Off-line Foveated Compression and Scene Perception: An Eye-Tracking Approach

Nyström, Marcus

2008

Off-Line Foveated Compression and Scene Perception: An Eye-Tracking Approach

Marcus Nyström

LUND UNIVERSITY

Ph.D. Thesis, September 26, 2008

Marcus Nyström
Department of Electrical and Information Technology
Lund University
Box 118
S-221 00 Lund, Sweden
e-mail: marcus@eit.lth.se
http://www.eit.lth.se/

ISBN: 1654-790X
ISRN: 9

Abstract

With the continued growth of digital services offering storage and communication of pictorial information, the need to efficiently represent this information has become increasingly important, both from an information theoretic and a perceptual point of view.

There has been a recent interest to design systems for efficient representation and compression of image and video data that take the features of the human visual system into account. One part of this thesis investigates whether knowledge about viewers' gaze positions as measured by an eye-tracker can be used to improve compression efficiency of digital video; regions not directly looked at by a number of previewers are lowpass filtered. This type of video manipulation is called *off-line foveation*. The amount of compression due to off-line foveation is assessed along with how it affects new viewers' gazing behavior as well as subjective quality. We found additional bitrate savings up to 50% (average 20%) due to off-line foveation prior to compression, without decreasing the subjective quality.

In off-line foveation, it would be of great benefit to algorithmically predict where viewers look without having to perform eye-tracking measurements. In the first part of this thesis, new experimental paradigms combined with eye-tracking are used to understand the mechanisms behind gaze control during scene perception, thus investigating the prerequisites for such algorithms. Eye-movements are recorded from observers viewing contrast manipulated images depicting natural scenes under a neutral task. We report that image semantics, rather than the physical image content itself, largely dictates where people choose to look. Together with recent work on gaze prediction in video, the results in this thesis give only moderate support for successful applicability of algorithmic gaze prediction for off-line foveated video compression.

Contents

Preface

Most of the work in this thesis has been published in journals and presented at conferences and workshops. The main contributions are:

- Nyström, M. & Holmqvist, K. (2008, submitted). Effect of compressed off-line foveated video on viewing behavior and subjective quality. *ACM Trans. on Multimedia Computing, Communications, and Applications.* (14 pages)

- Nyström, M. & Holmqvist, K. (2008, in press). Semantic override of low-level features in image viewing – both initially and overall. *Journal of Eye Movement Research.* (11 pages)

- Nyström, M. & Holmqvist, K. (2007). Variable resolution images and their effects on eye-movements. In B. Rogowitz, T. Pappas & S. Daly (Eds.), *Human Vision and Electronic Imaging.* San Jose, CA: SPIE. (9 pages)

- Nyström, M. & Holmqvist, K. (2007). Variable resolution images and their effects on eye-movements. *European Conference on Eye Movements*, Potsdam, Germany. (Abstract)

- Nyström, M. & Holmqvist, K. (2007). Deriving and evaluating eye-tracking controlled volumes of interest for variable-resolution video compression. *J. Electron. Imaging*, 16(013006). (8 pages)

- Nyström, M. & Holmqvist, K. (2007). Perceptual effects of off-line foveated video. *Scandinavian Workshop on Applied Eye-Tracking*, Lund, Sweden. (Abstract)

- Nyström, M. & Holmqvist, K. (2006). Perceptual effects of off-line foveated video. *Scandinavian Workshop on Applied Eye-Tracking*, Lund, Sweden. (Abstract)

- Nyström, M. & Holmqvist, K. (2005). A quality function for variable resolution video based on eye-tracking measurements. Poster, *Bioinspired Information Processing*, Lubeck, Germany. (Abstract)
- Nyström, M. & Holmqvist, K. (2005). Allocating overt attention in video based on eye-tracking measurements. *European Conference on Eye Movements (ECEM)*, Bern, Switzerland, August 2005. (Poster)
- Nyström, M. Novak, M. & Holmqvist, K. (2004). A novel approach to image coding using off-line foveation controlled by multiple eye-tracking measurements. *Picture Coding Symposium*, San Francisco, CA. (6 pages)

The information from the listed contributions is not included in the thesis in its original form, but has, when needed, been altered to enhance clarity and to promote the transformation of a collection of papers into a more comprehensive and naturally coherent thesis.

The work originates from my five years as a Ph.D. student and has been performed jointly at the department of Electrical and Information Technology (EIT) and the Lund Humanities Lab (LHL), Lund University.

Acknowledgments

Without the help and support from my academic collaborators, colleagues and friends, this thesis could not have been written.

My advisor Kenneth Holmqvist has been the biggest support during my years as a PhD student. Each meeting with Kenneth is followed by enthusiasm and creativity, which have inspired much of the work in this thesis. He has kindly shared his expertise in eye-tracking theory and methods as well as in scene perception.

Mirek Novak introduced me to image and video compression, and emphasized the importance of pedagogical skills in teaching. Mirek always had time to sit down and talk about small and big issues in life. His early demise left a big void, both academically and personally.

I am grateful to John B. Anderson for kindly accepting me as his student, and for helping me to finish this thesis. I am particularly glad that he gave me the opportunity to visit the ViVoNets lab, where Prof. Jerry D. Gibson guided me toward new and exciting ideas in image and video compression.

More than once I had to explain and defend my engineering approaches to humanities research for the ET-seminar group. Neither questionable experimental methods nor bad research ideas slipped through the scrutinizing eyes of this group. Without doubt, this was cross-disciplinary research at its best!

Many are the people who have been participating in my experiments, and from whom I have collected eye-tracking data. I would not replace you with any gaze prediction algorithm in the world!

Technical and administrative support at the Department of Electrical and Information Technology (EIT) has been top notch. Not only do you offer professional support, but you are also very friendly and nice!

Fortunately, life at EIT has not been all about research and course work. Past and present PhD students have become good friends, and have given the department an open and cheerful working environment. The lunches and coffee breaks we have are always welcome daily elements.

Finally, I would like to thank my family for their love and support.

Chapter 1

Introduction

DIGITAL information is increasingly more pictorial in nature, and we are constantly fed with visual impressions through television, the Internet, and our cellular phones, to name a few important examples. Consequently, it is an ongoing challenge to find improved methods for efficient representation and storage of this voluminous data. Today, there are a number of quite mature methods for data compression popularized in standards such as the JPEG and MPEG families of codecs. However, a common denominator for these standards is that they take only few of the properties of the human visual system (HVS) and perception into account. It is therefore likely that many of tomorrow's improvements in these standards lie in optimizing images and videos not only in a mathematical framework, but over the end-to-end optimization between image acquisition and the viewer(s) at the receiving end.

Surely, the time is now right to further cross-fertilize knowledge from information theory and cognitive psychology to facilitate improved data compression. In this thesis, we will investigate whether current state-of-the-art methods for compression of pictorial data can be improved by taking into account where people look as measured by an eye-tracker. Since regions outside the central line of sight cannot be seen with high detail, the quality of such regions can be reduced without this being noticed. Clearly, this opens a large potential for improved data compression.

Of course, it would be of great advantage if it was possible to algorithmically predict where people would look, without having to perform time-consuming eye-tracking experiments. In view of this, a part of the thesis is devoted to empirical investigations of the cognitive mechanisms behind gaze control in image viewing. For example, we address questions like: 'Where do people look when presented to natural scenes?' and 'Why do they look toward these regions?'. The answers to these questions are crucial first steps toward future successful algorithms for gaze prediction in video.

The thesis is divided in two parts. Using a new experimental paradigm, we investigate in Part I the mechanisms behind gaze control. In Part II, we measure how off-line foveation affects compression, subjective quality, and eye-movements. Below, an overview of the contents and main results of the thesis are given.

1.1 Overview of Part I: Gaze Behavior in Images

Gaze behavior in scene viewing has been investigated for over a century, with important pioneering work in the early twentieth century by Buswell (1935) and later by Yarbus (1967). Some important aspects in scene perception concern the cognitive mechanisms behind eye-movement control: to which scene regions do viewers look, and why do they look toward these particular regions of the scene? Although much is known from the vast amount of published research on the subject, there is currently an intense debate of how higher- and lower cognitive factors interact to control where people direct their gazes. While this type of research is well motivated solely to increase the general understanding about the HVS and visual cognition, accurate models of gaze control and prediction would have direct practical applications within fields such as computer vision, image and video compression, marketing, and automobile safety. There have been numerous efforts to develop computational tools to predict human fixation locations, many relying on the basic structure outlined by Koch and Ullman (Koch & Ullman, 1985). Although many of these models seem promising, they are currently quite far from mimicking the behavior of a human viewer in terms of accurate modeling of fixation locations and, in particular, fixation durations.

In this part of the thesis, partly to highlight the limitations of models predicting visual attention, we aim to better understand the causes behind gaze shifts during inspection of natural images. We use a new experimental paradigm where low-level image statistics are manipulated to dissociate objects from their low-level signal strength. Eye-tracking experiments are then performed to elicit the spatial and temporal contributions of lower and higher cognitive factors to gaze guidance. Figure 1.1 gives an example of the stimuli we used in the experiments. The image is contrast manipulated such that the face is blurred, and the circles represent fixations collected from a number of observers free-viewing the image. The diameter of each circle is proportional to the fixation duration. Notice that the face attracts many (long) fixations despite its low contrast and thus lack of detailed facial features. In this case a typical,

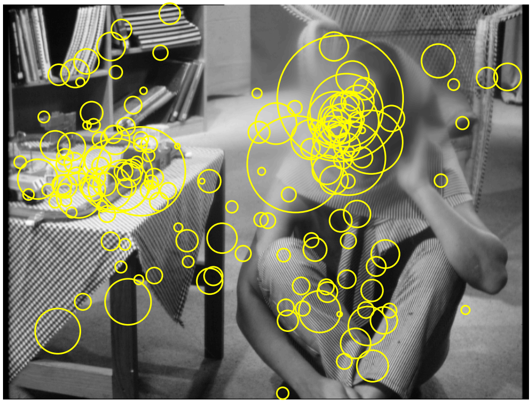


Figure 1.1: Distribution of fixations over a contrast manipulated image. Each circle represents a fixation, and the diameter of each circle is proportional to the fixation duration.

image-driven algorithmic predictor would fail miserably to predict human fixations.

The highlights of our findings reveal that the interplay and relative contribution between lower and higher cognitive factors on gaze guidance are linked with the semantics of the viewed image; fixated content in images with neutral semantics correlates quite well with image features whereas semantically important objects are gazed upon despite a weak feature signal strength.

Part I of the thesis is outlined as follows. Chapter 2 gives an introduction of some properties of the HVS and also a brief overview of eye-movements and visual attention as well as how they are coupled. That is, does the position of the eye also indicate where attention is located? If so, how tight is this relationship? In Chapter 3, we review some key papers on gaze behavior in image viewing. Specifically, we address what is previously known about where people look, why they choose to look at these regions, and for how long. The following three chapters (4, 5, 6) present our work which is mainly based on material from the papers Nyström and Holmqvist (2007b) and Nyström and Holmqvist (2008, in press). Our main findings are summarized and discussed in Chapter 7.

1.2 Overview of Part II: Off-line Foveation for Video Compression

In most practical application, compression is essential to manage and store image and video data. Compression efficiency is a trade-off between bitrate, quality, and computational complexity, and today's standards for compression have addressed these issues quite successfully. A typical image coder such as JPEG can compress an image to about 1/30 of its original size and still produce acceptable quality. Video coders can further improve this ratio due to significant temporal redundancies present in video data. Despite these substantial capabilities for data compression, there is a constant demand for improved compression efficiency due to factors such as ever larger picture formats, increasing costs for bandwidth, etc.

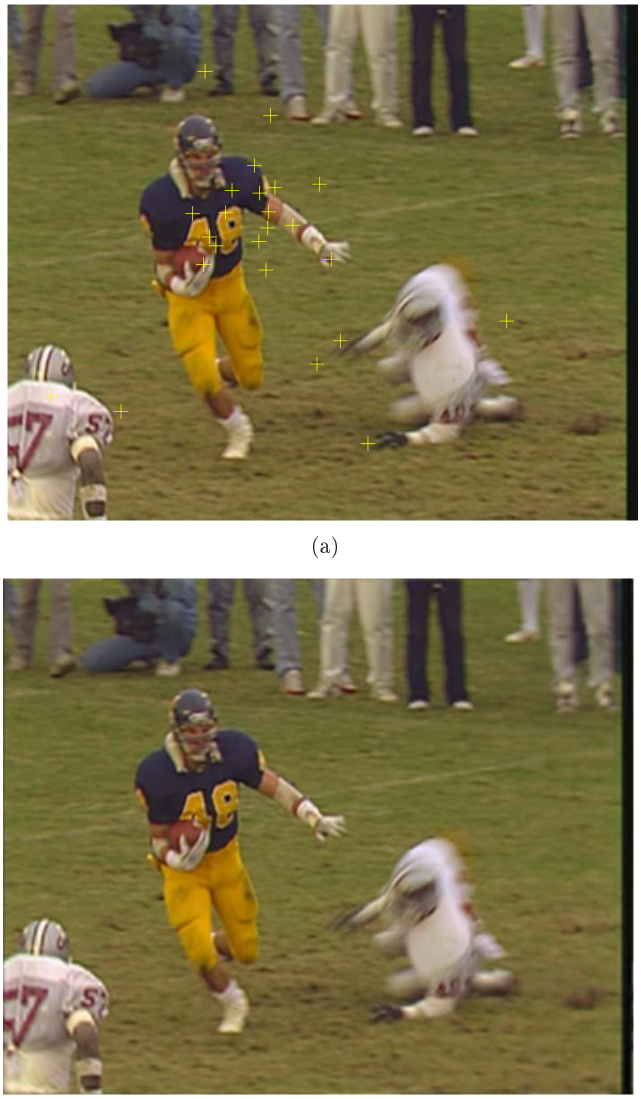
In this part of the thesis, we investigate how knowledge about where people look can be utilized to improve compression efficiency of digital video. If we knew where people looked while viewing video, unattended parts could be degraded in quality and, due to the inability of the HVS to resolve fine detail in peripheral vision, this would not be noticed. Since regions low in spatial detail generally require fewer bits to represent digitally, this opens a large potential for improved data compression. The following questions are addressed: Where do observers look? Do observers look toward similar regions? If we know where people look, how much can we degrade regions where people do not look (and thus decrease the bitrate) without decreasing the subjective quality and changing where people initially look?

In our work, eye-tracking is utilized to collect eye-movements from a number of observers while free-viewing images and videos. This eye-movement data is then used to study observers' viewing behavior as well as to control the bit-allocation such that visually attended regions are given more bits than regions not visited by peoples' high-acuity, foveal vision. We have dubbed this approach *off-line foveation*. An illustrative example is shown in Figure 1.2. Figure 1.2(a) depicts a frame from a video shown to a group of viewers. Each crosshair represents one viewer's gaze position. Figure 1.2(b) shows this frame after off-line foveation. Notice the peripheral blurring in unattended regions.

We will address the design, implementation and evaluation of off-line foveated image and video coding. Specifically, we focus on a number of central challenges. First, a method is proposed to transform collected gaze positions to regions of interest (ROIs) for images, and volumes of interest (VOIs) for video applications. Second, we address the problem of how the ROI/VOI could be used to implement off-line foveation. Third, we target off-line foveation in a framework of video coding. Fourth, we devise new methods to evaluate off-line foveation subjectively.

Our results show that off-line foveation can yield substantial bitrate savings without decreasing subjective quality. In some of the tested videos, bitrate reductions of up to 50% due to off-line foveation were found compared to unfoveated video. However, the degree of bitrate savings largely depends on the type of the video, and what type of viewing behavior the video elicits.

Part II is structured as follows. Chapter 8 provides a brief introduction to image and video compression, gives an overview of viewing behavior while watching video, and presents previous work in foveated image/video coding. Chapter 9 presents our initial work on off-line foveation video coding, where we get an estimate of its potential in compression. The chapter is based on results from Nyström, Novak, & Holmqvist, 2004. The highlights of Part II are given in Chapter 10, originating from the work published in Nyström & Holmqvist, 2007a, 2008. Here, a full-scale implementation and evaluation of off-line foveated video is undertaken. Finally, Chapter 11 summarizes our findings and discusses the practicability and potential of using off-line foveation in real-world applications.



(a)

(b)

Figure 1.2: Example of off-line foveation. The upper video frame shows where people look in the original frame and the lower video frame depicts the same frame after off-line foveation. Each crosshair represents one viewer's position of gaze.

Part I

Gaze Behavior in Images

Chapter 2

Human Visual System

KNOWLEDGE about the evolutionary optimized design as well as the functionalities of the human visual system (HVS) is key to understanding, implementing and evaluating systems for visual communications. This chapter gives a brief overview of some properties of the HVS. It describes the anatomy of the eye, visual acuity, the visual pathway, eye-movements, visual attention, and reviews evidence for a coupling between eye-movements and visual attention.

2.1 Physiology of the HVS

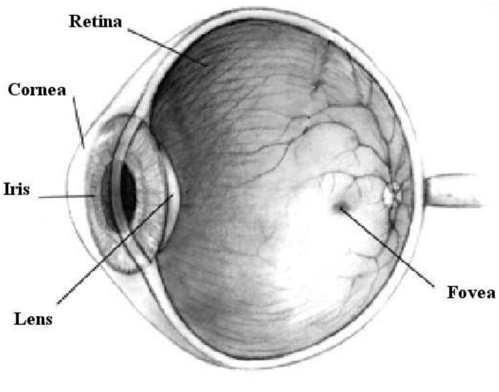


Figure 2.1: Structure of the human eye (Modified from Wikipedia, 2008b)

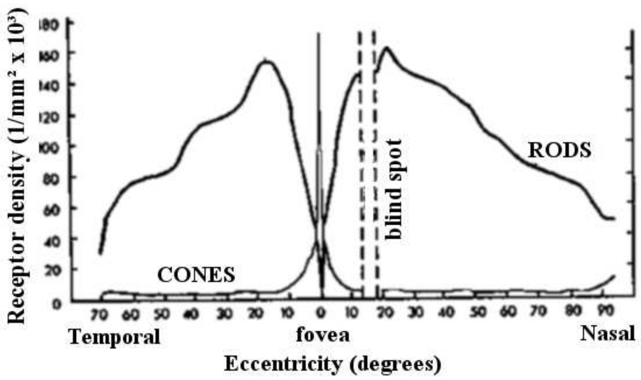


Figure 2.2: Distribution of rods and cones on the retina (Adapted after Osterberg, 1935)

Figure 2.1 shows a cross-section of the human eye. At the first stage of processing, incoming light reaches the cornea, which together with a flexible lens focuses the light on the retina. The cornea has more refractive power than the lens; approximately 70% of the refractive power is provided by the cornea.

On the inside of the eye ball lays the retina, which comprises a set of neural layers. The retina is sensitive to light and holds two different types of photo-receptors involved in vision: *rods* and *cones*. Rods are sensitive to illumination, total 70-150 million per eye and are found over the entire retinal surface. Since many rods can share the same nerve ending they reproduce visual details quite poorly, typically yielding a coarse, gray scale image of the world. However, rods are invaluable due to their sensitivity to dim light, and provide night vision. In addition to rods, about seven million cones serve high acuity color vision. Cones are densely packed within a small part of the retina called the *fovea*, and are increasingly more sparse away from the fovea. With cones humans can resolve fine details in fovea since each cone is connected to one nerve end. Figure 2.2 illustrates the distribution of rods and cones on the retina. The fovea subtends approximately 2° of visual angle. In other words, if we look straight ahead, we have sharp vision only in the central 2° of vision. Regions outside the fovea are usually divided in two different parts: the *parafovea* and the *periphery*. The parafovea is the area outside of the fovea extending over 2-5° of the visual angle. Due to the steep drop of cones, vision is reduced in the parafovea compared the fovea itself. The periphery suffers from very poor acuity, and no detailed spatial information can be acquired from this part of the retina. However, peripheral vision has other important functionalities such as guiding eye-movements, and is also very sensitive to motion. One can rather easily get a feeling for foveal acuity (and the lack of detail outside the fovea): fixate a word in the book you are reading and then try to read the next or previous two words. This is a very difficult task without moving the eyes.

In early retinal processing, rods and cones translate incoming light to action potentials, which are propagated to higher neural layers in the retina where bi-polar cells provide some basic visual processing such as edge detection. In later stages of retinal processing, ganglion cells transmit neural signals to the brain through the optic nerve. They leave the eye through a part of the retina where no receptors exist. Thus we cannot see an object falling onto this part of the retina, hence the name 'blind spot'. Figure 2.3 depicts how the visual input is transmitted to the visual cortex through dedicated pathways; information leaving the retina through the optic nerve is passed to the lateral geniculate nucleus (LGN), which forwards the input primarily the the visual cortex, even though smaller pathways directly to the superior colliculus (SC) exist. Neurons in primary visual cortex, V1, are typically activated by simple features such as orientation, color, intensity, and contrast. V2-V5 represent regions of the visual cortex that facilitate higher level interpretations of the visual input. Typically, direct sensory information together with information processed in higher regions of the visual cortex are combined in the SC to trigger eye-movements. The exact topology of the visual cortex and how it activates motor control for eye-movements currently remains a hot topic of research.

2.2 Eye-Movements – Basic Facts

A general problem in biological systems is information overflow, that is, large amounts of sensory information are constantly fed to the system, which does not have the resources and time for processing and interpretation. The HVS is no exception; the retina has been estimated to receive up to 10^9 bits of (Shannon) information per second (Kelly, 1962). The evolutionary design to handle this huge amount of information is solved by a foveated system, which uses sparse visual input from the periphery to guide the fovea to regions with potentially important or relevant information through eye-movements. In fact, we constantly move our eyes three to four times per second for this purpose. Foveal information is not only acquired with higher detail than other regions on the retina, but is also processed by a disproportionally large part of the visual cortex. This is known as cortical magnification.

