22	0.21	0.17	0.13	0.20	0.19	0.18	0.04	–						
orientation	0.43	0.27	0.11	0.46	0.55	0.56	0.48	0.25	0.47	0.56	0.22	0.28	0.06	–					
edges	0.35	0.28	0.27	0.37	0.31	0.31	0.30	0.40	0.32	0.23	0.28	0.20	0.11	0.16	–				
entropy	0.58	0.46	0.34	0.66	0.70	0.70	0.69	0.48	0.58	0.56	0.38	0.34	0.20	0.43	0.47	–			
mcontrast	0.39	0.21	0.10	0.27	0.33	0.33	0.25	0.19	0.43	0.66	0.06	0.29	0.07	0.35	0.17	0.37	–		
s_phase	0.33	0.10	0.10	0.43	0.53	0.53	0.41	0.26	0.34	0.42	0.20	0.09	0.13	0.35	0.28	0.37	0.21	–	
s_simple	0.46	0.49	0.37	0.34	0.37	0.36	0.32	0.14	0.41	0.26	0.32	0.39	0.06	0.14	0.24	0.50	0.14	0.06	–

Table B.1: Results of the correlations between all methods. A pairwise comparison between the methods was performed by calculating the correlation between the saliency maps produced by the two methods for a certain image. This was done for all images in the mugs category and the results were averaged.

	s_isotropic	s_color	s_rradial	xlike	log	csurround	dct	wavelet	saliency	saliency_I	saliency_C	saliency_O	sift	orientation	edges	entropy	mcontrast	s_phase	s_simple
s_color	0.95	–																	
s_radial	0.95	0.92	–																
xlike	0.53	0.49	0.50	–															
LoG	0.56	0.50	0.50	0.63	–														
csurround	0.53	0.47	0.47	0.62	0.98	–													
DCT	0.37	0.34	0.35	0.53	0.63	0.64	–												
wavelet	0.27	0.25	0.26	0.51	0.38	0.38	0.48	–											
saliency	0.21	0.17	0.17	0.17	0.36	0.36	0.13	0.10	–										
saliency_I	0.15	0.10	0.10	0.11	0.36	0.37	0.10	0.08	0.80	–									
saliency_C	0.15	0.14	0.15	0.09	0.17	0.17	0.09	0.06	0.70	0.28	–								
saliency_O	0.15	0.15	0.12	0.18	0.23	0.23	0.10	0.08	0.60	0.39	0.08	–							
sift	0.10	0.08	0.09	0.17	0.28	0.29	0.23	0.12	0.17	0.16	0.08	0.12	–						
orientation	0.12	0.07	0.11	0.14	0.28	0.28	0.14	0.11	0.29	0.34	0.13	0.11	0.13	–					
edges	0.43	0.40	0.41	0.43	0.38	0.38	0.37	0.36	0.02	0.00	0.02	0.04	0.05	0.08	–				
entropy	0.51	0.46	0.50	0.54	0.52	0.53	0.54	0.41	0.03	-0.00	0.03	0.02	0.12	0.09	0.53	–			
mcontrast	-0.13	-0.17	-0.13	-0.15	-0.08	-0.08	-0.16	-0.11	0.36	0.50	0.05	0.22	-0.01	0.31	-0.16	-0.22	–		
s_phase	0.24	0.15	0.18	0.31	0.54	0.56	0.32	0.26	0.28	0.33	0.14	0.10	0.14	0.25	0.21	0.24	0.03	–	
s_simple	0.47	0.47	0.41	0.30	0.31	0.30	0.12	-0.02	0.15	0.05	0.02	0.32	0.03	-0.00	0.25	0.33	-0.05	-0.01	–

Table B.2: Results of the correlations between all methods. A pairwise comparison between the methods was performed by calculating the correlation between the saliency maps produced by the two methods for a certain image. This was done for all images in the nature category and the results were averaged.

	s_Isotropic	s_color	s_radial	xlike	log	csurround	dct	wavelet	saliency	saliency_I	saliency_C	saliency_O	sift	orientation	edges	entropy	mcontrast	s_phase	s_simple
s_isotropic	–																		
s_color	0.94	–																	
s_radial	0.87	0.87	–																
xlike	0.62	0.59	0.60	–															
LoG	0.69	0.64	0.59	0.78	–														
csurround	0.69	0.63	0.57	0.77	0.99	–													
DCT	0.63	0.60	0.56	0.72	0.83	0.83	–												
wavelet	0.53	0.52	0.56	0.72	0.58	0.57	0.59	–											
saliency	0.75	0.72	0.66	0.55	0.67	0.67	0.55	0.47	–										
saliency_I	0.70	0.63	0.59	0.51	0.70	0.70	0.54	0.42	0.91	–									
saliency_C	0.70	0.68	0.59	0.51	0.59	0.58	0.49	0.42	0.92	0.72	–								
saliency_O	0.50	0.51	0.56	0.39	0.41	0.40	0.37	0.46	0.67	0.61	0.44	–							
sift	0.25	0.23	0.28	0.29	0.37	0.37	0.35	0.29	0.24	0.27	0.18	0.21	–						
orientation	0.47	0.40	0.38	0.38	0.56	0.57	0.42	0.27	0.57	0.67	0.45	0.31	0.28	–					
edges	0.64	0.62	0.59	0.62	0.60	0.60	0.60	0.54	0.54	0.45	0.53	0.38	0.20	0.30	–				
entropy	0.75	0.72	0.64	0.68	0.72	0.72	0.73	0.56	0.61	0.54	0.61	0.33	0.24	0.39	0.72	–			
mcontrast	0.56	0.49	0.51	0.31	0.41	0.42	0.34	0.23	0.71	0.77	0.56	0.47	0.18	0.60	0.30	0.41	–		
s_phase	0.59	0.48	0.47	0.50	0.62	0.62	0.47	0.40	0.55	0.56	0.49	0.33	0.18	0.41	0.48	0.51	0.34	–	
s_simple	0.30	0.34	0.18	0.24	0.28	0.28	0.27	0.13	0.26	0.20	0.25	0.22	0.04	0.13	0.26	0.42	0.14	0.08	–

Table B.3: Results of the correlations between all methods. A pairwise comparison between the methods was performed by calculating the correlation between the saliency maps produced by the two methods for a certain image. This was done for all images in the objects category and the results were averaged.

	s_isotropic	s_hcolor	s_rradial	xlike	log	csurround	dct	wavelet	saliency	saliency_I	saliency_C	saliency_O	sift	orientation	edges	entropy	mcontrast	s_phase	s_simple
s_isotropic	-																		
s_color	0.92	-																	
s_radial	0.93	0.89	-																
xlike	0.49	0.43	0.47	-															
LoG	0.56	0.47	0.50	0.67	-														
csurround	0.54	0.44	0.47	0.65	0.99	-													
DCT	0.35	0.29	0.32	0.55	0.67	0.68	-												
wavelet	0.24	0.19	0.23	0.55	0.42	0.41	0.47	-											
saliency	0.31	0.25	0.27	0.23	0.38	0.37	0.16	0.19	-										
saliency_I	0.31	0.21	0.25	0.22	0.44	0.44	0.20	0.20	0.81	-									
saliency_C	0.12	0.10	0.11	0.05	0.11	0.11	0.02	0.02	0.69	0.28	-								
saliency_O	0.29	0.28	0.26	0.28	0.31	0.30	0.17	0.23	0.67	0.48	0.18	-							
sift	0.14	0.11	0.15	0.20	0.30	0.30	0.26	0.14	0.15	0.18	0.04	0.12	-						
orientation	0.04	-0.01	0.03	0.10	0.23	0.23	0.13	0.11	0.26	0.31	0.09	0.19	0.13	-					
edges	0.39	0.35	0.36	0.38	0.35	0.35	0.34	0.31	0.06	0.06	-0.00	0.09	0.08	0.01	-				
entropy	0.43	0.37	0.42	0.50	0.50	0.51	0.55	0.37	0.03	0.05	-0.02	0.04	0.14	-0.08	0.49	-			
mcontrast	-0.14	-0.18	-0.13	-0.14	-0.08	-0.08	-0.14	-0.08	0.35	0.43	0.14	0.20	0.00	0.36	-0.18	-0.30	-		
s_phase	0.31	0.16	0.22	0.33	0.55	0.56	0.31	0.25	0.26	0.35	0.09	0.12	0.15	0.18	0.22	0.23	-0.01	-	
s_simple	0.46	0.48	0.40	0.31	0.31	0.30	0.16	0.04	0.15	0.06	0.03	0.31	0.03	-0.03	0.26	0.33	-0.11	-0.02	-

Table B.4: Results of the correlations between all methods. A pairwise comparison between the methods was performed by calculating the correlation between the saliency maps produced by the two methods for a certain image. This was done for all images in the symmetry category and the results were averaged.