To move our eyes, different types of eye-movements are employed; shifting our gaze from one location to another is called a *saccade* and between these shifts the eye remains relatively stable in a *fixation* (typically around 300 ms). However, the eye is not completely stable during a fixa-

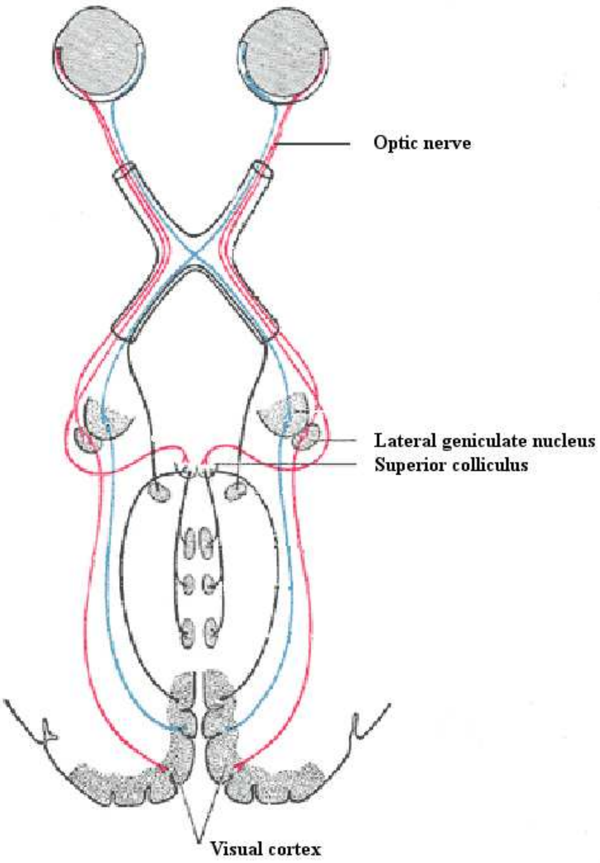


Figure 2.3: The visual pathway (Modified from Wikipedia, 2008a).

tion, but three types of small fixational eye-movements occur (Martinez-Conde, Macknik, & Hubel, 2004): *tremor*, *drifts*, *micro-saccades*. Tremor are the smallest eye-movements, having a frequency around 90 Hz. The role of tremor in vision is unclear, but is believed to help maintaining vision by preventing retinal stabilization. It has been observed that by stabilizing visual input on the retina, the impression of vision slowly fades away. Drifts are movements that slowly move the eye away from the point of fixation, possible due to lack of precision or fatigue of the oculomotor system. This is compensated for by micro-saccades, small corrective movements, which rapidly guide the eye back to its initial position.

A saccade is a rapid eye-movement and therefore sensitivity to visual input is significantly impaired. The speed of a saccade can be up to 1000 degrees per second and its length varies over 1-30°. However, this depends on factors such as task and stimuli. Typically, the time it takes for the eye to move from one location to another during a saccade is 30-70 ms. As expected, long saccades take more time than short.

Pursuit eye-movements occur when the eye follows slow moving objects. Compared to saccadic eye-movements, pursuit eye-movements are considerably slower. This type of eye-movement is generally not possible to evoke without a moving target for the eye to track.

Another type of eye-movement is called *vergence* and occurs when the eyes move toward each other in order to fixate on close objects. If the head moves, but the gaze is kept on the same target, *vestibular* eye-movements have been used to compensate for head movements.

Depending on type of task (silent reading, oral reading, visual search, scene perception, music reading and typing), fixation duration and saccade length can vary considerably.

For a more comprehensive overview on the basic properties of eye-movements and their significance in visual cognition, refer to the overviews by Rayner, 1998; Henderson & Ferreira, 2004; Rayner & Castelhano, 2007.

2.3 Eye-Tracking and Its Applications

For quite some time it has been known that eye-movements provide valuable insights in cognitive processes. However, high precision eye-trackers are relatively new and the number of papers using eye-tracking as a measurement tool are quickly increasing. There is a range of available techniques and apparatus as well as methodological concerns using eye-tracking, and accurate eye-tracking is of course essential to get valid data. Further information about eye-tracking and related issues can be found in the books by Duchowski (2003) and Holmqvist (2009).

Eye-tracking applications have been reported from for a wide range of disciplines, for example neuroscience, psychology, industrial engineering and human factors, and has been divided in two broad areas: *diagnostic* and *interactive* (Duchowski, 2002). In diagnostic applications, the tracked eye-movements are analyzed off-line in order to assess or to obtain objective and quantitative measures of a viewer's overt visual attention. For example, successful studies have been performed on subjects with schizotypy (O'Driscoll, Lenzenweger, & Holzman, 1998) and autism (Klin, Jones, Schultz, Volkmar, & Cohen, 2002) where eye-tracking data show indications of sickness due to deviating eye-movement behavior. In applications where a system responds or interacts with recorded eye-movements in real-time, it is said to be interactive. An example of an interactive system is real-time, gaze-contingent foveation, where the resolution of a display changes contingent on viewers' position of gaze.

In psychology, eye-tracking has become an invaluable tool to study different aspect of visual cognition in *reading*, *scene perception*, and *visual* *search*. One comprehensive source of how the usage of eye-tracking has progressed over the years is the review by Rayner (1998). He compiles 20 years of eye-movement research mostly covering the cognitive mechanisms in reading, and provides a range of basic information about reading behavior: When reading English, the fixation duration is typically 225-275 ms and saccade length about 8 letters; readers do not exclusively go forward in the text but use small saccades to the left called *regressions*; good readers tend regress less frequently than bad readers and the number of regressions increase as texts grow more conceptually difficult; silent reading is faster that reading aloud. Obviously, these types of observations would be cumbersome without modern eye-tracking technology.

Scene perception is another field that has benefited significantly from the evolution of eye-tracking. Unlike reading, scene viewing produces less systematic eye-movement across viewers. In part, this can be explained by the classical observation made by Yarbus (1967) that task influences eye-movements. While the task in scene viewing is not always well defined, reading follows certain rules with the overall goal to comprehend the text. Eye-tracking in scene perception has particularly been used to investigate the influence of higher and lower factors to gaze guidance, which typically is done by analyzing fixated image content.

A more constrained type of scene perception is *visual search* (see Wolfe, 1998) where subjects are asked to search for targets until they are found, or until subjects are ensured that the target is absent in the display. While response buttons can measure search and reaction times, eye-movement data yield a rich collection of perceptual measures indicating the allocation of attention during the search.

Eye-tracking has been used in other areas such as monitoring eye-movements of drivers, pilots, in newspaper design and advertising, and also gaze contingent displays and computer graphics. As eye-tracking technology gets more portable, easier to use and cheaper, the potential for eye-tracking applicability is expected to grow substantially. One example of a future application with huge potential is to integrate eye-trackers with computer games, opening a whole new world of opportunities for rapid and intelligent game interaction. Further information about applications can be found in the overview by Duchowski (2002).

2.4 Visual Attention

Generally speaking, attention refers to the ability to focus most of our cognitive resources to limited or relevant parts in our environment, while largely ignoring other parts. In visual attention, these resources can refer to the ability of the HVS to focus on the most relevant and interesting visual elements in the environment, and allocate proper parts in the brain to process this information with priority. Visual attention is commonly divided into overt and covert attention. Overt attention is of a direct measurable nature, and is aligned with the eye-movement. Covert attention is a mental state of attention and cannot be measured explicitly; it is sometimes described as a mental 'spotlight' preceding overt attention (Posner, 1980).

2.4.1 Bottom-up and top-down processing

Perceiving visual information can be seen as a hierarchical process; visual input propagates from lower cognitive levels to higher, more complex levels where the information gets increasingly more tangible (Levine, 2000). Within this framework, attention may be responsible for integrating, or 'gluing' simple features into whole objects (Treisman & Gelade, 1980). Moreover, it is argued that higher cognitive levels can influence the decisions at lower levels through feedback. These two processes are often referred to as *bottom-up* respectively *top-down* processing. Bottom-up processing consists of rapid, spontaneous and automatic decisions and is purely stimulus dependent and computed in parallel. Top-down processing on the other hand reflects higher cognitive mechanisms controlled by factors such as task, context and linguistic input, and is believed to be slower than bottom-up processes. In scene viewing, bottom-up processing refers to a quick, involuntary response after image onset to saccade toward low-level features such as color, motion and contrast while top-down guidance is influenced by factors such as task-dependence (e.g., remember image objects, object search) as well as prior knowledge and experiences (e.g., faces are important in human communications).

Although the metaphorical model of bottom-up and top-down processing outlines an important conceptual model in cognitive psychology, it is also subject to quite some confusion. One key issue concerns which parts of the brain comprise the 'top' and, likewise, the 'bottom' (cf. Roepstorff & Frith, 2004). In an anatomical sense, the bottom can refer to the 'reptile' brain, whereas the top would comprise more developed mammalian parts of the brain. However, dividing the brain into sections responsible for top-down and bottom-up processing has shown to be elusive, partly since the functions within and interactions between different parts of the brain cannot be fully explained. The top and bottom can also refer to an organism and its sensory input. The top is then controlled by the organism's mental world, whereas bottom-up control is modulated by the organism's physical input. Today, the interplay between bottom-up and top-down processing in scene perception as well as how they contribute to different actions are not completely understood.

2.4.2 Coupling between eye-movements and visual attention

A recurring question directed to researchers using eye-tracking to study visual attention touches the relationship between the position of gaze (overt attention) and location of our internal (covert) attention. There is a large body of research devoted to the relationship between visual attention and eye-movements. While it has been shown that eye-movements quite easily can be separated from covert attention in simple discrimination tasks (Posner, 1980), there exists ample evidence that that this coupling generally is quite tight (Deubel & Schneider, 1996), especially when scenes grow more complex (see e.g., Henderson & Ferreira, 2004).

The connection between saccadic programming and shifts in covert attention has been extensively researched through clever visual search and discrimination task experiments. Deubel and Schneider (1996) used a letter discrimination task where subjects were asked to fixate a cross in the center of a display, and simultaneously prepare a saccade to a cued location. Before the saccade was initiated, the discrimination letter appeared briefly either at the cued location or adjacent to the cued location. Results showed that letter discrimination increased significantly when the cued location coincided with the position of the discrimination target. This finding supports the coupling hypothesis - that it is not possible to prepare a saccade to a target without first directing attention to it. If the contrary were true, attention could have been directed to the discrimination letter independently of the programmed saccade target. As a consequence, letter identification would be successful even if the location of the discrimination letter would differ from the intended saccade landing location. These results are in line with the widely believed claim that covert attention precedes saccadic eye-movements and thus is used to guide the eyes to interesting regions in a scene.

2.5 Summary

The foveated nature of the HVS is highly efficient and addresses the trade-off between the huge amount of information constantly available and the limited computational resources of information processing in the brain. Information from our visual surroundings is gleaned through eye-movements, directing high acuity vision to potentially relevant of interesting regions in our environment. This chapter described some key properties of the HVS and the types of eye-movements used by humans to explore the visual world, and also how visual information is transported to the brain for further processing. For the natural, complex images, which will be used in this thesis, we pointed to evidence that the coupling between attention and eye-movement is tight.

Chapter 3

Gaze Behavior in Natural Images – The *Where*, *Why*, and *For How Long*

UNDERSTANDING the subtle mechanisms behind eye-movements in scene viewing has shown to be a challenging and interesting problem, and has attracted an increasing amount of attention from researchers using eye-tracking as a measurement tool. Knowledge about visual attention and gaze behavior in scene perception has important application in, e.g., engineering and marketing, to render visual communications more precise and efficient. This chapter reviews the literature on scene perception and eye-movement, and presents some key findings gleaned over the last century.

3.1 Scene Perception and Eye-Movements

A *scene* usually refers to a depiction of an environment, which for example can comprise the real world, an artificial world, or line drawings illustrating real-world or artificial objects. In an experimental setting today, most scenes are viewed as digitized images on computer screens, where it is easy to control experimental parameters such as where the scene is located, the size of the scene, how long the scene is shown and the viewing distance from the scene to the observer. The goals when studying scene perception are multifaceted and involve how people understand and interpret scenes. In this chapter, we review what eye-movements can reveal about the perception about a particular type of scene: *Natural images.* In this thesis natural images refers to digitized photographs depicting natural environments from the real-world, that is, visual input that is typical for the everyday person.

An exciting part in scene viewing concerns the speed and mechanisms of perception, which have been lively debated issues over the past decades. Currently, many aspects of the early perceptual mechanisms remain unclear. There are however some general consensus. There exists evidence that the general semantic category, sometimes called gist, of a image is apprehended very quickly, well within a fixation after image onset but perhaps as quickly as 30-50 ms. Gist is rather ill-defined in the literature but is assumed to include the category of the image (e.g., indoor or outdoor), and some information of the objects and their spatial layout (Henderson & Ferreira, 2004). However, more detailed semantic information of individual objects is not likely to be acquired during this very brief period of time unless the object is large and close to the point of fixation. A recent study by Fei-Fei, Iyer, Koch, and Perona (2007) investigated the amount of information subjects could glean from a set of test images for a number of short presentation times (27 to 500 ms). They found that only a feature level representation of the images could be acquired from the shortest times (27 and 40 ms). However, presentation times well below a typical fixation duration showed to be sufficient to acquire a “rich collection of perceptual attributes” which “raises to conscious memory”. There are some evidence that low spatial frequencies facilitate, but are not mandatory for, initial scene identification, more so than high frequencies (Oliva & Schyns, 1997). Moreover, there is evidence that scene identification is faster when objects are presented in (natural) color rather than in gray scale (Oliva & Schyns, 2000).

To understand how scenes are perceived, it is necessary to understand how the eyes move to provide us with the information that optimally facilitates perception. Knowing the position of the eye and for how long it stays at each position provides valuable insight into what is sent to the brain, and thus comprising a basis for perception.

Studies of scene perception through eye-tracking have been conducted for over a century. Initial studies were based on direct observations of how humans moved their eyes while watching different stimuli. Two of the most frequently cited early studies in picture viewing were performed by Buswell (1935) and Yarbus (1967). Buswell used a simple but ingenious device to record eye-movements while participants viewed pictures, and made a number of important observations. For example, he noted that certain image regions attracted substantially more fixations than others, and that differences in eye-movement locations were large across subjects. Besides Buswell’s work, Yarbus’ book about eye-movements and vision is one of the most well-cited studies in the history of eye-movement research. To a large extent, he replicated and expanded the findings of Buswell. Perhaps the most cited of Yarbus’ observations is that the task heavily

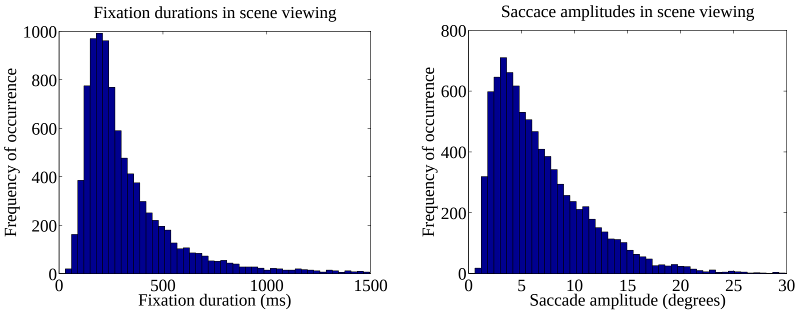


Figure 3.1: Typical fixation duration and saccade amplitude in scene viewing. Eye-movements were recorded from viewers looking at images presented on a 19 inch screen from a distance of approximately 70 cm.

influences where people look in pictures. Considering the limited technical equipment used by Buswell and Yarbus, the results of these early studies were remarkably fruitful and outline much of today's work.

Today, knowledge about how the eyes move in scene viewing is well documented (cf. e.g., Rayner, 1998; Henderson, 2003; Henderson & Ferreira, 2004). For example, scene viewing elicits somewhat different, more unconstrained, eye-movements than for example reading and visual search. Typically, both the fixation duration and saccade length are on average slightly larger in scene viewing. The fixation duration is usually around 300 ms and the saccade length 2-20 degrees. However, fixation duration and saccade length can vary significantly with the distribution of low-level image features, image semantics, task, size of stimulus, type of stimulus, etc. Figure 3.1 illustrates histograms of typical distributions of fixation duration and saccade amplitude. The figures are generated with data collected from subjects free-viewing natural images during five seconds.

3.2 Factors That Influence Where We Look

Eye-movements are generally guided toward a small portion of the total image area considered more interesting, relevant, or informative than other regions. What makes an image region have these inherently ill-defined attributes largely remains an open question, central in many recent studies aiming to unravel the causes behind fixation selection. Specifically, the interplay between bottom-up and top-down factors in fixation selection has been investigated in several recent eye-tracking studies.

In favor of a bottom-up perspective, there is some evidence that attention, and hence eye-movements, quickly and effortlessly are guided toward certain regions based on low-level features in the image (Treisman & Gelade, 1980). These features can be contrast, color, luminance, and spatial frequency. In agreement with this evidence, there are eye-tracking studies showing that fixations on average land on regions with higher feature densities than control regions. For example, it is known that fixated regions contain higher contrast (Reinagel & Zador, 1999; Parkhurst & Niebur, 2003; Parkhurst, Law, & Niebur, 2002; Einhäuser & König, 2003; Tatler, Baddeley, & Gilchrist, 2005; Henderson, Brockmole, Castelhano, & Mack, 2007; Rajashekar, Linde, Bovik, & Cormack, 2007) and edge density (Mannan, Ruddock, & Wooding, 1996; Tatler et al., 2005; Baddely & Tatler, 2006; Henderson et al., 2007) than control regions. It has also been reported that high levels of luminance correlate with fixation locations (Tatler et al., 2005; Rajashekar et al., 2007), although lower than control luminance at fixated regions was reported by Henderson et al. (2007).

The influence of bottom-up features on eye-movements has been studied through computational frameworks by computing a *saliency* map, i.e., the distribution of saliency over an image, and then measure how saliency coincides with human fixations. Saliency is defined as a weighted combination of a candidate set of low-level primitives, and peaks in a saliency map point to regions likely to be visually attended (Itti, Koch, & Niebur, 1998; Itti & Koch, 2000). Salience has shown to correlate with gaze positions better than at random (Parkhurst et al., 2002), and has recently been reported to coincide with image regions deemed as important by human viewers (Elazary & Itti, 2008). Parkhurst et al. (2002) and Itti (2006) argue that saliency is more influential early after stimulus onset than later in viewing. However, these findings are not supported by Tatler et al. (2005), who found that bottom-up features are equally influential over time, whereas top-down influences increase as a function of viewing time. Since the acuity of the HVS drops quickly as a function of eccentricity[1] and thus prevents high frequencies from being registered by peripheral vision, the correlation between feature content and fixation locations decreases as a function of saccade length (Rajashekar et al., 2007). Tatler, Baddeley, and Vincent (2006) found only short saccades (≤ 8 degrees) to be feature dependent, whereas longer saccades show no such tendencies. Obviously, the landing positions of long saccades are hard to predict given the feature content available in the periphery of a viewer when the saccade is initialized.

Despite the recent popularity of computational models of visual attention dominantly relying on bottom-up features, it is an undisputed fact that higher cognitive factors are highly involved in the attentional processes preceding eye-movements. Some factors known to influence where people look are short and long term episodic memory and scene schema

[1] Angular distance from the fixation point

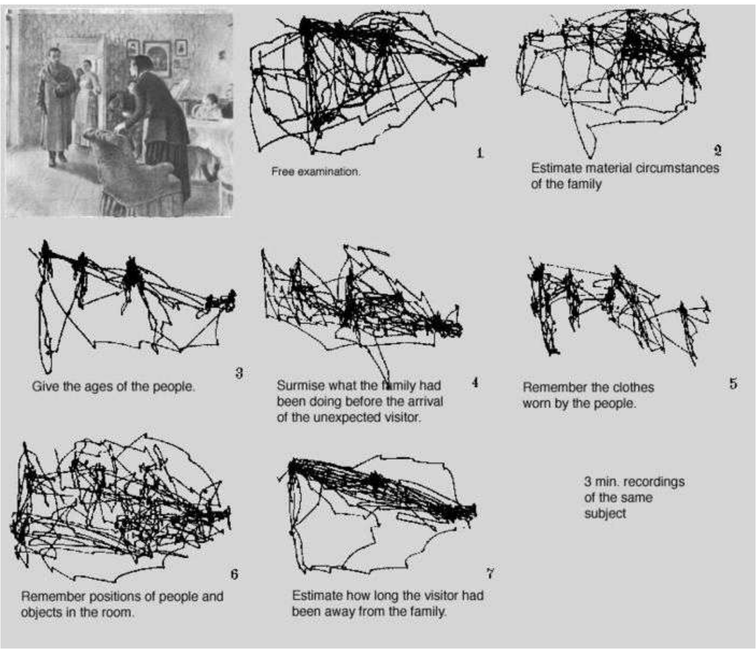


Figure 3.2: Influence of task on eye-movements - a classical example from Yarbus (1967).

knowledge (cf. Henderson & Ferreira, 2004 for an excellent review). An old, well known example of top-down influence on eye-movements is Yarbus' experiment using a painting named 'The unexpected visitor' containing a number of people in a room. Depending on the instruction given to the viewer prior to image onset, which could be to estimate peoples' ages or remember the positions of people and objects in the room, different viewing pattern were observed. Figure 3.2 illustrates the eye-movement pattern elicited by different viewer instructions. The significant influence of context and task on eye-movements has been replicated and extended by several other studies (Lipps & Pelz, 2004; Rothkopf, Ballard, & Hayhoe, 2007; Einhäuser, Rutishauser, & Koch, 2008). As when viewing images on computer screens, eye-movement guidance in everyday activities seems to be even more about task and context (M. Land, 2007). Differences in eye-movement behavior have also emerged due to gender (Rupp & Wallen, 2007) (men look more toward faces in sexually explicit images, whereas women look more toward genitals or the background); cultural differences (Chua, Boland, & Nisbett, 2005) ("Westerners attend more to focal objects, whereas East Asians attend more to contextual information."); and between experts and novices (Law, M. Atkins, Kirkpatrick, Lomax, & Mackenzie, 2004). All this despite being engaged in the same task and context. Moreover, it is also well known that eye-movements are reflected by linguistic input; this is extensively researched using the *visual world paradigm*, in which the interplay between, for example, when an object is mentioned and when this object is fixated is investigated. Evidence of linguistic control of eye-movements can be found in anticipatory eye-movements where objects expected to be uttered are gazed upon, or when the mentioning of an object elicits eye-movements to a part of a blank screen where this object previously was located (Johansson, Holsanova, & Holmqvist, 2006). Clearly, such eye-movements originate from internal mechanisms.

Lately, the saliency map hypothesis as well the empirical evidence showing a coupling between certain low-level features and fixations have been challenged by a series of studies. Einhäuser and König (2003), for example, show that moderate changes in local contrast at a number of image regions do not change where subjects fixate, as would be expected by a bottom-up predictor tuned toward contrast. Moreover, it has been shown that bottom-up predictors such as the one presented by Itti et al. (1998) easily can be cognitively over-ridden by changing the task instructions during viewing (Underwood, Foulsham, Loon, Humphreys, & Bloyce, 2006; Einhäuser et al., 2008). Interestingly, experiments by Henderson et al. (2007) report that fixated locations not only contain high densities of certain low-level features, but also are judged as more semantically important than control regions. Together, these results raise questions about the causes behind the measured correlations between low-level features and fixated image content. One specific question is whether this effect is simply *correlative* or in fact *causal*. A causal effect would imply that fixation locations are chosen as a direct consequence of the signal strength of one or a set of combined low-level primitives. A correlative effect, on the other hand, would mean that fixations land on regions that happen to contain high feature densities, but are in fact guided to these regions by other, higher level mechanisms. For example, objects may be fixated since they contribute to the semantic representation of the scene, and not because they happen to contain high contrast. It is hardly speculative to claim that certain objects are fixated due to their semantic contribution to the scene, and not mainly because they happen to contain, e.g., high contrast or edge density.

A well known observation is that eye-movements (positions) are strongly biased to the center of the display (see e.g., Tseng, Carmi, Cameron, & Munoz, 2007; Tatler, 2007). This tendency is shown in Figure 3.3, which plots fixation locations from eight subjects free-viewing 30 images. Interestingly, Tatler (2007) found this central bias to be largely independent from both feature distribution and task. Instead, he suggests three alternative explanations: “First, the center of the screen may be an optimal

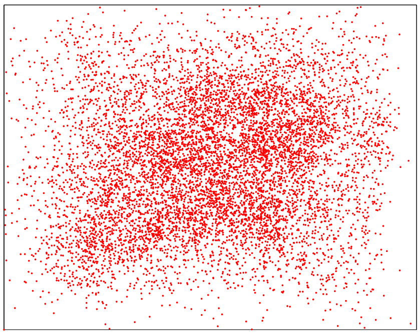


Figure 3.3: Central bias effect of eye-movements. Each dot represents a fixation.

location for early information processing of the scene. Second, it may simply be that the center of the screen is a convenient location from which to start oculomotor exploration of the scene. Third, it may be that the central bias reflects a tendency to re-center the eye in its orbit.". Besides that gaze positions are biased toward the center of the display, previous and future eye-movements influence where we look (Tatler & Vincent, 2008, in press). For example, long fixations tend to be followed by long fixations and we have a tendency to execute the current saccade in the same or the 180 degree opposite direction as the previous saccade. Overall, Tatler and Vincent suggest a global and local relocation of gaze; long global saccades take us to new image regions whereas short saccades are employed in local scanning to scrutinize a limited image area in detail.

3.3 Summary

What controls where we look and for how long we look there? There is ample evidence that eye-movement guidance in scene viewing is determined by a combination of bottom-up, external factors, i.e., the physical properties that compile the scene, and top-down, internal factors, which reflect a complicated interplay between higher cognitive processes. However, the spatial and temporal manners in which these factors interact are still elusive. Currently, the attentional mechanisms behind eye-movement control are slowly starting to unravel, but unanimity among explanations is surprisingly low considering the large number of papers published on the subject.

Chapter 4

Effects of Contrast Manipulations on Gaze Locations

PREVIOUS chapters have shown that there exists a large body of research on gaze behavior in natural images, and that the results are somewhat incongruent: on the one hand, people emphasize the contribution of image based saliency to gaze guidance while at the same time it is known that top-down factors largely influence where people look.

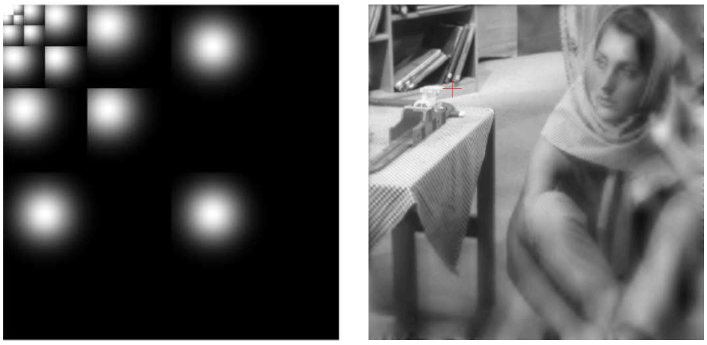
It has be argued that one of the problems in eliciting the causes behind fixation selection is the lack of experimental manipulation of the natural images (e.g., Henderson, 2007). While it is common to use different viewing instructions, which are known to influence eye-movements, to argue for the important role of higher level factors to gaze guidance, it is much less common to use a neutral task and instead alter the low-level content of the image. An exception is the work done by Einhäuser and König (2003) who used a new experimental paradigm where natural images were contrast manipulated at five randomly chosen points; contrast was either decreased or increased smoothly around these points. Eye-movements were recorded from viewers watching the contrast manipulated images and an analysis revealed that contrast by itself was not a good predictor of fixation locations. They observed that moderate changes in contrast affected fixated locations very little, whereas strong reductions in contrast attracted fixations. This is inconsistent with previous research that found a significant correlation between high contrast and image content at fixations. However, their results were disputed by Parkhurst and Niebur (2004), who pointed to a number of methodological flaws. First, the same image (with slight modification) was seen by each subject multiple times. This gave subjects the possibility to encode the images as well as the locations of the manipulated image patches into memory over the trials, and potentially use this top-down information during later inspections. Second, Parkhurst and Niebur criticized the lack of stimulus control; while changing the local contrast, Einhäuser and König also altered the local luminance in this regions, making is difficult to relate the changes in fixation locations to contrast manipulations alone. Finally, Parkhurst and Niebur criticized the introduction of undesired changes in second order statistics due to first-order contrast manipulations. Specifically, they argue that 'texture contrast', defined as the 'contrast of the contrast', was altered and thus acted as a causal attractor for fixations. In fact, using a bottom-up model (Itti et al., 1998) tuned toward texture contrast to predict fixations on the image set used by Einhäuser and König, Parkhurst and Niebur found texture contrast to predict fixations quite well.

Despite the criticism by Parkhurst and Niebur, we believe that proper use of the contrast manipulation paradigm can serve as a useful tool to dissociate objects from their low-level signal strength, and therefore elucidate possible relationships between gaze guidance and image features from a new perspective. In the current and following two chapters, we will use contrast manipulated images to estimate the relationship between bottom-up and top-down processing on eye-movements in image viewing; eye-movement will be measured from subjects viewing natural images with manipulated low-level statistics while engaged in rather neutral tasks (“free-view the images”, “inspect the images carefully”). We will address the issues brought up by Parkhurst and Niebur in our experimental design.

This chapter presents two experiments. In the first, *Experiment I*, eye-movements are collected from viewers watching 39 images. Thirty of these are shown in their original form whereas three of the images are shown both with and without contrast manipulations. Each of the three images is displayed in three version: One unprocessed and two versions that are contrast manipulated at locations specified by the experimenter. In *Experiment II*, contrast is modified contingent on where people looked in the unprocessed images from the first experiment. A new group of test subjects then views these images under the same experimental conditions as in Experiment I. Besides investigating how contrast manipulation affects gaze behavior in these experiments, we analyze how contrast statistics around gaze positions are affected by the image manipulations.

4.1 Implementing Contrast Manipulations

Variable image contrast is implemented in the wavelet domain (cf. Appendix A) by multiplying a wavelet decomposed image with a Gaussian



(a) Wavelet mask used to introduce a varying contrast. Five levels of wavelet decomposition were used in this example.

(b) Variable contrast image computed by using the mask in (a).

Figure 4.1: Implementing a varying image contrast. The contrast is smoothly reduced away from the chosshair in the upper left corner.

mask. Such a mask with five levels of decomposition is exemplified in Figure 4.1(a). The brightest areas in the mask represent unit values whereas the dark areas represent values close to zero. In order to achieve a smooth contrast degradation, the mask is generated by centering a 2-D Gaussian function with standard deviation $\lambda\sigma$ in each wavelet subband at the position marked by a crosshair in Figure 4.1(b). λ denotes the decomposition level, where $\lambda = 1$ represents the highest frequency level. Figure 4.1(b) illustrates the resulting variable contrast image after inverse transformation. If instead the region around the crosshair is to be degraded, each Gaussian function is inverted, normalized to unit height, and its standard deviation is set to $(L - \lambda + 1)\sigma$. L denotes the number of decomposition levels. The choice of parameters (σ and number of decomposition levels) were experimentally tuned to introduce noticeable blur in desired parts of the image. When implementing varying contrast in color images, each color component (R,G, and B) was manipulated separately as just described.

4.2 Experiment I – Manually Controlled Contrast Reduction

The purpose of Experiment I is to investigate how eye-movements are affected by contrast manipulations. We observe qualitatively how gaze guidance to objects with high cognitive saliency, such as human faces, interplay with lower level features such as high/low image contrast.

4.2.1 Subjects

Eight naive subjects (two females, 32.6±7.6 (M ± SD) years old), students and staff at Lund University, volunteered to take part in the experiment. All subjects had normal or corrected-to-normal vision.

4.2.2 Stimuli

In total 39 images (in gray scale and color and of various dimensions) were used in the experiment. They are commonly used by the image compression community and depict a range of different image types such as natural outdoor scenes, humans, and computer generated images, as shown in Figure 4.2. Among the images, there are three images that each is represented in three different versions: One original version and two versions with different configurations of variable spatial contrast. Larger prints of these nine images can be seen in the left columns in Figures 4.3, 4.4 and 4.5. The reason for using more images than those with manipulated contrast was threefold. First, since three different versions of three of the images are shown during the presentation, there will be undesirable memory-driven influences on eye-movements if the versions were shown directly after each other. To alleviate this effect other images are mixed in with the contrast manipulated versions. Second, eye-movements are collected from all images in preparation for the second experiment where contrast is manipulated contingent on gaze density instead of subjective decisions. Third, recorded gaze positions from unaltered images are used as a baseline measure during the analysis in Experiment II.

In the current experiment high and low contrast regions were chosen to compose the facial/non-facial regions in the two images containing faces (`Barbara` and `Kodak`) and two arbitrarily defined regions in `Peppers`, which contains no obvious region of interest. Contrast manipulations were implemented as described in the previous section by centering Gaussian/inverted Gaussian functions with $\sigma = 0.10M$ at the desired regions. M denotes the horizontal image dimension. Five levels of wavelet decomposition were used.

4.2.3 Procedure

Subjects were seated in front of a 19 inch (37.7×30.5 cm active display area) flat screen (of resolution 1024×768 and an update rate of 75 Hz) where the screen area subtended a visual angle of 27.7 degrees horizontally and 22.5 degrees vertically. They were asked to place their heads on a chin rest positioned 76.5 cm from the screen.

A session started with a 13-point calibration and after verifying the accuracy of the calibration, the 39 test images were displayed one by one in a random order. Each image was displayed for five seconds and

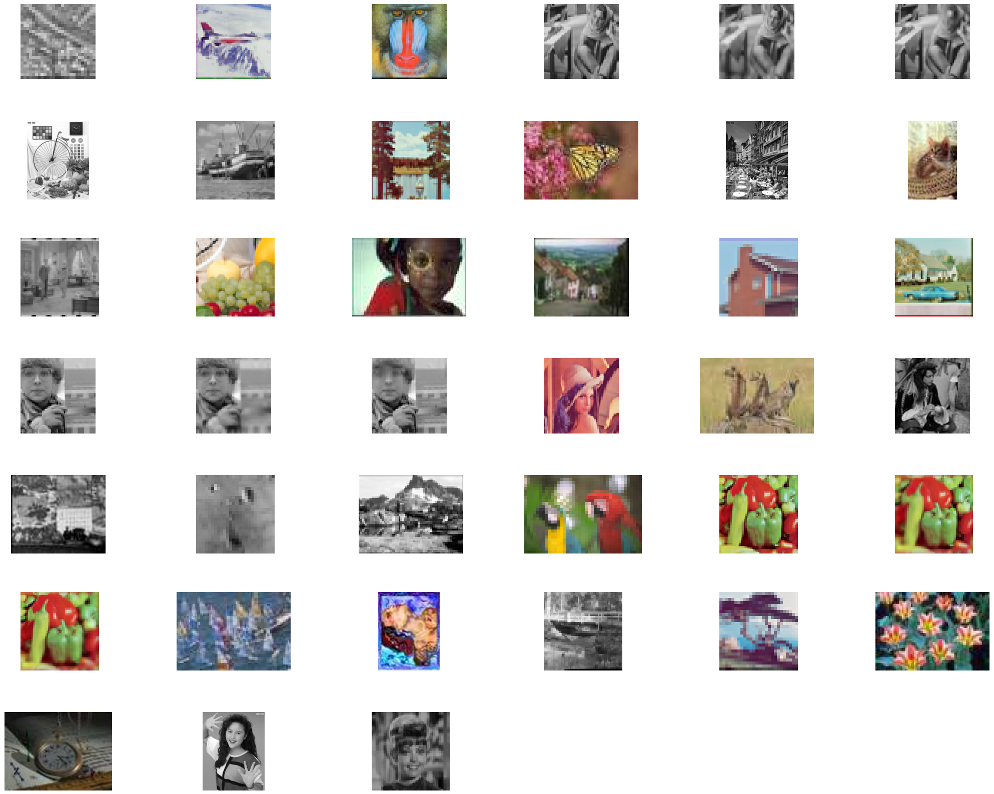


Figure 4.2: Test images used in Experiment I.

between two subsequent images a mid-gray image was shown for one second. Images were displayed in full screen while maintaining their aspect ratio. No pre-stimulus fixation marker was used to constrain the position of subjects' initial gaze position.

Subjects were given no specific task and were asked to 'free-view' the images. Before a session started, they were introduced to the presentation setup and were shown a trial presentation with images not contained in the set of test images. Eye-movements were recorded monocularly with an iView X Hi-Speed eye-tracker, sampling gaze positions at 240 Hz with gaze position accuracy 0.2°. A Matlab program using ActiveX scripting to communicate with the Quicktime media player was developed to control the eye-tracker, display the stimuli and control the accuracy in timing throughout the experiments.

4.2.4 Data representation

Subjects' visual interest is represented and visualized by centering a 2-D Gaussian function at the location of each gaze point and then superimposing all functions belonging to the set of gaze points to be visualized. The variance of each Gaussian function is set such that the full width at half maximum height spans the foveal and para-foveal regions (approximately five degrees of visual angle) of a subject viewing the stimuli presentation. The aggregate Gaussian functions represent the gaze density and are therefore referred to as *gaze density functions* (GDFs). Examples of GDFs represented as so called heat maps are shown in the middle columns of Figures 4.3 through 4.5.

4.2.5 Results

Figures 4.3, 4.4 and 4.5 show different versions of the three manipulated test images (first column). The second and third image columns depict GDFs generated by gaze positions collected during short time intervals; the second columns show where attention is located after subjects typically have launched their first saccade (300-350 ms) and the third columns visualize the distribution of subjects' gaze locations after about twice this time. The fourth image columns correspond to the cumulative distribution of GDFs composing a representative set of collected gaze positions from all viewers over the whole five seconds of viewing.

First, we observe that introducing a variable contrast affects the way subjects look at an image; *total* dwell time is increased in regions remained in high contrast and decreased in regions reduced in contrast. This effect is present in all three tested images. Second, from the second and third columns in Figures 4.3-4.5, it can be observed that the location of the first saccade target seems largely unaffected by a change in image contrast. Rather is it consistent, even when the saccade is directed toward

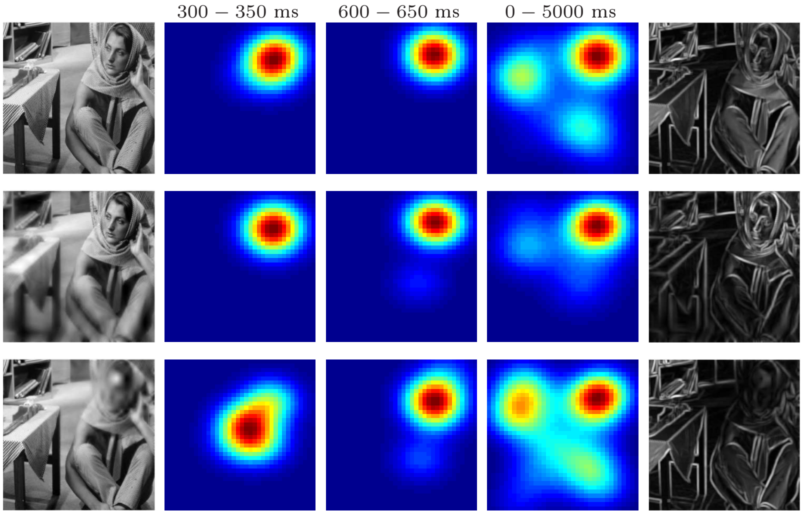


Figure 4.3: Test image `Barbara`. The heat maps visualize how gaze positions from seven viewers are distributed over different time intervals. The rightmost column illustrates the local image contrast.

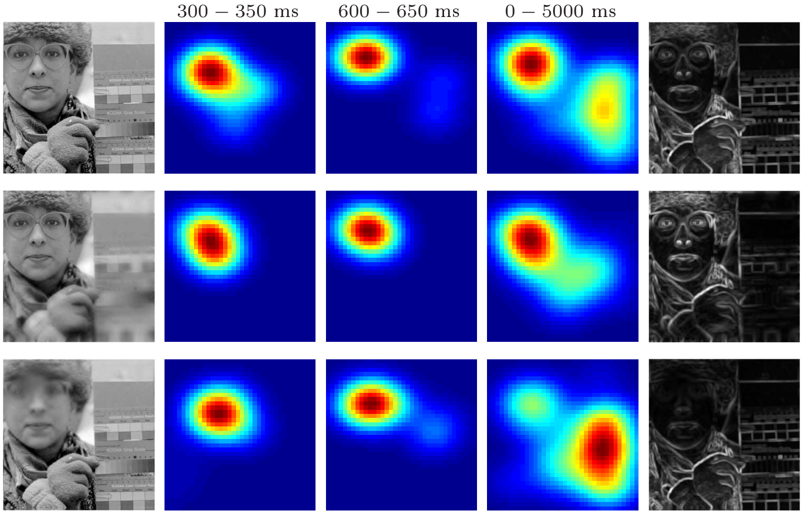


Figure 4.4: Test image `Kodak`.

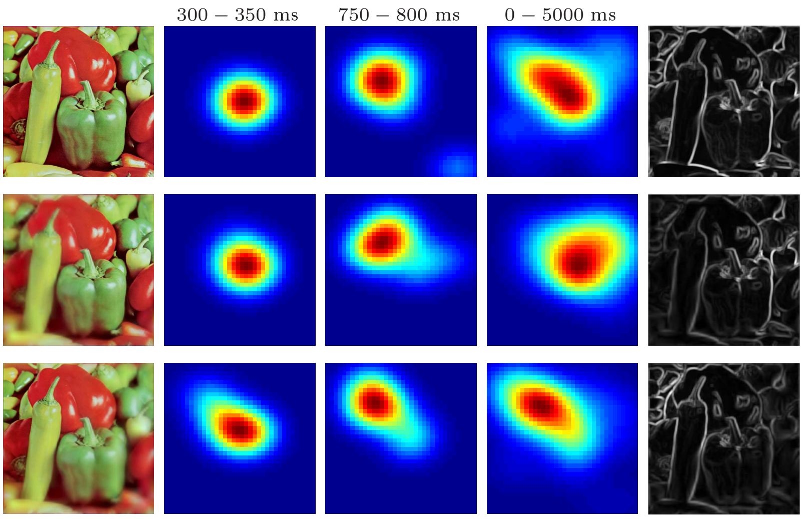


Figure 4.5: Test image `Peppers`.

a region where the image contrast is heavily degraded. In the images containing faces, eye-movements are quickly directed toward the blurred facial regions. Even in the `Peppers` image, which contains no obvious regions of interest, subjects' gaze directions are initially not drawn to the regions of high contrast but instead follow a similar path as in the same, unaltered image. The third observation concerns the initial saccade latency; GDFs reveal that the initial saccade is launched more quickly when it is directed directed toward a high contrast region and at the same time away from a low contrast region. Also, the initiation of a saccade seems to slow down when the saccade target is of low contrast relative the overall image.

The rightmost columns in Figures 4.3-4.5 illustrate the local image contrast, which for a pixel at location (m, n) is defined as the standard deviation within a 15×15 pixel square centered at (m, n). These illustrations clearly show that contrast per se does not have a dominant influence on the location of the initial saccade target, but seems to shift the overall gaze density toward regions kept in high contrast.

4.3 Experiment II - Gaze Density Controlled Contrast Reduction

In this second experiment, we further investigate the results from Experiment I by asking the following questions: 1) What happens with subjects' gaze behavior if regions known to attract overt visual attention are de-

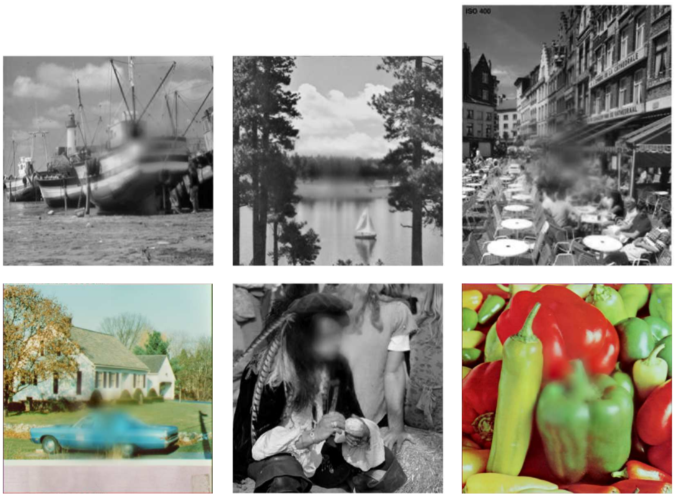


Figure 4.6: Variable contrast images used in Experiment II.

graded in contrast? 2) How do these manipulations affect contrast statistics around viewers' positions of gaze? The reason for degrading regions with a known high probability of attracting gaze is to quantify how features and semantics interact to guide eye-movements toward informative regions. Since the experimental setup and procedure in Experiment II follow that in Experiment I, only differences from the first experiment are described below. If nothing else is mentioned it is assumed that the conditions from Experiment I are fulfilled.

4.3.1 Subjects

15 naive subjects (nine females) of ages 30.2±16.1 (M ± SD) years.