	s_isotropic	s_color	s_radial	xlike	log	csurround	dct	wavelet	saliency	saliency_I	saliency_C	saliency_O	sift	orientation	edges	entropy	mcontrast	s_phase	s_simple
s_isotropic	-																		
s_color	0.92	-																	
s_radial	0.89	0.87	-																
xlike	0.55	0.48	0.48	-															
LoG	0.61	0.51	0.48	0.71	-														
csurround	0.59	0.49	0.46	0.70	0.99	-													
DCT	0.44	0.38	0.35	0.61	0.73	0.74	-												
wavelet	0.33	0.30	0.33	0.58	0.44	0.44	0.48	-											
saliency	0.47	0.42	0.38	0.36	0.50	0.50	0.32	0.27	-										
saliency_I	0.43	0.32	0.30	0.32	0.53	0.53	0.32	0.24	0.82	-									
saliency_C	0.33	0.31	0.28	0.23	0.30	0.29	0.21	0.16	0.76	0.38	-								
saliency_O	0.35	0.36	0.33	0.30	0.32	0.31	0.22	0.27	0.65	0.49	0.22	-							
sift	0.17	0.14	0.18	0.20	0.29	0.29	0.25	0.16	0.18	0.19	0.10	0.12	-						
orientation	0.22	0.15	0.14	0.24	0.37	0.38	0.26	0.17	0.38	0.44	0.20	0.22	0.15	-					
edges	0.44	0.40	0.40	0.44	0.40	0.40	0.39	0.38	0.20	0.16	0.17	0.16	0.10	0.11	-				
entropy	0.54	0.48	0.47	0.58	0.59	0.59	0.61	0.44	0.26	0.24	0.20	0.16	0.17	0.15	0.54	-			
mcontrast	0.11	0.04	0.05	0.04	0.10	0.10	0.03	0.03	0.44	0.56	0.19	0.28	0.05	0.40	-0.01	-0.00	-		
s_phase	0.36	0.21	0.24	0.38	0.56	0.56	0.37	0.28	0.34	0.41	0.20	0.15	0.15	0.28	0.28	0.32	0.11	-	
s_simple	0.43	0.45	0.35	0.30	0.32	0.31	0.21	0.07	0.23	0.13	0.13	0.31	0.03	0.04	0.25	0.38	0.01	0.02	-

Table B.5: Results of the correlations between all methods. A pairwise comparison between the methods was performed by calculating the correlation between the saliency maps produced by the two methods for a certain image. This was done for all images in the symmetry category and the results were averaged.

Bibliography

Attention: Contemporary theory and analysis. (1970). In D. I. Mostofsky (Ed.), (p. 99-124). New York: Appleton-Century.

Bacon, W., Egeth, H. (1994, May). Overriding stimulus-driven attentional capture. *Percept Psychophysics*, *55*(5), 485-496.

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

Blind, I. American Printing House for the. (2004). *Visual pathway.* Website. (`http://www.aph.org/cvi/brain.html`)

Canny, J. (1986). A computational approach to edge detection. *Pattern Analysis and Machine Intelligence*, *8*, 679-714.

Dakin, S., Herbert, A. (1998). The integration region for symmetry detection. *The spatial region of integration*(265), 659-664.

Daubechies, I. (1992). *Ten lectures on wavelets (c b m s - n s f regional conference series in applied mathematics).* Soc for Industrial & Applied Math.

Engel, S., Zhang, X., Wandell, B. (1997). Colour tuning in human visual cortex measured with functional magnetic resonance imaging. *Nature*, *388*(3), 68-70.

Eriksen, C., St. James, J. (1986). Visual attention within and around the field of focal attention a zoom lens model. *Perception & Psychophysics*, *40*(4), 225-240.

Eriksen, C., Yeh, Y. (1985). Allocation of attention in the visual field. *Journal of experimental psychology: human perception & performance*, *11*(5), 583-597.

Essen, D. C. V., Maunsell, J. H. R. (1983). Hierarchical organization and functional streams in the visual cortex. *Trends in Neuroscience*, *6*(370-375).

Heidemann, G. (2004, July). Focus-of-attention from local color symmetry. *IEEE Transaction of Pattern Analyis and Machine Intelligence*, *26*(7), 817-830.

Hochstien, S., Ahissar, M. (2002, December). View from the top: hierarchies and reverse hierarchies in the visual system. *Neuron*, *36*, 791-804.

Hubel, T. N., David H.and Wiesel. (1959). Receptive fields of single neurons in the cat's striate cortex. *Journal of Physiology (London)*, *148*, 574-591.

Hubel, T. N., David H.and Wiesel. (1974). Sequence regularity and geometry of orientation columns in the monkey striate cortex. *The Journal of Comparative Neurology*, *158*(3), 267-293.