4.3.2 Stimuli

Stimuli consisted of six of the images used in Experiment I, each having its contrast modified in in accordance to the gaze density (as found in the first experiment) from all viewers between $t = 500 - 600$ ms such that regions of high gaze density were reduced, whereas other regions were kept in high contrast. Contrast modifications were implemented as described in Section 4.1, but with a GDF replacing the single Gaussian function in the wavelet mask. The resulting six stimuli images are shown in Figure 4.6. These images were presented to the subjects. Again, the images were shown with another 35 images, not included in the current analysis. $t = 500 - 600$ ms was chosen since the similarity between different viewers'

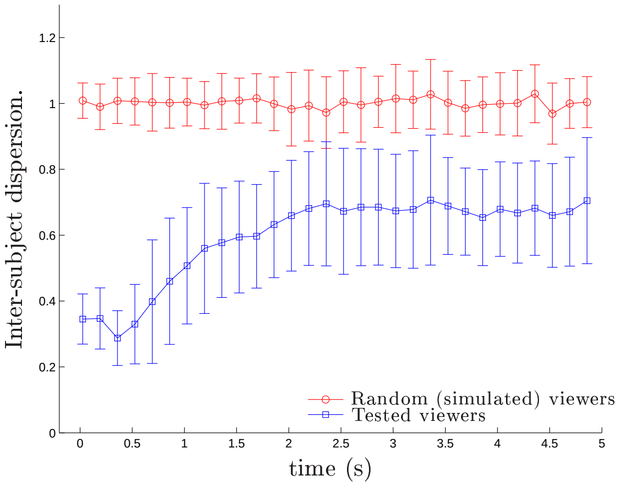


Figure 4.7: Inter-subject dispersion across eight people free viewing the 33 original test images from Experiment I. Error bars span one ± one standard deviation.

gaze positions typically peaks around that time (cf. Tatler et al., 2005), hence identifying regions of particular visual interest. Figure 4.7 confirms this observation for data collected from all 33 (unprocessed) test images in Experiment I. The figure illustrates the degree to which subjects' gaze positions coincide as a function of time after image onset, defined by the *inter-subject dispersion*, S_t, which at time t is calculated as

$$S_t = \frac{1}{P} \sum_{i=1,2,\ldots,P} \frac{G^{i'}_{t,max} - G^{i'}_{t}(m_i, n_i)}{G^{i'}_{t,max} - G^{i'}_{t,avg}} \tag{4.1}$$

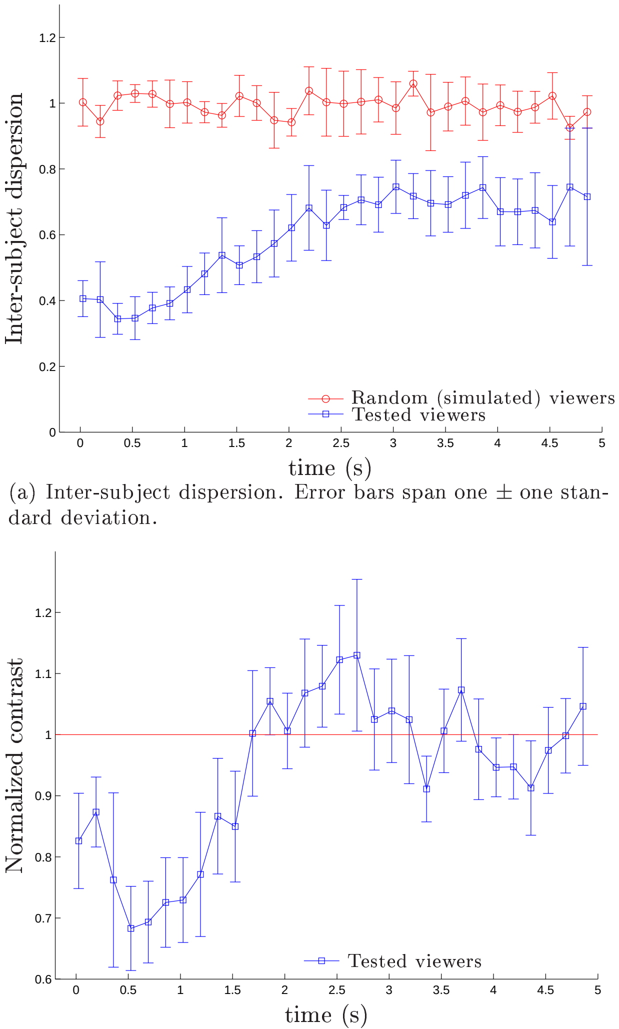
where $G^{i'}_t(m,n)$ denotes a GDF at time t that has been generated by $P-1$ gaze positions collected during the time interval $[t-\Delta t, t+\Delta t]$, excluding the i^{th} gaze location (m_i, n_i). $G^{i'}_{t,max}$ and $G^{i'}_{t,avg}$ denote the maximum and average value of $G^{i'}_t(m,n)$, respectively. To obtain a robust measure of dispersion at time t, Δt was set to 40 ms. Following this notation, $S_t = 0$ indicates that all gaze positions are located at the same spatial location, whereas $S_t = 1$ represents a random distribution of gaze positions. The bottom curve depicts the inter-subject dispersion across the collected gaze positions. Notice the dip in dispersion around 500 ms. As a control, the top curve in Figure 4.7 represents simulated random viewers (whose gaze positions were drawn from a uniform distribution).

4.3.3 Results

Figure 4.8(a) shows the inter-subject dispersion between viewers watching the images in Figure 4.6. As for the unprocessed images in Experiment I, similarity peaks around 500 ms, which typically coincides with subjects' fixation locations after the first voluntary eye-movement. This indicates that, after reducing the contrast in regions where people normally look early after image onset, subjects still largely agree on where to initially move their eyes. However, we cannot tell whether people look at similar regions as the viewers from Experiment I or if they have decided to look at a region elsewhere in the image. One way to approach this issue is by analyzing the image content at fixation. Specifically, how is fixated image content correlated to contrast densities?

The analysis of contrast statistics was limited to gray images. Images presented in color were therefore converted to gray images through an RGB to YUV transformation, where the Y component composed the gray image after transformation. Each image was then resized to match the display resolution it was presented at. After resizing, a pixel subtended the same visual angle in all images. Contrast at each pixel location (m, n) was defined as the local standard deviation of pixel intensities within a square region centered at (m, n). We used squares of size 15×15 pixels. Symmetric padding was used at the image borders. In the analysis below, we have extracted the average contrast from 35×35 pixel squares (roughly corresponding to the foveal part of the visual field) around gaze positions recorded during a range of temporal interval, and normalized it with the average contrast of the whole image. Other square sizes for contrast calculation and analysis of contrast were tested with similar results as those presented below.

Figure 4.8(b) presents how contrast statistics around viewers' gaze positions change as a function of viewing time t. Each box represents the average normalized contrast around each gaze position recorded during time $[t - \Delta t, t + \Delta t]$. The analysis reveals that after about 500 ms, gaze positions land on image regions with lower than average contrast, and are after a while drawn to regions with higher than the average contrast. This suggests that the region(s) attracting many subjects' gaze some hundred milliseconds after stimulus onset indeed are those where contrast has been degraded. For comparison, normalized contrast at gaze locations collected from the 33 test images in Experiment I is given in Figure 4.9. It confirms findings from earlier work that contrast is elevated at gaze locations compared to random locations, which give a unit normalized contrast as indicated by the solid red line in the figure. The low values of contrast right after image onset occur because subjects have not yet completed their first voluntary eye-movement. An interesting, and maybe somewhat surprising observation from Figures 4.8(b) and 4.9 is their large differences in normalized contrast after a few seconds of viewing. This happens since



(a) Inter-subject dispersion. Error bars span one ± one standard deviation.

(b) Contrast statistics at gaze. Error bars span ±1 standard error.

Figure 4.8: Statistics at gaze positions collected from 6 subjects free viewing the 6 variable contrast test images from Experiment II.

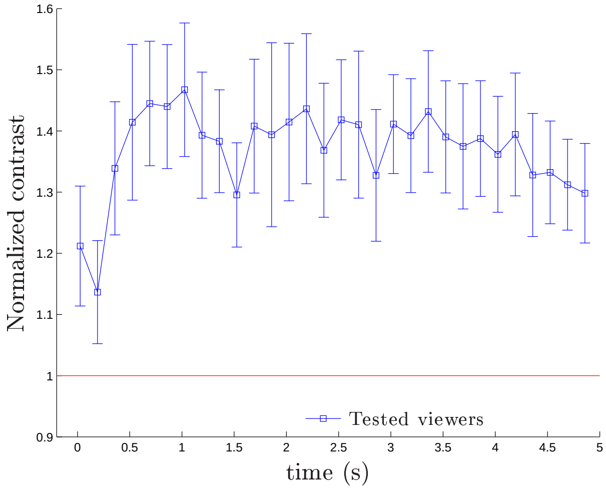


Figure 4.9: Statistics at gaze positions collected from 8 subjects free viewing the 33 original test images from Experiment I.

visually interesting regions are positioned close to the center of the image, where fixations generally are biased (Parkhurst et al., 2002; Tatler, 2007).

4.4 Summary

Earlier studies have shown that while free-viewing images people tend to gaze at regions with a high local density of bottom-up features such as contrast. In particular, this tendency was found to be more emphasized during the first few fixations after image onset. In this chapter, we used a new experimental paradigm to investigate how gaze locations are chosen; image contrast was modified and we measured how this affected eye-movement behavior during free viewing. Results showed that gaze density overall is shifted toward regions presented in high contrast over those reduced in contrast. However, initial saccade targets are largely unaffected by a change in contrast and certain image regions seem to attract early fixations regardless of display contrast. These results suggest that cognitive factors, instead of image features, are dominant in guiding eye-movements early after image onset.

Chapter 5

Effects of Contrast Manipulations and Image Semantics on Fixation Behavior

SINCE the contrast manipulation paradigm proved to be an interesting and efficient experimental method to study gaze control in images, we continue in this chapter to pursue the causes behind gaze control using this paradigm. We make a number of important modifications and extensions compared to the previous chapter. First, it is investigated how *image semantics* influence the relative contribution of lower-, and higher level cognitive mechanisms to viewing behavior. We propose a method to quantify image semantics dubbed *semantic information dispersion (SID)*. Second, Gaussian pyramids, instead of wavelets, are used to implement contrast manipulations because they yield smoother contrast reductions and fewer undesired contrast artifacts. Third, we analyze event based measures in the form of fixations instead of solely observing peoples' gazing behavior. Fourth, since edge density and contrast arguably are the two most investigated low-level features in earlier works, both of these features are analyzed in this chapter. Fifth, the level of discrimination for contrast and edge density between fixated and control locations is analyzed using receiver operating characteristics (ROC), which lately has arisen as a popular method in such analyses. Moreover, we use a slightly modified viewing instruction to alleviate the undesired top-down adoptions reported by participants in the previous chapter. Finally, due to the importance faces have in human communication and interaction, one section is devoted to the effect contrast manipulations have on face perception.

In this chapter, we will investigate how contrast and edge density contribute to fixation selection, and how this effect varies over time. Unlike the majority of previous studies, test images are contrast manipulated prior to display. Meanwhile, we aim to keep their semantic content intact. We believe that by decoupling objects (or regions) from their low-level signal strength, an analysis is more likely to elicit causal relationships between where subjects fixate and the reason they choose to look there. Besides manipulating the image statistics, three image categories are used: Images naturally embedding faces, images with man-made objects, and images depicting scenes with neutral semantics (trees, leaves, etc.). Each class is chosen to represent images with different *semantic information dispersion (SID)*, a concept we define as follows:

Definition 1 *Semantic information dispersion (SID) measures how scattered the information is that subjectively best conveys the information of the whole image.*

For example, a face generally contributes more to the core meaning of an image then does a leaf on a tree. Consequently, an image has a low SID if a small aspect of the image (such as a face) is judged to contain the majority of conveyed information. The rationale for using different image categories is to introduce a varying top-down influence without using an explicit task, a strategy employed by a range of earlier works. For example, the task *look at regions with uniform texture* would yield a low correlation between edge density and fixated image content, but would hardly reveal much about the mechanisms behind gaze guidance. To verify that the images chosen for the experiment indeed represent different levels of SID, an experiment is performed where subjects are asked to identify a fixed size region that best conveys the information of the whole image. The average overlap between the regions chosen by the subjects is then used to estimate the SID.

The remainder of this chapter is organized as follows: Section 5.1 describes the materials and methodology of the eye-tracking and data recordings. Specifically, we describe the images and how they are experimentally modified, the experimental setup, and the procedure for data collections. Results are given in Section 5.2 and discussed in Section 5.3.

5.1 Methods

5.1.1 Test images

Three semantic image categories are used. In the first category, we use images containing faces; it is known that faces are very semantically important image regions and therefore frequent fixation targets (e.g. Yarbus, 1967). The second category comprises images with neutral scene semantics and depicts scenes with motives from nature such as trees and bushes (from Einhäuser & König, 2003), grass, and a picture of a brick wall. The last category falls between the first two categories and contains man-made objects embedded in natural environments. Six images from each category are used. Images were converted to eight bit gray scale and resized to dimension 1024 × 768 through the Matlab functions `rgb2gray` and `imresize` (bilinear), respectively. The test images are shown in Figure 5.1. They comprise: Face images (top two rows), images with neutral semantics (row three and four), and images containing man-made objects (bottom two rows). As can be seen, each image comes in two versions where contrast has been modified differently.

Face images are modified to form two subcategories. In the first subcategory faces were retained in high contrast, whereas other regions were gracefully reduced in contrast away from the facial region. In the second subcategory, these contrast modifications were inverted; only the facial regions were reduced in contrast. Figure 5.2 exemplifies this. For the other two categories, each image was transformed into two different versions as follows: Four candidate positions, same for all images, were available as shown in Figure 5.3. One of these positions was selected at random, and the first version was generated by reducing the contrast smoothly away from this position. The other version was generated in a similar manner, but now with the contrast being reduced away from the point diagonally opposite to the randomly selected position.

5.1.2 Image manipulation

Contrast manipulation was implemented by means of variable resolution image processing using Gaussian pyramids. A five level pyramid was created by iterative lowpass filtering and downsampling of the original image, followed by upsampling and (bi-linear) interpolation back to the original image resolution (1024 × 768). Lowpass filtering was implemented by an ideal filter with a cutoff frequency adjusted to avoid aliasing given a subsampling factor of two pixels. These operations resulted in a collection of images where the original image comprised the bottom layer and higher layers were copies of the original image with increasingly lower contrasts. To create images with variable contrast, high resolution regions were selected from the bottom layer of the pyramid, whereas low resolution regions originated from the higher layers in the pyramid. Regions from different levels were then synthesized through a Gaussian shaped blending function. Let $I_\ell(m, n)$ denote an image at level ℓ in the lowpass pyramid. m and n span the image dimensions and $\ell = \{1, 2, 3, 4, 5\}$, where $\ell = 1$ denotes the bottom layer comprising the original image. Then the implementation can be described by Algorithm 1. $I(m, n)$ is the output

Figure 5.1: Test images.

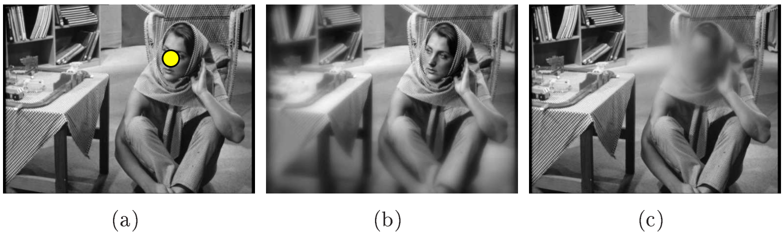

(a) (b) (c)

Figure 5.2: Contrast manipulation for face images. (a) shows the original image. In (b), the contrast is decreased away from the marker in (a), positioned over the woman's face. The figure in (c) illustrates the case where contrast instead is reduced toward the face area by inverting the contrast manipulation function in (b).

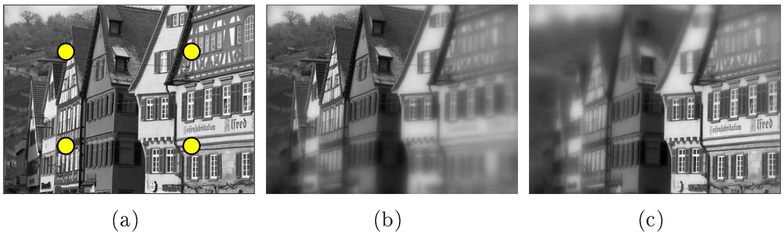

(a) (b) (c)

Figure 5.3: Contrast manipulation for images not containing faces. Figure (a) shows the original image with four candidate markers. One of these markers is chosen at random, and (b) illustrates the case when contrast is reduced away from this marker (in upper left corner). In Figure (c), the marker diagonally opposite the randomly picked one is instead used as the point from were contrast is reduced.

image, and $G(m, n)$ denotes a Gaussian function

$$G(m,n) = e^{-\left(\frac{(m-m_i)^2}{2\sigma_m^2}+\frac{(n-n_i)^2}{2\sigma_n^2}\right)} \tag{5.1}$$

where (m_i, n_i) represents the point where the Gaussian function is centered, i.e., the point from where the image is increasingly reduced in contrast. The $\hat{}$ operator denotes normalization to unit height. To introduce a noticeable amount of blur, σ_m and σ_n were set to 1024/2 and 768/2 pixels, respectively. These parameters were chosen simply by pilot testing where contrast reduction was deemed as significant without changing the semantics of the image. It has been pointed out in an earlier study (Parkhurst & Niebur, 2004), that when using contrast manipulations to study fixation selection, it is important to implement smooth contrast degradations to avoid undesired variation in higher order image statistics, which could explain possible changes in fixation behavior. Our implementation accounts

Algorithm 1 Implementing a variable contrast

```
$I(m,n) = I_1(m,n)$ {Initialize}
for $\ell = 2$ to 5 do
    $I(m,n) \leftarrow I(m,n) \cdot \hat{G}(m,n) + I_\ell(m,n) \cdot (1 - \hat{G}(m,n))$
end for
```

for this observation.

Contrast manipulation for face images was implemented with the above parameters when contrast was reduced away from the face. However, in the opposite case, when contrast was reduced toward the face region (the face was blurred), then the blending function was modified as

$$G_{inv}(m,n) = 1 - G(m,n), \{\sigma_m, \sigma_n\} = \{1024/2^3, 768/2^3\} \tag{5.2}$$

in order to better limit the contrast reduction effect to the facial region.

5.1.3 Subjects

13 naive test subjects (25.7±4.9 (M±SD) years old, one female) were recruited to participate in the experiment. Their visions were normal or corrected to normal. Compensation was given in the form a lottery ticket and subjects consented to use of their data by signing a form.

5.1.4 Experiment I: Viewing contrast manipulated images

Contrast manipulated images from all three categories were shown one at the time in full screen. Before the presentation of an image, a central dynamic fixation marker in the form of solid black circle was shown on a mid-gray screen. The diameter of the circle was decreasing as a function of time. After one second, the circle disappeared and an image was displayed in full screen during a time randomly drawn from the interval $t = [3, 4, 5, 6]$ seconds. This procedure was repeated for all images, which were shown in random order. Varying display time was used to prevent subjects from adopting top-down strategies such as systematic scanning of the images. Prior to each image was displayed, subjects were asked to look at the fixation marker.

The instruction given to the subjects was to *please study the images carefully.* Supposedly, being a fairly general instruction, it prevents subjects from adopting individual viewing strategies that try to guess the purpose of the tests. For example, we saw in an earlier study (Nyström & Holmqvist, 2007b), where subjects were given the more neutral instruction solely to *watch the images*, that subjects adopted a top-down strategy avoiding to look at the blurred regions a bit into the presentation. We believe that the task instruction used in this chapter will alleviate this undesirable adaption.

5.1.5 Experiment II: Image semantics evaluation

In a second experiment, that followed right after the first, subjects were shown the 18 unprocessed (no contrast manipulation) images (in eight bit gray scale of dimension 1024×768), one by one in full screen. For each image, their task was to position a box, controlled by the mouse cursor, over a region in the image that had the highest semantic importance. The exact instruction was given in writing before the experiment started: 'Position the box over a region that best conveys the information of the whole image'. There was no time constraint to finish this task, and when the final box position was decided, a mouse click ended the semantic rating to proceed to the next image. The size of the box was chosen large enough to encapsulate whole objects or parts of objects, so that the meaning of the box content would be clear without access to the whole image. We used a box size that spanned four degrees (128×128 pixels). Subjects were not informed about Experiment II until after the first experiment was completed.

5.1.6 Eye-tracking

Eye-tracking was preformed monocularly during both experiments with an SMI iView X Hi-Speed 1250 Hz system. Subjects were seated 0.67 m away from a 19 Inch Samsung GH19PS screen with the resolution and update rate set to 1024×768 pixels and 60 Hz. The physical dimension of the screen was 380×300 mm, spanning 32×25 degrees of visual angle. Each recording started with a 13-point calibration. Stimuli presentation, communication with the eye-tracker, and data analysis were performed with Matlab and its Psychophysics Toolbox Version 3 extension (Brainard, 1997). A saccade based detection scheme developed by SMI (IDFconvert.exe) was used to filter out event based measures such as fixations and saccades. Gaze positions were classified as saccades if the eye velocity was $\geq 75°/s$ and if the saccade duration lasted ≥ 10 ms. If these assumptions were violated, and the eye was stable for ≥ 50 ms, a fixation was detected.

5.2 Analysis and Results

The analyses address the following questions: Are contrast and edge density different at fixated regions compared to control regions for contrast manipulated images? Do contrast manipulations change where people look, and how is the magnitude of change related to image semantics?

If contrast manipulations change where people look, do they also change what people look at? Finally, we target how one category of images, namely those containing faces, is affected by contrast manipulations.

5.2.1 What do we look at? – Feature analysis

It is known from several previous studies that certain low-level features are elevated at fixated positions. For example, fixated locations tend to have higher contrast and edge density than non-fixated, control regions. We begin our analysis by testing whether these observations still hold using contrast manipulated images. Contrast at the image location (m, n) is defined as the standard deviation within a 3×3 neighborhood centered at (m, n). Edge density is extracted by convolving the image separately with horizontal and vertical Sobel operators, and then computing the average of these filtered outputs.

In the analysis, an approximately 1 degree (32×32 pixel) region is extracted from the feature maps around each fixation location. For comparison, 1 degree regions are also extracted from control locations, and the difference between fixated and control feature contents is analyzed. Instead of using uniform sampling over the image area to simulate a random viewer, we use control fixations collected from other images used in the experiment. This way, a simulated 'random' fixation pattern coincides with the distribution of fixations, which is known to be non-uniform with a bias to the center of the display. It has been argued that the central bias may give rise to artificially high features values at fixation (e.g., Tatler et al., 2005), and should therefore be carefully accounted for in the analysis.

An increasingly popular method to estimate the degree to which fixated and control feature content can be differentiated from each other is the receiver operating characteristics (ROC) analysis (e.g., Hanley & McNeil, 1982). A ROC curve plots the fraction of *true positives* (TPs) against the fraction of *false positives* (FPs). In our case, TPs consist of fixated feature content, whereas FPs comprise feature content at control locations. The area under the ROC curve varies between zero and one, and is a robust measure of how well image features can be discriminated between fixated and control locations; if the ROC area is significantly larger than 0.5, a tested feature is said to discriminate fixated locations from control locations. A ROC area that equals 1 is said to give perfect classification.

Figure 5.4 plots the average ROC areas for contrast and edge density. Black bars represent results considering the first fixation (from all subjects in all images) only, whereas the white bars represent a similar analysis over all fixations. By the *first fixation*, we mean the fixation following the initial saccade after image onset and not the first registered fixation is the data file, which is constrained to the center of the screen by

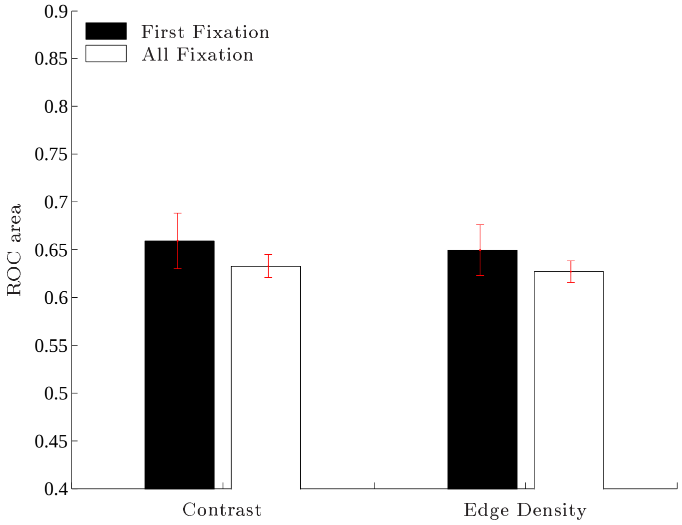


Figure 5.4: ROC areas for discrimination between image features at fixated and control locations. Black bars show ROC areas for the first fixation whereas all fixations are included in the white bars. Error bars span standard errors of the mean. A ROC area larger than 0.5 indicates a difference.

a fixation marker. As reported by several previous studies, feature densities at fixated locations are significantly higher (ROC area > 0.5) than feature densities at control locations ($p < 0.01$, t-test, for both contrast and edge density). Apparently, this is also true for contrast manipulated images. Moreover, there is a tendency, although non-significant, that initial fixations discriminate contrast and edge density better than fixations do over the whole time course of viewing.

5.2.2 Do image semantics and feature manipulations influence where we look?

To this point, our empirical findings are in line with previous results emphasizing bottom-up control over fixation selection. The findings show, *on average*, that contrast and edge density are higher at fixated positions than at other, control positions. In this section, it is investigated whether these general tendencies are consistent when analyzing images with regard to their semantic information dispersion (SID) as well as their direction of contrast reduction. What happens with peoples' allocation of fixations, for example, if a region deemed as semantically important is reduced in low-level signal strength? Obviously, a saliency based framework would predict an obligatory shift in fixation density away from this region.

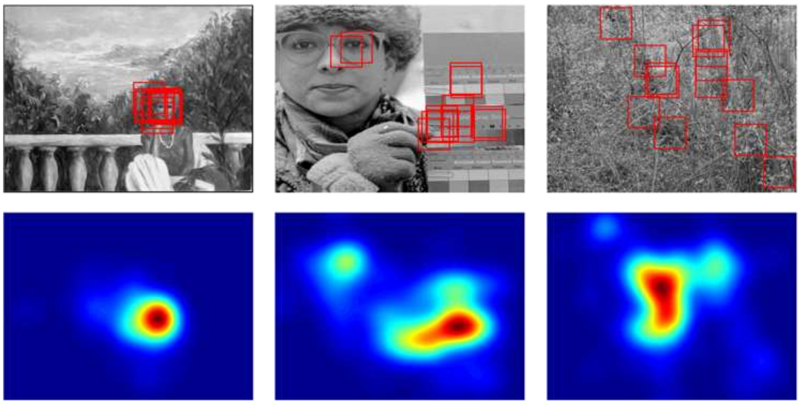


Figure 5.5: Images in order of increasing semantic information dispersion (SID). The top row shows where subjects have positioned a box that 'best conveys the information of the whole image'. The bottom row illustrates the fixation density of the same subjects while performing this task. As can be seen, the inter-subject agreement between fixation density and the regions judged to best convey the information of the whole image is large.

Using data collected from the second experiment, we found the SID for each image, calculated as the average overlap between box locations within an image. Thus, if $B_{i,j}$ denotes a box in the image i positioned by subject j, the SID for image number i is defined as

$$\mathrm{SID}_i = \left[\frac{2}{P(P+1) - 2P} \sum_{\substack{j=1,\ldots,P-1 \\ k=j+1,\ldots,P-1}} B_{i,j} \cap B_{i,k} \right]^{-1} \tag{5.3}$$

where $\cap$ denotes the intersection between the boxes in pixels, and P is the number of viewers. The inverse is computed such that a large SID value represents a spread out semantic information and vice versa. The top row in Figure 5.5 shows three of the unprocessed test images and the boxes as positioned by the test subjects. Out of the 18 unprocessed images used in the experiment, images with the lowest, midmost, and highest SID are shown in the figure. Unsurprisingly, the image with the lowest SID contains a face, and the image with the highest SID contains rather neutral semantics. For the sake of comparison, the fixation density of the same subjects performing the SID detection task is given in the second row in the figure. For these images, the overlap between where subjects fixated and where they positioned the box is quite large. As expected, the image categories were tightly couple with SID; five of the six images containing faces were among the images with the lowest SID (boxes were dominantly positioned over the face), and all the six images from the

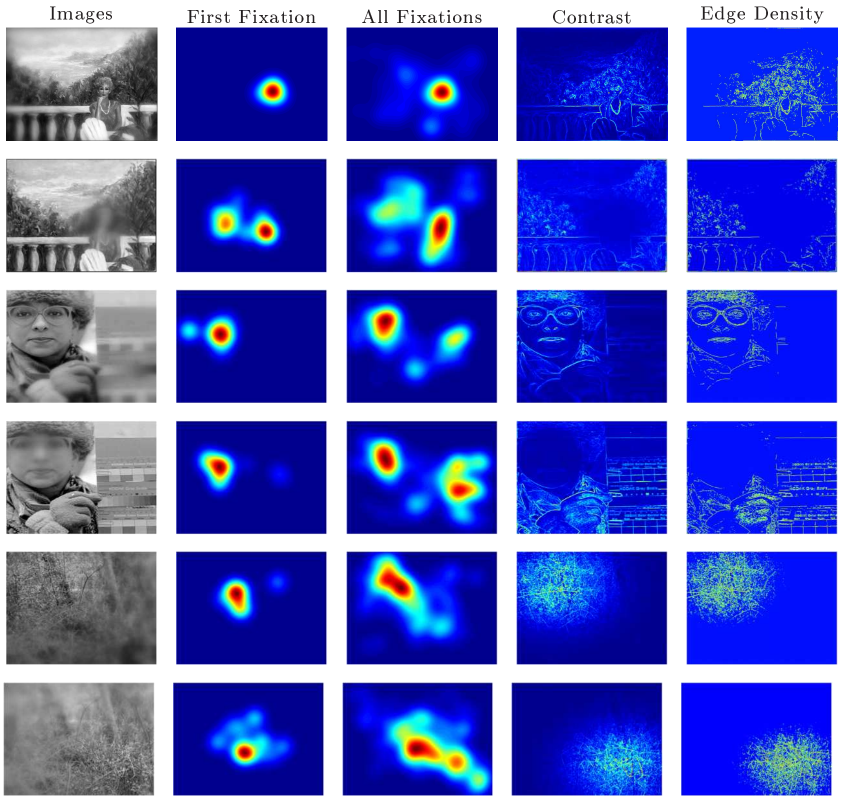


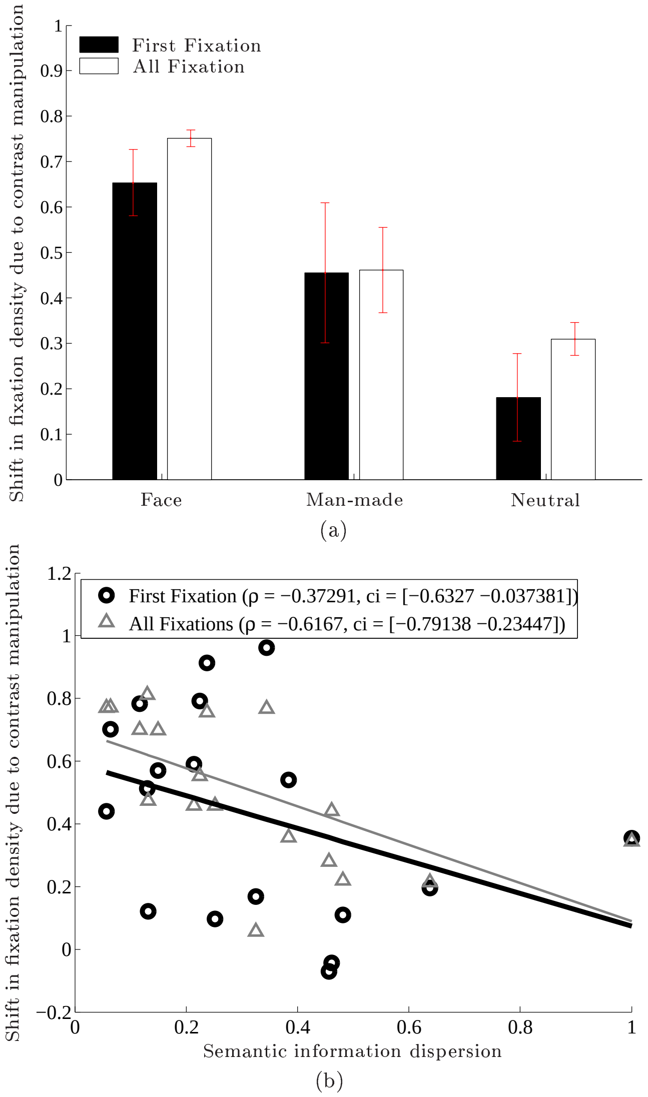
Figure 5.6: Effect of contrast manipulation on fixation behavior.

'neutral' category had the highest SIDs. Consequently, five images from the 'man-made object' class were located in the mid-SID section along with one face image.

Figure 5.6 illustrates how the fixation density changes as a result of contrast manipulations for images with low, medium, and high SID. The fixation densities are visualized as heat maps, where Gaussian functions have been centered at each fixation location and then superimposed. The variance of each Gaussian function has been set such that the width at half its maximum height approximates the size of the foveal span of a viewer in the current experimental setup. In addition, the height of each Gaussian function has been scaled in proportion to the fixation duration. As a consequence the fixation densities not only reflect where people have fixated, but also their level of cognitive processing during each fixation, hence providing more sensitive and detailed information. Henceforth, we refer to the heat maps as fixation density functions (FDFs), in order to better capture what the heat maps represent. The second column in Figure 5.6 depicts FDFs for all subjects during the first fixation, and the third column illustrates corresponding fixation densities collapsed over all fixations. This can be compared with the two rightmost columns, where contrast and edge density are visualized. An inspection of the plots indicates that contrast and edge manipulations clearly influence where subjects look. However, the magnitude of change seems to differ depending on the image type; the images containing faces undergo relatively small changes in fixation placement due to contrast manipulation whereas fixations in the images that contain more neutral semantics seem to be more influenced by the manipulations.

To quantify how fixation locations change as a function of contrast manipulation and SID, the two-dimensional correlation coefficient between FDFs belonging to the two contrast manipulated versions of each image is computed. This metric has been used in other works for the same purpose (Rajashekar, Linde, Bovik, & Cormack, 2008). Although it is not clear how accurately the 2-D correlation coefficient, or any other metric for that matter, captures the difference between people's fixation locations, it gives an estimate that helps us to interpret the magnitude of change. For a reference of other metrics used to estimate the similarity between fixations, see for example Mannan, Ruddock, and Wooding (1995); Privitera and Stark (2000); Tatler et al. (2005). Since images' SID-values almost perfectly matched the initial division of images into three semantic categories, the analysis is preformed with respect to the image categories, which henceforth are referred to as 'Face', 'Man-made', and 'Neutral'. Figure 5.7(a) depicts the average 2-D correlation between FDFs generated from the initial fixation (black bars) and all fixation (white bars) within each category. It can be seen that the image category influences the degree to which contrast manipulations trigger shifts in fixation densities; images containing regions of high semantic importance, such as faces, are less sensitive to the manipulations than other images and in particular those from the 'Neutral' category. This tendency is present for both the initial fixation and for fixations over the time course of viewing.

Another way to represent how fixation locations are affected by contrast manipulations and semantics, shown in Figure 5.7(b), is to plot the shift in fixation density (2-D correlation coefficient between FDFs) against images' SID. Circles and triangles represent how the initial fixation and all fixations, respectively, are shifted in location as a function of SID. The lines are least square fits to the data points. Considering all fixations, it can been seen that SID clearly influences the magnitude of shift in fixation density, having a correlation of $\rho = -0.62$. This tendency is weak, or hardly present at all, considering the first fixation only. It may be the case since fewer fixations are used to generate the first fixation FDFs, giving individual fixations more weight. Consequently, a fixation that is not aligned with other fixations has a large impact on the shape



(a)

(b)

Figure 5.7: Influence of fixation selection on image category, SID, and contrast manipulations. (a) Bars represent the average shift in fixation density due to contrast manipulations within each image category. Error bars span one standard error around the mean. (b) The solid lines are least square fits to the data points. 95% confidence intervals of the correlation coefficient ρ are generated using bootstrapping with 1000 resampled sets (Matlab's `bootstrp` function).

of an FDF, and therefore also the value of the 2-D correlation coefficient between two FDFs. In summary, the results from Figure 5.7 clearly illustrate that the degree to which fixation locations are influenced by contrast manipulations depends on SID and image category.

5.2.3 Do image semantics and feature manipulations influence what we look at?

Since contrast manipulations change where people look with different magnitudes depending on images' SID, one would expect this to be reflected in fixated image content across the image categories. For example, in the category that was least influenced by the image manipulations, we would expect a lower discrimination for contrast and edge density between fixated and control locations than for the other two categories. Figure 5.8(a) plots average ROC areas for contrast and edge density over the three image categories. Results for both the first fixation and all fixations are given for each feature and category. As expected, the discrimination of features between fixated and control locations was the lowest in the 'Face' category and increasingly higher for the 'Man-made' and 'Neutral' categories. However, it was significantly ($p < 0.05$, *t*-test) better than chance (ROC area > 0.5) in all cases. Also notice how ROC scores in the 'Neutral' category are significantly ($p < 0.05$) higher for first fixation than all fixations, whereas this tendency was not significant in the other two categories. Figure 5.8(b) differs from Figure 5.8(a) in that only images from the 'Face' category where contrast was reduced toward the face, i.e., where the faces were blurred, were included in the analysis. Since people still looked at the face regions after being reduced in contrast, the discrimination was reduced to a chance level, considering both the first and all fixations. Interestingly, discrimination was worse for feature content fixated at the initial fixation, contrary to the finding by Parkhurst et al. (2002).

Both image semantics and features determine what we look at. There is a clear effect, however, that semantically important regions are looked at largely independent of their feature content in terms of contrast and edge density.

5.2.4 What so special about faces?

In agreement with previous findings faces seem to attract viewers' gazes, and do so largely regardless of their contrasts. So, what is so special about faces, and what can we learn about face perception using the contrast manipulation paradigm developed in this thesis? Considering all face images regardless of contrast, and if facial regions are defined by the black boxes in Figure 5.9, we found initial fixations to be located within these

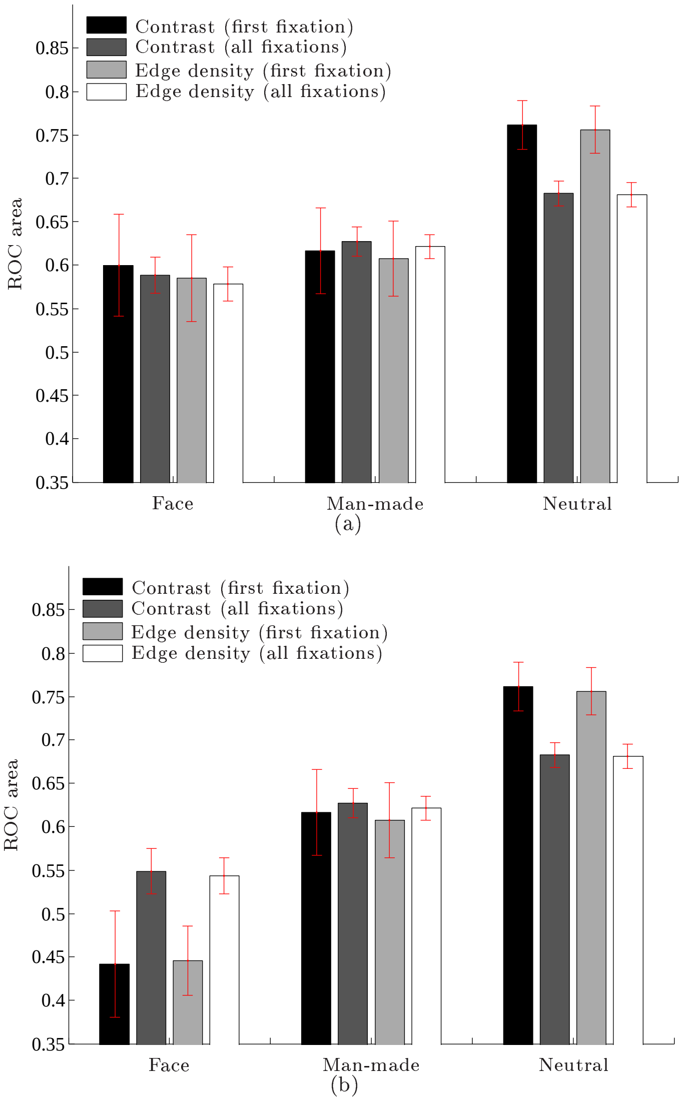


Figure 5.8: ROC analysis of contrast and edge density over different image categories. (a) All images from each category are included. (b) From the 'Face' category, only images with blurred faces are included.

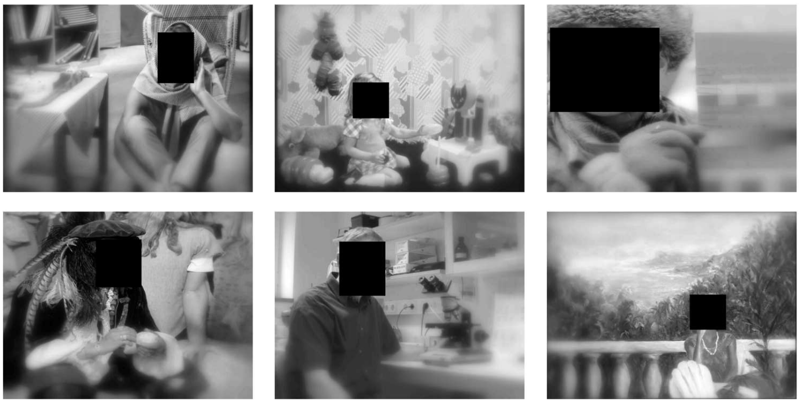
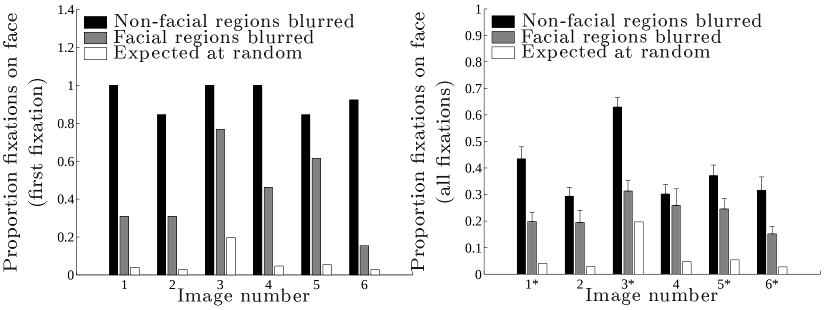


Figure 5.9: The images showed here depict the versions where contrast has been reduced away from the face regions, i.e., in 'non-facial' regions. The black boxes define the face regions used in the analysis

regions in 68.6% of the trials, and in 30.9% when taking all fixations into account. In the case non-facial regions are reduced in contrast, faces are fixated initially 93.6% of the times and overall in 39.1% of the trials. When the faces instead are reduced in contrast these numbers decrease to 43.6% and 22.7% , respectively. Faces are expected to be fixated with 6.5% chance if fixation locations are drawn from a uniform distribution.

Figure 5.10 breaks down the analysis to an image by image basis; Figure 5.10(a) plots the proportion of initial fixations located on the face region, and Figure 5.10(b) contains similar plots taking all fixations on the face into account. The x-axis lists the images in Figure 5.9 numbered from left to right starting from the upper left corner. It can be seen that subjects' initial fixations are overrepresented in face regions in all the tested images, and that fewer fixations are located on a face when its contrast is reduced. The same trend is found when considering the proportions of all fixations on the face regions. However, in this case many fixations are located on non-facial regions. In particular, this is true in images where other semantically important regions compete for attention with the faces; in the image numbered '2', there are toy animals whose faces attract many fixations and in image '4' the hands of the man are a strong competitor to the face region.

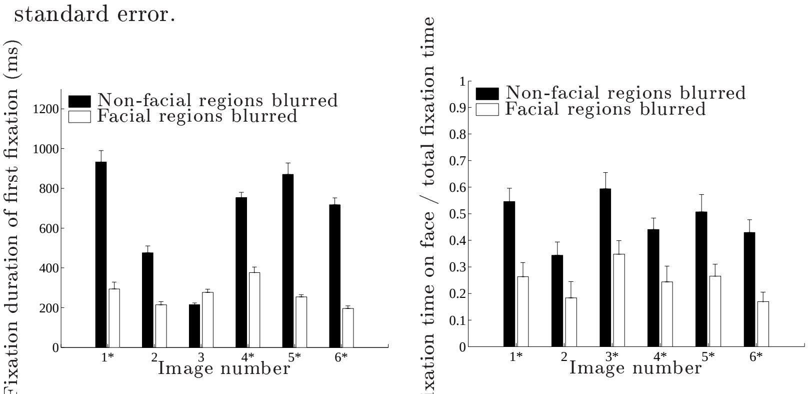
Besides knowing the position of a fixation, the fixation duration is another important measure that reflects ongoing visual and cognitive processes (Rayner, 1998; Henderson & Ferreira, 2004). Initial fixation durations are given in Figure 5.11(a), whereas all fixations located on the face are plotted in Figure 5.11(b). Apparently, if the initial saccade lands on the face, the duration of the following fixation is longer when the contrast



(a) Proportions of first fixation on the face

(b) Proportions of fixations on the face. A (*) indicates a significant difference in mean for significance level $\alpha = 0.05$.