Irwin, D. E. (1992, March). Memory for position and identity across eye movements. *Journal of Experimental Psychology*, *18*(2), 307-317.

Itti, L., Koch, C., Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions and Pattern Ananlysis and Machine Intelligence*, *20*(11), 1254-1259.

Kovesi, P. (1997). Symmetry and asymmetry from local phase. In *Ai'97, tenth australian joint conference on artificial intelligence* (p. 185-190).

Krieger, G., Rentschler, I., Hauske, G., Schill, K., Zetzsche, C. (2000). Object and scene analysis by saccadic eye-movements: an investigation with higher-order statistics. *Spatial Vision*.

Kuffler, S. W. (1953). Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology*(16), 37-68.

Lehrer, M. (1999). Shape perception in the honeybee: symmetry as a global framework. *International Journal of Plants Sciences*, *160*(S6), S51-S65.

Le Meur, O., Le Callet, P., Thoreau, D. (2006). A coherent computational approach to model bottom-up visual attention. *IEEE transactions on pattern analysis and machine intelligence*, *28*(5), 802-815.

Livingstone, S. M., Hubel, D. H. (1984). Anatomy and physiology of a color system in the primate visual cortex. *The Journal of Neuroscience*, *4*(1), 309-356.

Lochner, P., Nodine, C. (1989). The perceptual value of symmetry. *Computers and Mathemtics with Applications*, *17*(4-6), 475-484.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*.

Møller, A. (1993). Female pereference for apparently symmetrical male sexual ornaments in the barn swallow *Hirundo rustica*. *Behavioural Ecology and Sociobiology*, *1993*(32), 371-376.

Noton, D., Stark, L. (1971, June). Eye movements and visual perception. *Scientific American*(224), 35-43.

Ouerhani, N., Wartburg, R. v., Hügli, H., Müri, R. (2004). Emperical validation of the saliency-based model of visual attention. *Electronic Letters on Computer Vision and Image Analysis*, *3*(1), 13-24.

Parkhurst, D., Law, K., Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Visual Research*, *42*(107-123).

Pashler, H. (1988). Cross-dimensional interaction and texture segregation. *Percept Psychophysics*, *43*(4), 307-318.

Photons to phenomology. (1999). The Mit Press.

Pierrot-Deseillingny, C., Milea, D., Müri, R. (2004). Eye movement control by the cerebral cortex. *Current opinion in neurology*, *17*, 17-25.

Posner, M. (1978). *Chronometric explorations of mind*. Hillsdale, N.J: Lawrence Erlbaum Associates.

Privitera, C., Stark, L. (2000). Algorithms for defining visual regions-of-interest: comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(9).

Reinagle, P., Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Computational Neural Systems*(10), 1-10.

Reisfeld, D., Wolfson, H., Yeshurun, Y. (1995). Context-free attentional operators: the generalized symmetry transform. *International Journal of Computer Vision*(14), 199-130.

Shannon, C. (1948, July and October). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379-423 and 623-656.

Shulman, G., Wilson, J. (1987). Spatial frequency and selective attention to local and global information. *Perception*, *16*(1), 89-101.

Silverman, M. S., Grosof, D. H., De Valois, R. L., Elfar, S. D. (1988). Spatial-frequency organization in primate striate cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *86*(2), 711-715.

Stark, L. W., Privitera, C. M., Yang, H., Azzariti, M., Fai Ho, Y., Blackmon, T., et al. (1999). *Representation of human vision in the brain: How does human perception recognition recognize images?* (Tech. Rep. No. UCB/ERL M99/49). EECS Department, University of California, Berkeley.

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

Treisman, A. M., Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97-136.

Tsal, Y. (1983). Movement of attention accross the visual field. *Journal of experimental psychology: human perception & performance*, *9*(4), 523-530.

Wagemans, J. (1997, December). Characteristics and models of human symmetry detection. *Trends is cognitive sciences*, *1*(9), 346-348.

Wenderoth, P. (1994). The salience of vertical symmetry. *Perception*(23).

Wenderoth, P. (1995). The role of pattern outline in bilateral symmetry detection with briefly flashed dot patterns. *Spatial Vision*, *9*(1), 57-77.

Wolfe, J., DiMase, J. (2003). Do intersections serve as basic features in visual search? *Perception*, *32*, 645-656.

Wolfe, J. M., Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *International Journal Of Pharmacy Practice*, *12*(2), 73-81.

Yarbus, H., A.L. (1968, September). Eye movements and vision. *The quaterly review of biology*, *43*(3), 360.