Figure 5.10: Proportion of fixations located on faces. Error bars span one standard error.

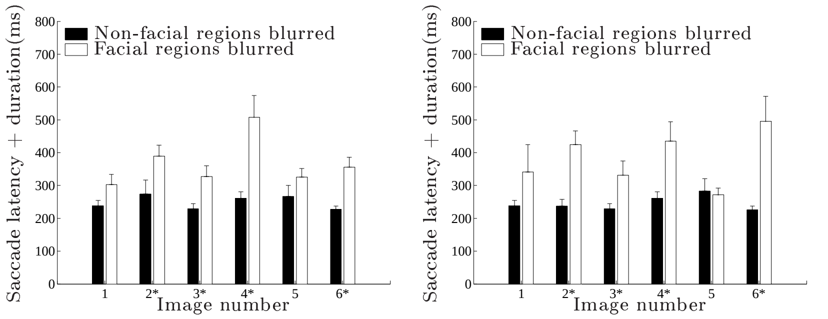


(a) Fixation duration of first fixation

(b) Overall fixated time on face

Figure 5.11: (*) indicates a significant difference in mean for significance level $\alpha = 0.05$ (t-test).

of the face is higher than its surrounding regions. Figure 5.11(b) gives the total fixation time on the face as a proportion of the total viewing time. Again, it can be seen that the faces are looked upon more when they are kept in high contrast. Since fixation durations not only depend on foveal information available to the viewer but also on peripheral information, we analyze initial saccade latencies, in the case a saccade is directed toward a face region. Latencies are measured as the time from image onset until the first saccade lands in a fixation. Thus, included in the saccade latencies is the time it takes to execute the saccade, which typically is 50 ms. Figure 5.12 shows the initial latencies when all initial saccades, regardless of final destination, are considered (Figure 5.12(a)), and when only those landing on the face are considered (Figure 5.12(b)). The figures tell us



(a) Saccade latency for all initial saccades

(b) Saccade latency when initial saccade is directed toward the face

Figure 5.12: Initial saccade latencies + durations for all saccades (a), and only for the saccades directed toward the face region. (*) indicates a significant difference in mean for significance level $\alpha = 0.05$ (t-test).

that a reduced face contrast yields an increase in saccade latency.

5.3 Summary

This chapter extends the work from the previous chapter using the experimental paradigm based on contrast manipulations; it investigates the contribution of the low-level features contrast and edge density as well as image semantics to the selection of fixations in images. Overall, both contrast and edge density were elevated at fixated image patches compared to control patches within an image. However, image content actively chosen by subjects' gazes varied significantly with a number of factors. First, when regions of high semantic importance were reduced in contrast, subjects still looked at these regions, causing contrast and edge density to be *lower* at fixated locations compared to control locations. This tendency was particularly strong when faces were reduced in contrast, and was found both early after image onset as well as later in viewing. Second, image content at fixation proved to correlate better with the tested low-level features when the semantic information dispersion (SID) was high. In other words, when an image does not contain any specific regions of high semantic importance, bottom-up features correlate quite well with image content around fixation locations. Overall, the results in this chapter do not support a causal link between bottom-up features and image content at fixation.

Chapter 6

Assessing Fixation Prediction Algorithms on Contrast Manipulated Images

ALGORITHMS for fixation prediction have recently attracted considerable attention from researchers across different fields. One reason for this interest is the potential benefit such algorithms would have in a range of research disciplines and future technical systems. Accurate algorithms for gaze prediction could replace time consuming eye-tracking experiments and hence, for example, be used to automatically assess if people look at the desired product in a commercial, or provide relevant visual input to a robot. The ability of some of the proposed algorithms to predict human fixations has been reported to be quite good under certain conditions (Itti & Koch, 2000; Parkhurst & Niebur, 2002; Itti, 2004), despite using only low-level features as a basis for prediction. Given the current, intense debate on gaze control and fixation prediction in natural images, we will in this chapter take a closer look at two algorithms that predict human fixations solely based on low-level image input: One, by Itti et al. (1998), is based on the concept of a *saliency map* and is well established and evaluated against human fixations in several previous works (see e.g., Parkhurst & Niebur, 2002). The other algorithm is a very recent contribution by Rajashekar et al. (2008).

Previous chapters did not support the hypothesis that low-level features per se provide causal cues to fixation selection in natural images. Instead, regions with a high semantic importance, such as a face, could rather easily cognitively override manipulations in image contrast.

To put the predictive accuracy of the two algorithms to a test, they are used to find fixations in some of the contrast manipulated images we used in last chapter. The similarity between algorithmically generated fixations and human fixations will be compared. Again, the main novelty in this chapter lies in, as opposed to the majority of previous work, using stimuli manipulations to naturally separate image semantics from its low-level signal strength. By applying the algorithms to the manipulated images, we will measure how they contribute to fixation selection under task-neutral viewing.

6.1 Predicting Fixations

We present in this section two different approaches to algorithmic prediction of fixations. It is not intended as a comprehensive description of the algorithms, but merely an overview of their major components and functionalities. For details, refer to the references given.

6.1.1 Saliency map approach

The concept of a *saliency map* and its relevance in attentional guidance was first proposed by Koch and Ullman (1985). According to a saliency map, visual importance is represented by a two-dimensional map predicting how likely each location of an image is to be visually attended by a viewer; peaks in the saliency map point to regions likely to be gazed at, and vice versa. By successively moving to the highest peak in the saliency map, a sequence of fixations can be predicted. To prevent the algorithm from halting at the largest peak, it is endowed with an inhibition-of-return mechanism, which reduces the saliency at previously visited peaks. The saliency at these regions is restored after a period of time such that the same image location can be visited multiple times over the course of viewing.

A saliency map is computed by first decomposing an image into a set of feature channels, typically comprising luminance, orientation, and color. Each feature channel is then transformed into a feature map by feeding it through center-surround extracting filters and a mechanism that allows spatial competition between neighboring feature content. Finally, all features maps are combined into a single saliency map. The choice of features are motivated by early psychological research, e.g., that by Treisman and Gelade (1980), suggesting that some features trigger attentional selection quickly and obligatorily by 'popping out' from their surrounds.

Algorithmic implementations following the framework outlined by Koch and Ullman have been proposed in several papers, e.g., (Itti et al., 1998; Itti & Koch, 2000; Walther & Koch, 2006). We have used the Matlab based Saliency Toolbox by (Walther & Koch, 2006) to compute the first 15 fixations in each tested image. Besides generating a saliency map from an image, this implementation identifies salient object-based representations in a image. We do not use this extension of the implementation, but use the saliency map directly to predict fixated locations.

Implementations of saliency maps have been validated against human fixations in some earlier papers (see e.g., Parkhurst et al., 2002; Henderson et al., 2007; Rothkopf et al., 2007; Foulsham & Underwood, 2008). There is some evidence that peaks in saliency coincide with fixation locations when the viewing task is neutral, but also ample evidence that task and context can override such a relation.

6.1.2 Gaze attentive fixation finding engine (GAFFE)

The gaze attentive fixation finding engine (GAFFE), which is designed by Rajashekar et al. (2008), uses an approach based on machine learning. Using an image set comprising gray scale, natural images, fixations from a large number of subjects are collected to find statistical differences between fixated and control image locations using a *foveated* image analysis. In a foveated analysis, an image is blurred away from the *current point* of fixation in accordance to the spatial sensitivity of the human visual system (HVS). Then a region around the location for *next* fixation is analyzed in terms of feature content. This way, a foveated analysis uses the information available to a human viewer at the time a saccade to the next fixation location is planned. Rajashekar et al. report that luminance and contrast as well as bandpass outputs of these features are significantly higher at locations fixated by human viewers compared to control locations. Consequently, these features are chosen as the basis for prediction.

Fixation prediction is initiated by foveating an image away from its center. This foveated image then is filtered with respect to the four features mentioned above, and the next fixation target is decided by combining the filtered feature maps based on parameters empirically found by the initial analysis. The algorithm proceeds by updating the foveation point to the next (predicted) fixation and repeats the filtering procedure at this new fixation. GAFFE permanently inhibits previously fixated positions from becoming fixated again. Also, it does not attempt to predict the temporal order of the fixations.

As for the saliency map approach, we use GAFFE to find 15 fixations (we do not use the central, initial fixation). Before applying GAFFE to predict fixations on our set of images, parameters were modified to fit the experimental setting we used while recording eye-movements.

6.1.3 Eye-tracking on human subjects

To validate the algorithmic predictions, fixations were extracted as described in the previous chapter, Section 5.1.6. To allow for a fair comparison between human and algorithmic fixations, the first 15 fixations (excluding the initial in the center of the display) from each viewer were selected to comprise the human baseline measure. In case fewer than 15 fixations were recorded from one viewer, these $N_f < 15$ fixations were used in the analysis.

6.1.4 Stimuli

Images belonging to the categories 'Face' and 'Neutral' from the last chapter were used. They were chosen since they represent images with different semantics; faces are known to convey much information in human interaction whereas images from the 'Neutral' category contain no objects of particular informative semantics. Stimuli are shown in Figure 5.1 (p. 42).

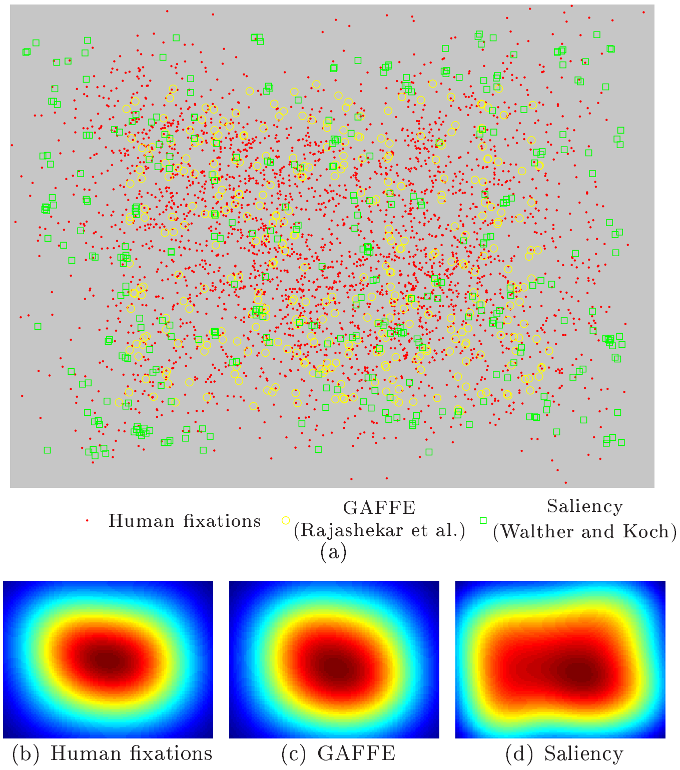
6.2 Analysis and Results

6.2.1 Qualitative analysis

Figure 6.1(a) illustrates how human and algorithmically predicted fixations from all images (and subjects) are distributed. Consistent with what has been reported in previous works, human fixations show a clear bias toward the center of the image as illustrated by the heat map in Figure 6.1(b). It can further be noted that human fixations tend to have an oval distribution, being extended more in the horizontal direction than in the vertical direction. GAFFE also shows a strong central tendency in fixation distribution (Figure 6.1(c)), but with more equally extended horizontal and vertical biases. Lastly, Figure 6.1(d) visualizes how fixations computed from saliency maps are distributed; substantially more fixations are located toward the edges in the images compared to the other two cases.

Figures 6.2 and 6.3 show a comparison between human and algorithmic fixations for images belonging to the 'Neutral' category. As described in previous chapter, contrast has been reduced in a Gaussian-like manner away from a (different) location in each version of an image. Dots represent human fixations from all tested subjects, squares point to locations predicted by a saliency map, and circles indicate fixations generated by GAFFE. As we reported from previous chapter, the distribution of fixations recorded from human viewers is shifted toward regions kept in high contrast. The general tendency for both algorithmic predictors is similar. Interestingly, GAFFE seems to overemphasize the bias toward regions of high contrast whereas the opposite is true for prediction made from saliency maps.

In Figures 6.4 and 6.5, algorithmic prediction is compared to human fixations on the images containing faces, which we from the previous chap-



(a)

(b) Human fixations (c) GAFFE (d) Saliency

Figure 6.1: (a) Distribution of fixations collected from human viewers and predicted algorithmically. (b)-(d) Fixation density functions represented as heat maps.

ter know attract human fixation regardless of the tested contrast. As expected, the limitations of purely bottom-up predictors are made explicit when reducing the contrast in the facial regions; the predictions deviate strongly from human fixation in these situations. Both algorithms fail systematically to predict that fixations will land on a blurred face. In fact, they are in most cases not even close to the faces. Interestingly, also when they are kept in high contrast, faces are sometimes missed by the algorithms.

6.2.2 Quantitative analysis

To quantify the strengths of the tested algorithms' abilities to predict human fixations, we use two different methods to estimate the similarity between two sets of fixations: The 2-D correlation coefficient and a

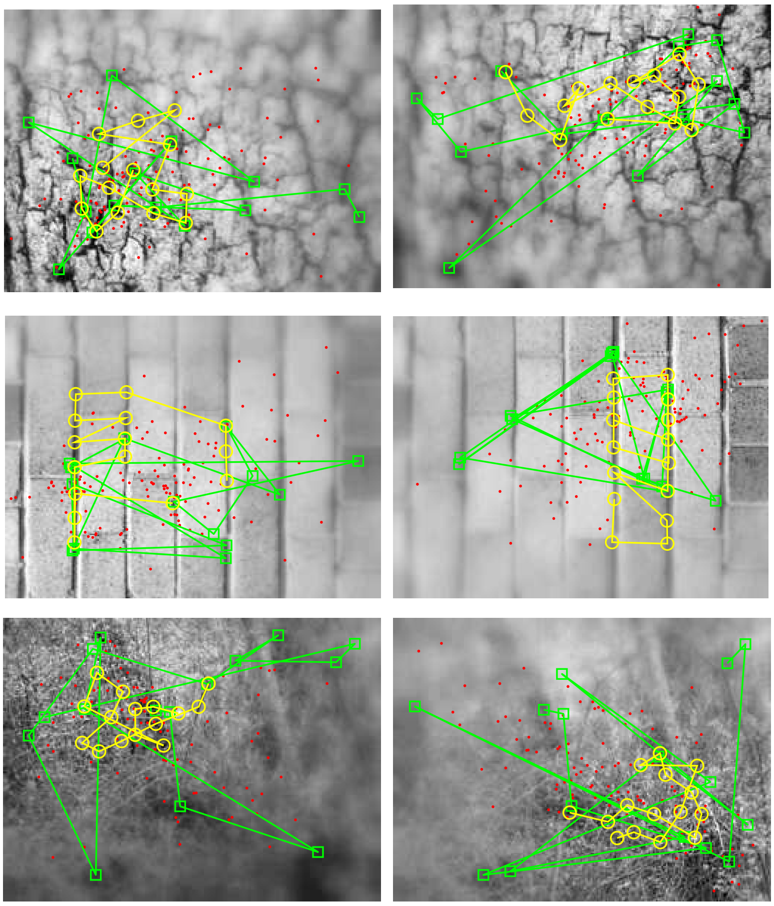


Figure 6.2: Algorithmically predicted fixations and human fixations (dots). Algorithmic predictions are made by GAFFE (circles) and by using a saliency map (squares).

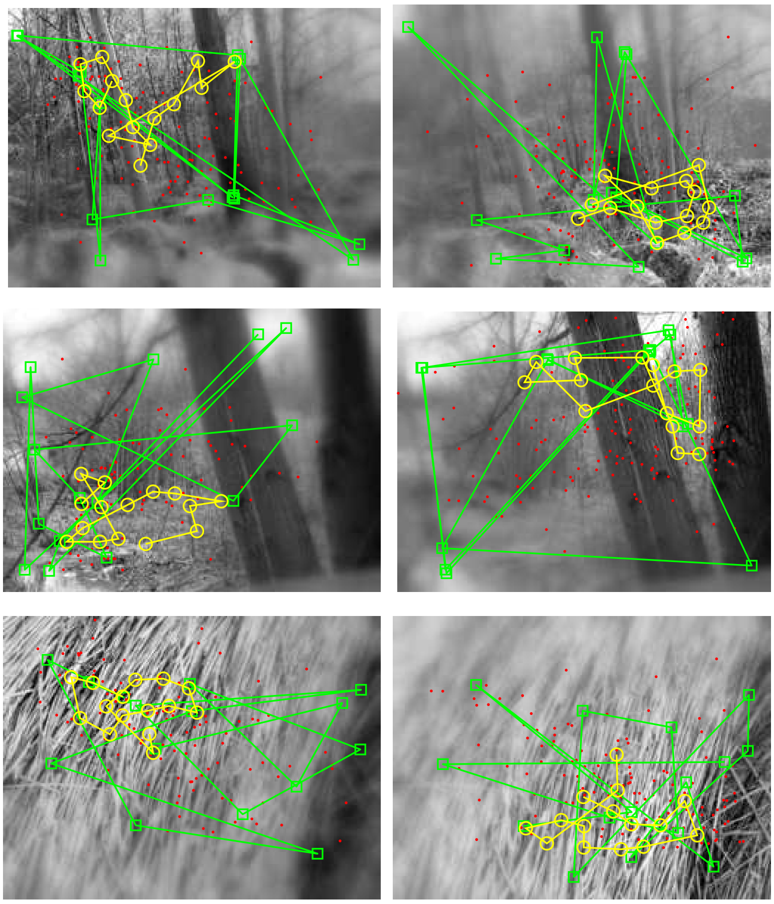


Figure 6.3: Algorithmically predicted fixations and human fixations (dots). Algorithmic predictions are made by GAFFE (circles) and by using a saliency map (squares).

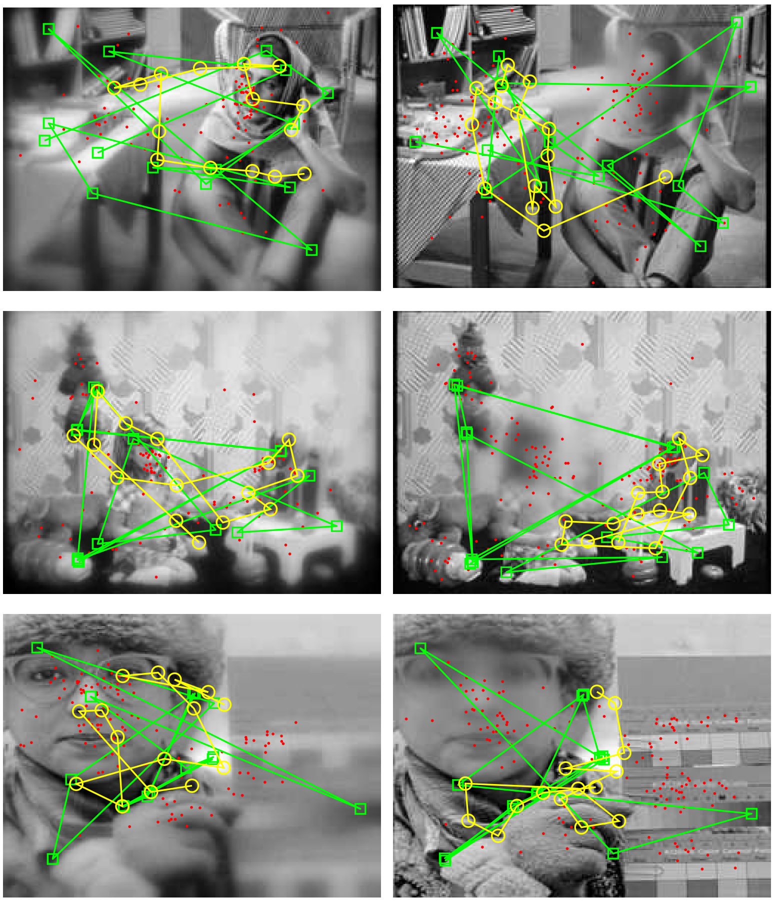


Figure 6.4: Algorithmically predicted fixations and human fixations (dots). Algorithmic predictions are made by GAFFE (circles) and by using a saliency map (squares).

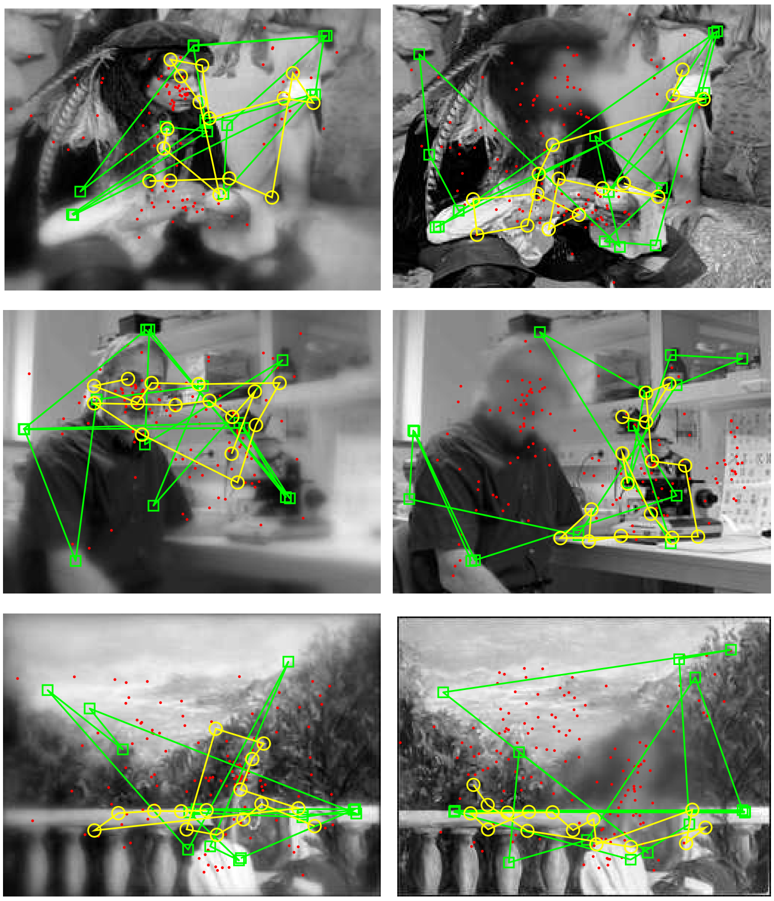


Figure 6.5: Algorithmically predicted fixations and human fixations (dots). Algorithmic predictions are made by GAFFE (circles) and by using a saliency map (squares).

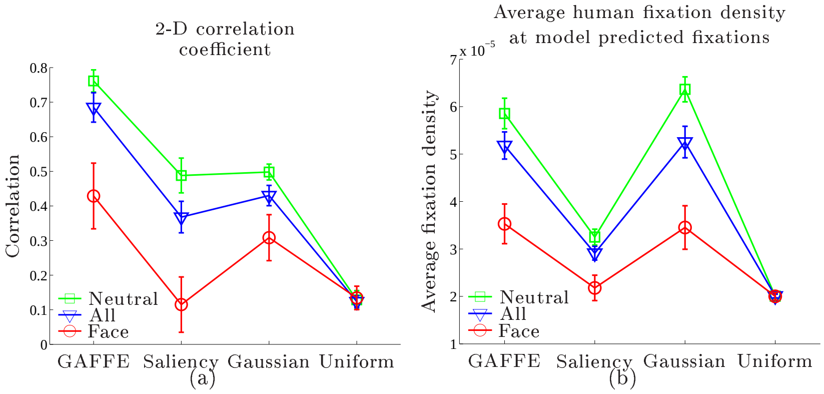


Figure 6.6: Comparison between human fixations and model generated fixations using two different methods. Error bars span one standard error.

dispersion measure that we defined in Chapter 4, Eq.(4.1). Initially, fixations collected from humans and predicted by GAFFE and saliency maps are used to create fixation density functions (FDFs) for each image. The FDFs were generated using $\sigma = 20$ pixels. Since neither of the algorithms attempts to predict the duration of a fixation, human FDFs are generated without taking fixation duration into account.

First, the correlation coefficient between human FDFs and algorithmic FDFs is computed. Figure 6.6(a) illustrates these correlations for all the images together, and those from the 'Neutral' and 'Face' categories separately. For comparison, FDFs for 15 fixations drawn from a uniform and a Gaussian (to model the central bias) distribution are compared against human fixations. Samples drawn from the latter distribution were generated by Matlab's `randn` function and then scaled by σ. In order to get more robust comparisons, uniform and Gaussian samples were compared to human fixations over 10 trials, and the average value over these comparisons was used.

Using the correlation coefficient to estimate the similarity between FDFs, it can be seen from Figure 6.6(a) that algorithmic prediction performs best on images coming from the 'Neutral' category and worst on images containing faces. These results are verified in Figure 6.6(b), where similar comparisons have been made using the dispersion measure. Remarkably, it seems like fixations generated from a Gaussian distribution, that is, fixations that are biased toward the center of the image, are compatible or outperform the algorithmic fixation predictors. Remember that this is the case despite that contrast manipulations explicitly are implemented 'off-center', i.e., the kept high-contrast regions in the manipulated images are deliberately positioned a bit away from the center of the image.

Overall, GAFFE seems to predict fixations better than a saliency map.

However, much of this effect derives from the central bias that the designers of GAFFE have built in. The bias originates from two sources. First, GAFFE always begins its prediction at the center of the image, and since the distance between the current and the next predicted fixation typically is quite small[1], it may take a while for the algorithm to reach the borders of the image. Second, a mask attenuating features along the borders is applied before prediction. This prevents fixations from appearing close to the image borders, as can be seen from Figure 6.1(a).

6.3 Summary

We evaluated the performance of two bottom-up driven algorithms for fixation prediction against human fixations recorded from viewers watching images with manipulated contrast. While previous work has shown that certain task instructions can override predictions made by bottom-up algorithms, we show that by using a more neutral task in combination with contrast manipulated images, the same effect can be elicited. In view of these observations, our results strongly question the causal contribution of bottom-up algorithms to fixation prediction.

[1] The correlation between features and fixated image content is significant only for short saccades, typically ≤ 8 degrees.

Chapter 7

Discussion of Part I

WE investigated gaze control in natural images using a new experimental paradigm where contrast manipulated images were inspected during task neutral viewing. A measure called semantic information dispersion (SID) was devised to estimate the compactness of an image's semantics and to classify images into semantic categories, and we quantified how both contrast manipulation and SID influenced where people looked. Finally, using images before and after their contrasts were manipulated, we compared two state-of-the-art algorithms for fixation prediction against fixations collected from participants. Over all subjects and images, we found a net effect that contrast manipulations changed where people looked; their gazes were repelled from regions where contrast had been reduced. Interestingly, we also found that the degree to which contrast manipulations affect participants' gazes depends on an image's semantic category; semantically informative regions attract visual attention despite being reduced in contrast. Fixations made on images containing faces, in particular, were rather insensitive to the manipulations, and participants looked at the face regions regardless of tested contrasts. In agreement with these results, our comparative study revealed that algorithms using bottom-up features to predict human fixations sometimes perform well, but many times fail miserably.

In Chapter 4, the effect contrast manipulation has on gaze locations was analyzed. Over all images and types of manipulations, we found that subjects' gaze positions were affected by contrast manipulations; gaze density was shifted toward regions in high contrast over those reduced in contrast. We also found that participants on average looked at regions with contrast higher than what was found at control regions. This is consistent with the hypothesis of preattentive selection, i.e., that attention is drawn to local image cues based on their physical signal strength.

The bulk of previous works emphasize the contribution of such low-level features to gaze guidance. For example, it has been shown that contrast (Mannan et al., 1996; Reinagel & Zador, 1999; Tatler et al., 2005), edge density (Baddely & Tatler, 2006), and saliency (Parkhurst et al., 2002) are higher at fixated than control regions. A saliency based framework, in particular, predicts an obligatory shift in fixation density toward regions where the low-level signal strength is high (Koch & Ullman, 1985).

By analyzing images from different semantic categories, we found in Chapter 5 that the degree to which contrast manipulations affect fixation selection heavily depends on the semantic content of an image, as well as how this content is distributed over the image area. In our experiments, face regions attracted attention regardless of their tested contrasts, whereas fixations in images with more neutral semantics, such as a photograph of a brick wall or a forest, were shifted toward regions where the contrast remained high. For the semantic category comprising photographs of man-made objects, we observed a moderate change in where people looked; gaze locations were affected more than in the face images but less than for images containing neutral semantics. These results suggest a semantic override of low-level features, in case the semantic information dispersion (SID) is low and points to regions with high semantic relevance such as a face. In images with high SID, on the other hand, contrast manipulations seem to dominantly influence where people look. However, even though the correlation between bottom-up features and fixated image content is higher in the latter case, it cannot be ruled out that other high-level mechanisms still control fixation selection. It is possible, for example, that the contrast manipulations affect images' semantic content, which then is responsible for shifts in fixation density. This is by no means a controversial hypothesis since there is ample evidence supporting that eye-movements are guided by cognitive factors such as context and semantics, where the physical image components interplay cognitively to give the raw image content a higher meaning (review evidence from Chapter 3).

In Chapter 6, we tested two popular algorithms to predict fixations, implemented by Walther and Koch (2006) and Rajashekar et al. (2008), and compared the predicted locations with those collected from participants watching the contrast manipulated images. Although there exist evidence supporting the contribution of low-level saliency to eye-movement guidance in both static (Parkhurst & Niebur, 2002) and dynamic (Itti, 2005) scenes, it has been shown that top-down factors such as task and context can override such contribution. For example, Underwood et al. (2006) used a search task where subjects were instructed to detect the presence of a low saliency target. This task yielded a low spatial overlap between saliency and fixation locations. Rothkopf et al. (2007) studied the deployment of gaze in a virtual environment during different tasks and found that task and context, instead of saliency, dominate gaze allocation. Everyday activities such as food preparation seem largely independent of objects' low-level properties (M. F. Land & Hayhoe, 2001). While the dominant influence of task on eye-movements has been long known (Buswell, 1935; Yarbus, 1967), significantly less work has been done using the opposite experimental strategy with neutral, free-viewing tasks and images with manipulated low-level statistics, which we use in this thesis. Overall, we found the algorithms being remarkably poor at predicting human fixations, in particular for low SID images where contrast had been reduced at semantically informative regions. These results together strongly question that the low-level features used by these algorithms contribute causally to fixation selection.

It is currently debated whether regions are looked at because they are informative with respect to their physical image properties (such as saliency) or due to their semantic informativeness. Henderson et al. (2007) reported that, besides having higher saliency than control regions, fixated locations were deemed as more semantically informative than control regions. Salient regions have also been shown to overlap with regions labeled as interesting (Elazary & Itti, 2008). By reducing the coupling between saliency and semantic informativeness, we found that semantically informative regions are looked at despite having a weak low-level signal strength. Therefore, the previously reported (correlative) link between fixation selection and saliency may in fact reflect the causal link between semantic informativeness and fixation selection. A predictor based on saliency can in other words output predictions that coincide with actual fixations collected from humans, but does so not because saliency attracts attention, but since underlying, semantically informative objects happen to contain features with high saliency. As we have seen, if such objects are reduced in saliency, they nevertheless attract fixations.

A number of studies have recently investigated the mechanisms controlling the first fixation, which usually refers to the fixation following the initial saccade after image onset. A general observation (and consensus) is that the position of the first fixation largely coincides across viewers (Tatler et al., 2005). However, the explanation for this observation varies. Whereas early studies reported that objects inconsistent with the general semantic category of the image (Loftus & Mackworth, 1978) and regions deemed as informative by viewers (Antes, 1974) attracted a disproportional amount of initial fixations, some later works have emphasized the contribution of image features. For example, Parkhurst et al. (2002), suggested that saliency contributes more to fixation selection during the first fixation and thereafter contributes less. Tatler et al. (2005), on the other hand, argue that the contribution of bottom-features does not change with viewing time. Instead, top-down influences do. Using our data, we again found that contrast manipulations affect the location of the initial fixation differently depending on the image category. The effect reported by Parkhurst et al. (2002) was found in the high SID, 'neutral' category. However, the opposite effect was found when regions rated as semantically important, such as faces, were reduced in contrast; subjects' initial fixations instead landed on regions with low contrast and edge density. Consequently, our results do not support the hypothesis that initial saccades causally are driven by saliency. Instead, it is likely that the gist of the scene provides enough information to guide the initial saccade. In fact, recent research has shown that an image's gist can be apprehended very quickly after image onset and includes "a rich collection of perceptual attributes" and "rises to conscious memory within a single fixation" (Fei-Fei et al., 2007).

We have in this part of the thesis analyzed fixated content at rather high spatial frequencies. For example, the filters we used in Chapter 5 were of size 3×3 pixels and operated on images of size 1024×768 pixels. Consequently, only image variations with high detail were extracted, whereas coarser variations were not captured by these filters. Mannan et al. (1995, 1996) investigated how lowpass filtering of an image affects where people look. They found that during the first 1.5 seconds of viewing, people fixate the same locations in the original image as in the lowpass filtered version of this image. Since only the low frequency content is shared between these versions, this suggests that a representation based on low spatial frequencies could be responsible to guide early fixations. In this sense, a saliency map operating on lower spatial frequencies could account for the results found in this paper. This line of argument has some support considering images from the 'face' category only; contrast manipulations dominantly attenuating higher frequencies have little influence on where people look, and faces are looked at regardless of their contrast levels. However, it seems more plausible that face regions are looked at because of their known semantic importance than because of some low-level account. Moreover, images from the 'neutral' category directly overthrow this assumption since fixation locations showed to be directly affected by the contrast manipulations in this case.

In summary, the results from this part of the thesis do not support the hypothesis of a causal relationship between fixation selection and image features, i.e., bottom-up features do not obligatorily attract visual attention.

Part II

Off-Line Foveation – Implementation and Evaluation

Chapter 8

Compression and Foveation

THE lack of spatial detail in peripheral vision allows a display to be reduced in quality at locations where a viewer does not look directly, without this being noticed by the viewer. In image and video compression, this fact can be exploited by allocating bits in accordance to the spatial sensitivity of the HVS; more bits are given to fovea-near regions than to peripheral regions. This is called foveated compression.

This chapter serves as an introduction and motivation to foveated compression. It begins with an overview of traditional methods for image and video compression, followed by an introduction to foveation; what it is, how it is implemented, and how it can be (and has been) used to improve image and video compression. Unlike commonly known methods for foveated coding relying on *real-time* implementations, we introduce an approach called *off-line* foveation where gaze data collected from several previewers are used to predict where later observers, watching the same videos, will look.

8.1 Some Words on Source Coding

As the digital information age matures, technological advances have allowed an increasing number of people to use a range of multimedia services. Video applications, in particular, have recently undergone an explosive growth. For the practical applicability of video communications, source compression is crucial. Without compression, video file sizes would be too large to store on many devices and use excessive bandwidth during transmission. Since this section only scratches the surface of the wide area of source coding, the interested reader is referred to the textbooks by Haskell & Netravali, 1995 and Sayood, 2000 for a more comprehensive treatment of the subject.

There are two types of compression: *lossless* and *lossy*. As the names imply, lossless compression requires the reconstruction of the source to be an exact replica of the original source, while in lossy compression a certain amount of distortion, that is, a discrepancy between the reconstructed and original source, is acceptable. A common goal in compression is to remove so called *redundancies* in a source, that is, repeated information that we can discard and still keep the crucial source elements. The source can comprise digitized text, speech, an image, or a video. Mathematically, a source can be described by a statistical model with alphabet $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$, where letters in the alphabet occur with probabilities $\mathcal{P} = \{P(A_1), P(A_2), \ldots, P(A_n)\}$.

Lossless compression is necessary in a variety of applications where distortion is not acceptable. For medical purposes, for example, distortion in an X-ray image may lead to misinterpretations and confuse authentic fractures with compression artifacts. In text compression, a single letter that is lost or distorted may change the meaning of a word or a sentence drastically. Of course, there is a price for not allowing image distortion after reconstruction. Lossless schemes usually do not compress to less than about three times of the original source. The theoretical limit for how much a source without memory, i.e., where the source elements are independent, can be compressed is defined by the entropy of the source (Shannon, 1948)

$$H = -\sum_i P(X_i) \log P(X_i) \tag{8.1}$$

where $\{X_1, X_2, X_3, \ldots, \}$ denotes a sequence generated from the alphabet $\mathcal{A}$. There are many well known methods for lossless compression, for example Huffman (1952) and arithmetic codes (Rissanen, 1976), exploiting statistical properties of the source, and Lempel, Ziv and Welsh (LZW) (Welch, 1984) coding, taking advantage of repeated patterns in the sources. The LZW implementation can be found in, for example, Adobe's Portable Document Format (PDF).

Lossy compression addresses the trade-off between rate and distortion, with the overall goal to simultaneously minimize the rate and the distortion. Besides addressing statistical and structural redundancies, lossy compression targets *psycho-visual* redundancies by taking advantage of the very forgiving nature of human visual or auditory perception. According to this philosophy, a source can be compressed until the fidelity is violated as judged by human observers. The vagueness of this statement indicates the subtle nature of lossy source coding, which is even more complicated since individual, subjective differences exist between humans; one person can judge the source quality as poor while another person judges the quality of the same source as being fair or even good. In image com-

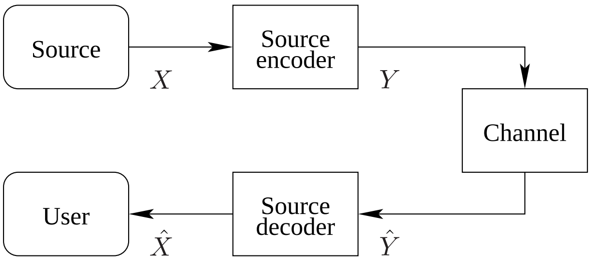


Figure 8.1: Overview of a compression scheme.

pression, lossy compression schemes can give about 30 fold compression of natural gray scale images with little or no perceived distortion.

Source coding is an important part of a *communication system*, which includes a source and channel encoder/decoder and sometimes also source encryption/decryption. We only consider source coding and assume that the channel is ideal and hence introduces no errors. As previously mentioned the source can be an image, speech, music etc. In this thesis we will consider only image and video sources. Figure 8.1 depicts a generic source coding scheme. A source X is fed into a source encoder which outputs a different (often binary) compressed representation Y of the original source X. To protect the encoded source from being corrupted when sent over the communication channel, redundancy can be added before transmission if the channel is not ideal. Since we only deal with ideal channels, in our case $\hat{Y} = Y$. For lossless compression we demand that $\hat{X} = X$ after source decoding while in lossy compression, we want to minimize the distortion. In other words, we want the reconstruction $\hat{X}$ to be as close to the original image X as possible, but at the cost of as few transmission bits as possible.

8.1.1 Image coding

Image coding is a special case of source coding. A typical system for image coding is outlined in Figure 8.2. It consists of an encoder and a decoder, which further are divided into a transform, quantization, and entropy coding stage. As a first step the image is transformed. The purpose of transformation is to decorrelate neighboring pixels and compact the majority of the image information into a small number of transform coefficients from an alphabet $\mathcal{C}$. Popular transforms are the discrete cosine transform (DCT) and the discrete wavelet transform (DWT), included in the standards JPEG (Wallace, 1992) and JPEG2000 (Taubman & Marcellin, 2001), respectively. The transform coefficients are then quantized, which can be defined as an operation that maps coefficients from $\mathcal{C}$ to

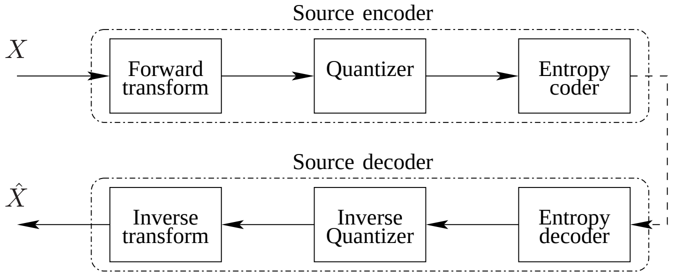


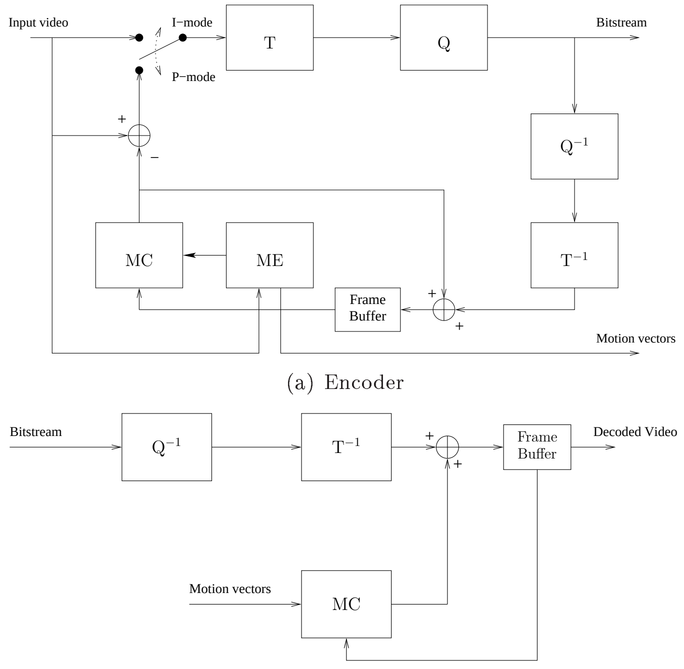
Figure 8.2: A lossy image coder.

another, coarser alphabet $\mathcal{C}_Q$. The purpose of quantization is to reduce the entropy of the coefficients. Lastly, an entropy coder is applied to the quantized coefficients. In JPEG, for example, the quantized output is entropy coded with run length coding (RLE) combined with Huffman coding. Decoding is generally straightforward, where entropy decoding is followed by inverse quantization and transformation. The degree of compression depends mostly on the quantization strategy, since both the transform and entropy coding stages are lossless or nearly lossless.

8.1.2 Video coding

A video consists of a sequence of images (called frames), each slightly different from its neighboring frames. Showing the frames quickly after each other creates the illusion of motion. In terms of compression, the most straightforward approach would be to code each frame as a still image. However, this approach is very inefficient. Instead, besides exploiting spatial redundancies as in image coding, a video coding scheme also exploits temporal redundancies through the fact that neighboring frames largely contain the same information. As a consequence, compression rates in video can be much higher than in still image compression.

The structure of a general video encoder/decoder is depicted in Figure 8.3. As a first step at the encoder, the input video is divided into a group of pictures (GOP), which typically consists of 8, 16, or 30 consecutive frames. In intra (I) mode, an input frame is directly transformed, quantized, and entropy coded, i.e., it is coded as a still image. In the predictive (P) mode, the (current) frame is first *predicted* from the previous decoded frame, and only the difference between the current predicted frame the current original frame, i.e., the prediction error (PE) is encoded. Prediction is made in two steps through *motion estimation* (ME) and *motion compensation* (MC). Figure 8.4 illustrates the general idea behind ME. Initially, two consecutive frames are divided into non-overlapping blocks. Each block in the current frame is then matched against blocks of the same



(a) Encoder

(b) Decoder

Figure 8.3: A typical video coder (without entropy coding).

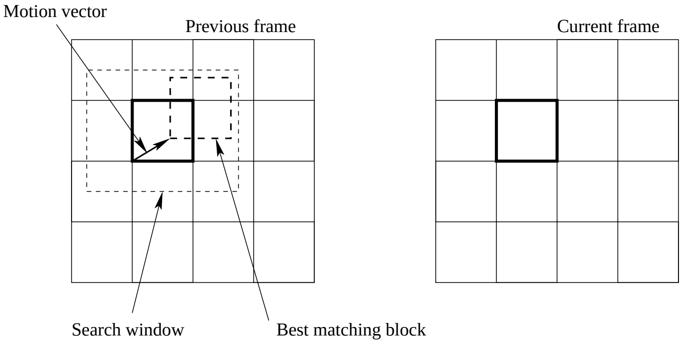


Figure 8.4: Block based motion estimation.

size in the previous frame within a search window. A vector describing the translational motion between the block in the current frame and the best matching block in the previous frame is stored. These motion vectors are used in MC to rearrange blocks of information in the previous frame to best describe the current frame. Algorithms for video coding following this basic framework have successfully been included in standards such as the moving picture experts group (MPEG) family of codecs (see, e.g., Gall, 1991; Wiegand, Sullivan, Bjontegaard, & Luthra, 2003).

8.1.3 Quality assessment in compression

To be able to design, implement, and evaluate an algorithm for compression, we need to be able to obtain accurate estimates of a compressed image's quality. The difference in quality between an original image X and its compressed representation $\hat{X}$ of dimensions $m \times n$ is typically measured with the mean squared error (*MSE*)

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} (X(i,j) - \hat{X}(i,j))^2 \tag{8.2}$$

or with the related peak signal-to-noise ratio (*PSNR*)

$$PSNR = 10 \cdot \log_{10} \left(\frac{[\max_{i,j}(X(i,j))]^2}{MSE} \right). \tag{8.3}$$

While these objective measures have been used extensively by researchers working with image and video compression, they have been found to correlate with the quality as perceived by human viewers quite poorly. This is hardly surprising since the HVS takes several aspects into account that are not considered by simple, pixel-based measures such as the *MSE*. For example, these aspects include (from Wang, Sheikh, & Bovik, 2003)

- Non-uniform retinal sampling.
- Light adaptation (luminance masking).
- Contrast sensitivity functions.
- Spatial frequency, temporal frequency and orientation selective signal analysis.
- Masking and facilitation.
- Contrast response saturation.

Unfortunately, there are today no objective measures that produce quality measures indistinguishable from those collected through subjective quality assessment. Currently, finding such objective methods is an active area of research (see e.g., Wang, Sheikh, & Bovik, 2003). Instead, to ensure reliable quality scores, experiments where several observers view and assess compressed images on rating scales (e.g., bad, poor, fair, good, excellent) or impairment scales (e.g., very annoying, annoying, slightly annoying, perceptible but annoying, imperceptible) are preformed. (ITU, 2002).

8.2 Using Foveation in Compression

As we saw in Chapter 2, humans have evolved a foveated system that combined with eye-movements is used for visual exploration. *Foveated compression* exploits the foveated nature of the HVS by removing undetectable high frequency content away from the foveation center as a function of eccentricity. Since high frequency content generally requires more information to represent digitally than low frequency content, foveation inherently improves compression. Although the huge potential to exploit foveation for the purpose of compression has been known for quite some time (formalized in e.g., Girod, 1988), it is today not a widespread technology. The reasons for this are mainly twofold. First, the compression system needs to know, or accurately estimate where the viewer looks. Second, in real-time applications, the delay introduced by coding and transmission is believed to exceed that acceptable to an observer. This delay causes a lag between the position of the current foveation point and the foveation center in the image currently being decoded. In other words, it cannot be guaranteed that the position where a viewer looks and the position where the decoded image has its best quality are aligned. Moreover, compression standards need to be extended to optimally code foveated image and video representations.

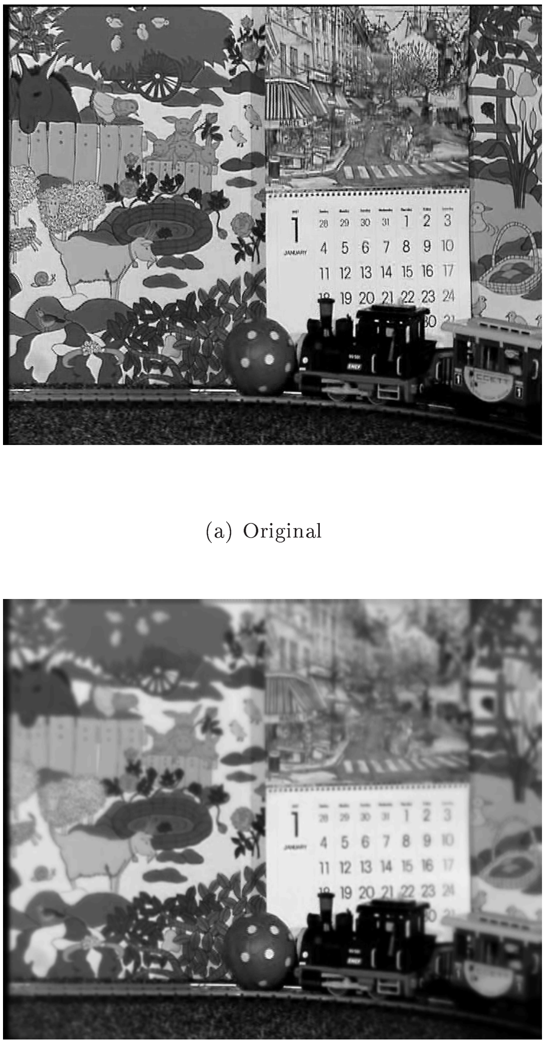
8.2.1 Foveation

Unlike the composition of a digital image as a uniform, two dimensional grid of pixels, acquisition of visual information on the retina is highly nonuniform with the highest sampling density in the fovea. The process of matching the image resolution in accordance to the sampling density of photoreceptors on the retina is called *image foveation* (Kortum & Geisler, 1996). Successfully implemented, foveation transforms an image such that a viewer looking at the foveation center cannot distinguish the foveated version from its original. Figure 8.5 illustrates image foveation; Figure 8.5(a) shows an unprocessed image and Figure 8.5(b) depicts the image after foveation, which is centered at the ball being pushed by the train.

There are a number of methods proposed to implement foveation. Early ones were based on adding pixels into larger elements, SuperPixels, which increase in size with increasing eccentricity from the point of gaze according to a resolution fall-off model consistent with anatomical measurements in the human retina and visual cortex (Kortum & Geisler, 1996). This type of implementation is simple and quick. However, borders between SuperPixels give rise to distinct blocking artifacts, which proved to be visually unpleasant. More recent implementations used multi-resolution pyramids (e.g., Geisler & Perry, 1998, 1999), where peripheral regions in the foveated image contain information from upsampled, higher pyramid levels, while regions closer to the foveation center comprise higher frequency content available from the low pyramid levels or from the original image itself. The borders between pyramid levels are typically smoothed with a blending function to avoid sharp transitions in the foveated image. Figure 8.5(b) has been generated using a foveated multi-resolution pyramid. The code is available online (http://svi.cps.utexas.edu/software.shtml).

Foveation has also successfully been implemented in the transform domain, using wavelets (Chang & Yap, 1997; Duchowski, 2000; Sheikh, Liu, Evans, & Bovik, 2001), and the discrete cosine transform (DCT) (Bergström, 2003) where appropriate transform coefficient scaling prior to inverse transformation produces foveated images. Typically, transform coefficients are scaled by a factor, $p \in [0, 1]$. Where p is small, the display quality is heavily degraded and where p is one, the quality is unaffected compared to the original display. Moreover, foveation has been implemented using polar down-sampling schemes (Juday & Fisher, 1989; Kuyel, Geisler, & Ghosh, 1999). The choice of implementation method depends on the application. Typically, smooth and artifact free resolution degradations are desirable.

In this thesis, we have chosen to implement foveation in the wavelet domain, mostly because the successful application of wavelets in compression. For example, the transform stage in many current state-of-the-art compression methods is based on wavelets (e.g., JPEG2000). Implement-



(a) Original

(b) Foveated

Figure 8.5: *Image foveation.* The bottom picture shows a foveated version of the original image on the top. Foveation center is located in the middle of the dotted ball. The foveated image was generated by the software publicly available from http://svi.cps.utexas.edu/software.shtml.

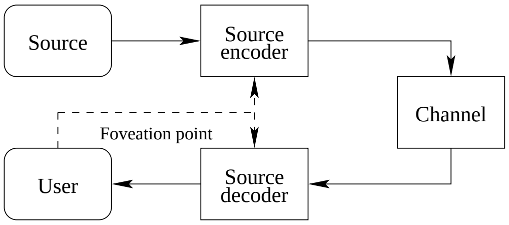


Figure 8.6: Overview of a foveated compression scheme.

ing foveation requires a number of parameters to be known (or accurately estimated). First, we need to know where a person looks. Second, we need a function approximating how visual sensitivity decreases as a function of eccentricity. Third, we need to know the distance from the image to the viewer. Fourth, we require knowledge about the resolution of the image and the screen on which it is presented, as well as the physical dimensions of the screen.

Foveated displays have been used for a number of purposes (see, e.g., Parkhurst & Niebur, 2002), for example to reduce computational resources in computer graphics rendering, to evaluate the perceptual span in scene perception, and to improve the compression efficiency of digital images and videos, which is the specific target of investigation in this thesis.

8.2.2 Foveated compression

Foveation improves compression efficiency by removing high frequency content, which typically consumes a substantial portion of the bit budget, from unattended parts of an image. We have identified two major categories of foveation-based, or foveation-like methods for improved image and video compression: *Real-time* and *off-line*. Maybe the most straightforward, intuitive approach to foveated coding is in real-time, first pointed out by Girod (1988), with potential applications in, e.g., surveillance, teleoperating of remote vehicles, telemedicine, and teleconferencing; these are situations where transmission bandwidth may be limited. In a typical situation shown in Figure 8.6, the position of the foveation center is sent to a remote location (camera) where the image is foveated in the spatial or transform domain, compressed with a standard coder such as JPEG, and transmitted back to the viewer where it is decoded and displayed. Of course, this type of setup requires a minimum delay from the time the foveation point is acquired until the image is decoded and displayed. Otherwise, it cannot be ensured that the foveation center and the region with best image quality coincide, which would reduce the subjective quality of the decoded image. Real-time foveated compression requires a fast and reliable link to transmit the foveation point to the encoder side, a quick algorithm to implement foveation, and symmetric coding schemes of relatively low complexity. Recent work has found that if a gaze-contingent display is updated within 60 ms after an eye-movement, blur due to foveation is not detectable (L. C. Loschky & Wolverton, 2007). One strategy to alleviate the effects of larger delays is to foveate an image while predicting how viewers' gazes change during the period of the delay (Khan & Komogortsev, 2006). The penalty is that a larger portion of the image needs to be represented in high quality than if the gaze positions would be known exactly. Real-time foveation has been reported to substantially improve compression. The bit rate savings depend on factors such as image size and viewing distance, but typically contributes with a factor ≥ 3 compared to standard 'unfoveated' compression (Geisler & Perry, 1999). With only minor changes in system design and implementation, real-time foveated compression can easily be extended to consider multiple foveation points (viewers).

A perhaps less intuitive way to use the fact that vision is reduced in the periphery, that we have named *off-line foveation*, is to beforehand predict where viewers will look and keep a high display fidelity only in these regions, while degrading other regions. Given that later viewers look within the predicted regions and that the peripheral degradation does not introduce visually unpleasant video distortions, off-line foveation will theoretically not reduce subjective quality. Obviously, besides exploiting peripheral image degradations to improve compression, off-line foveation relies on the assumption that different viewers will look at similar locations. If this was not the case, and if viewers' gaze positions were uniformly distributed, no region could be degraded without significantly reducing the perceived quality for an uncontrollable number of later viewers. Fortunately, there is ample evidence that different viewers look at largely similar video regions (Elias, Sherwin, & Wise, 1984; Stelmach, Tam, & Hearty, 1991; Tosi, Mecacci, & Pasquali, 1997; Goldstein, Peli, Lerner, & Luo, 2004; Dorr, Böhme, Drewes, Gegenfurtner, & Barth, 2005). Most of the time, these regions are confined to the center of the video display. The general structure of a system for off-line foveated compression is the same as in Figure 8.6. However, the foveation points are replaced by estimates of the locations where future viewers are likely to look. In our implementation, estimates come in the form of gaze density functions (GDFs) generated from superimposed Gaussian functions derived from empirical gaze data collected from previewers. Also, since the encoder is not constrained by any real-time computational demands, off-line foveation allows for a more non-symmetric construction where complexity can be shifted to the encoder. Off-line foveated compression is mainly suitable for, but not limited to, off-line, and semi real-time applications such as sports and news broadcast and streaming video over the Internet.

In addition to being real-time and off-line, foveated compression can

	Real-time	Off-line
Non-scalable	Juday and Fisher (1989) Kortum and Geisler (1996) Geisler and Perry (1998) Sheikh et al. (2001) Bergström (2003) Khan and Komogortsev (2006)	Itti (2004) Agrafiotis et al. (2006)
Rate scalable	⟵Wang and Bovik (2001)→ ⟵Wang, Lu, and Bovik (2003)→	

Table 8.1: Categorization of some papers on foveated image and video coding.

be rate scalable or not (see Wang & Bovik, 2001; Wang, Lu, & Bovik, 2003). Scalability in foveated compression refers to the ability to order the bit stream such that regions close to the foveation center are coded and transmitted with priority. As a consequence, when initial parts of the bit stream are received at the decoder side, the foveated region consumes bits almost exclusively, and is therefore reconstructed with higher fidelity than other regions. At this point, only a heavily foveated image version can be decoded. As more bits get available to the decoder, regions further away from the foveation center are successively refined. When the whole bit stream is decoded, the received image is fully 'unfoveated'. In foveated video compression, scalability can also refer to temporal scalability, where foveated regions are prioritized in frame rate. Table 8.1 lists a number of representative works from each category.

8.2.3 Off-line foveation: Open problems

One of the main challenges in off-line foveated video is how to accurately predict where future viewers will direct their gazes. There have been two main approaches: Using *eye-movements from a number of previewers* watching the video (Stelmach & Tam, 1994; Duchowski & McCormick, 1998), and using *computational algorithms* for automatic prediction (e.g., Osberger & Rohaly, 2001; Wang, Lu, & Bovik, 2003; Itti, 2004; Le Meur, Le Callet, & Barba, 2007).

Without explicitly targeting video coding applications, the use of previously recorded eye-movements to implement off-line foveation was presented and evaluated by Stelmach & Tam, 1994 and Duchowski & Mc-

Cormick, 1998. Stelmach and Tam manipulated each video frame such that the one region where most previewers looked remained in high resolution, whereas other parts of the frame were increasingly degraded away from this region by means of low-pass filtering or DCT coefficient quantization. The perceived quality of the manipulated, variable-resolution video was assessed by human observers and compared with three other versions of the same video; one unprocessed, one with an equal level of blur distributed uniformly over the frame, and one with a centrally fixed high-resolution region. As expected, the unprocessed video got the highest quality ratings and the uniformly blurred video the worst. The authors found, rather surprisingly, that the judged quality of the off-line foveated video was comparable to having a centrally fixed high resolution region throughout the video. In view of these results, Stelmach and Tam (1994) conclude that "Given the modest benefits and high cost of implementation ... gaze contingent processing is not suitable for general purpose processing". However, as they also discuss, the poor quality ratings of the off-line foveated sequence may derive from repeated viewings of the test sequences as well as the imposed task of quality evaluation, which could make subjects actively search for quality impairments. Either of these two reasons may disrupt the natural viewing behavior of subjects and hence cause them to gaze outside the regions of high resolution where the image quality is significantly decreased. A similar study by Duchowski and McCormick (1998) investigated the subjective quality of videos that were manipulated (off-line) such that high resolution was maintained around each previewer's position of gaze (from several viewers), whereas other regions were degraded in resolution. Results showed that eye-movements collected from subjects watching the manipulated videos deviated from eye-movements collected from the unprocessed, original video. The authors argue that new, suddenly appearing high-resolution regions may distract viewers' natural viewing patterns in the former case. Apparently, both Stelmach and Tam and Duchowski and McCormick came to the conclusion that off-line foveation is infeasible since it introduces video artifacts decreasing the subjective quality, and also seems to change the viewing behavior of new viewers.

Computational models for gaze prediction typically use low-level image features such as luminance, contrast, edge density, and motion (cf. Chapter 6 for gaze prediction in images), or use heuristic rules such as 'always choose faces'. Although there exist a few implementations using computational approaches for gaze prediction to generate off-line foveated videos (e.g., Osberger & Maeder, 1998; Itti, 2004) none, to the author's knowledge, has been subjectively evaluated. Interestingly, a recent study showed that the best among current state-of-the-art gaze predictors in video was one simply predicting that viewers would look at the center of the screen (Le Meur et al., 2007). As for gaze prediction algorithms in still images discussed in the first part of the thesis, automatic gaze prediction in video is currently quite far from producing data consistent with those recorded from human observers. This motivates the use of eye-tracking data, which define the 'ground truth', to predict future gazing behavior for the purpose of off-line foveation. In this thesis, therefore, we have adopted this approach.

There are a number of central challenges in off-line foveated compression that we will address in the coming chapters. First, it is an open question how recorded gaze positions best are transformed into a foveation function, that is, a function that manipulates the video quality such that peripheral degradations do not compromise the subjective quality experienced by later viewers. Imagine for example that gaze data is collected from 14 previewers; 11 look at an object in the upper left corner and the other three look toward a region in the lower right corner. When foveating and coding the video to be looked at by other viewers, how many bits do we want to spend in the lower right corner compared to the upper left corner? Second, given a foveation function, how is it used to efficiently allocate bits in a coding scheme? Finally, assuming the first two problems are solved, how can we estimate the quality of the foveated and coded video? Obviously, objective quality estimates such as the *PSNR*, which treats different image regions without regard either to the varying spatial nature of foveated images or to the collective viewing behavior, are not directly applicable to evaluate off-line foveated video. These and other issues will be the targets of investigating in the coming chapters.

8.3 Summary

Foveated compression exploits the non-uniform spatial acuity of the human visual system (HVS) by removing high spatial frequencies not detectable by our peripheral vision. By representing only the regions in a video where people look in high quality while degrading other regions, foveation has the potential to significantly improve today's state-of-the-art methods for compression. In a system for *real-time foveation*, a foveation point is sent from the viewer to a remote camera where the image is foveated, encoded, and directly transmitted back to the viewer. At the decoder side, the image is rapidly decoded and displayed. In a different approach to foveated coding that we have named *off-line foveation*, gaze positions are collected from a number of previewers. These gaze positions are then used to manipulate the image quality such that later viewers will not perceive the blur introduced by off-line foveation. Previous works on off-line foveated video argue against the feasibility of such an approach. In the coming chapters, we will revisit off-line foveation and evaluate its potential in compression by addressing a number of open research problems.

Chapter 9

A First Glance Toward Off-Line Foveated Compression

We begin to explore off-line foveated compression using eye-tracking experiments combined with a simple coding scheme. Foveated compression is applied to six short image sequences depicting natural scenes, where each image is foveated and compressed without regard to its neighboring images.

9.1 Overview

Figure 9.1 gives an overview of the system design. It consists of three main building blocks, each outlined by a dotted box. Initially, eye-movements are recorded from 17 people free-viewing the original image sequences. To implement foveation, each image from a sequence is wavelet transformed, and the wavelet coefficients are multiplied by a weighting function deriving from collected gaze positions. The foveated coefficients are finally quantized with a simple, uniform scalar quantizer and entropy coded with a Huffman coder. Decoding reverses the entropy code and transforms the wavelet coefficients back to the spatial domain. The degree of additional compression due to off-line foveation is calculated. In the evaluation phase, another 18 people look at the foveated, decoded image sequence under the same conditions as during the initial data collection. Again, their eye-movements are recorded. The purpose of a second recording is to compare where subjects look in the original sequence to where they look while watching the compressed off-line foveated sequence. Since standard methods for subjective and, in particular, objective quality evaluation are not directly applicable to off-line foveated video, we argue that comparing

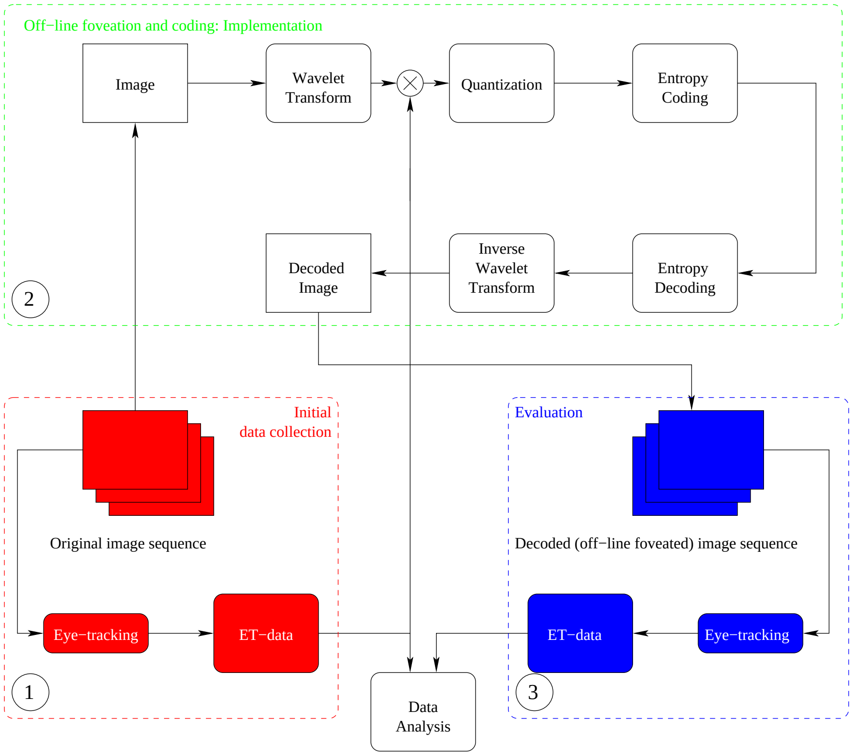


Figure 9.1: System overview

the distribution of gazes in the two conditions serves as an indicator of the perceived quality. For example, if people look at similar locations across the conditions, we know by definition that they gazed toward regions with high quality. Otherwise we know that they looked at regions degraded by foveation, which were thus of poorer quality. Besides analyzing the gazing behavior, we asked subjects some questions about their subjective viewing experience.

9.2 Methods

9.2.1 Data collection

Test subjects were seated one by one at a viewing distance of 75 cm in front of a computer screen. The screen extended 31×25 cm (23×19 degrees) and had a resolution and refresh rate of 720×576 pixels and 60 Hz, respectively. All observers had normal or corrected-to-normal vision. Image sequences were played with the Quicktime 6.3 player at 25 frames per second (fps).

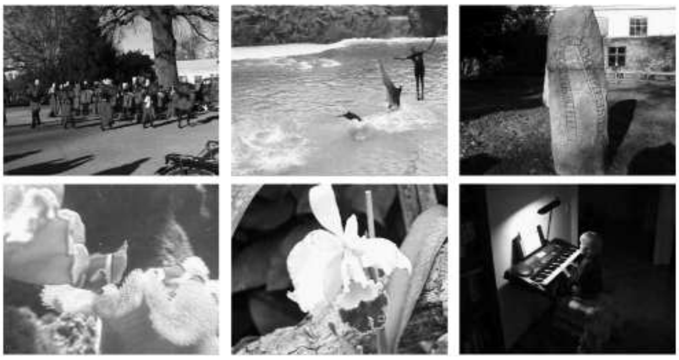


Figure 9.2: Representative image (Y-component) from each tested sequence.

To enable fast and accurate display, the image sequences were encoded at a high bitrate. Stimuli consisted of six short image sequences depicting natural scenes, and had a total duration of 3 min and 30 seconds. The resolution of the images was the same as the screen resolution. The images were represented in 24 bit color (RGB with 8 bit in each color channel). A representative image from each sequence is shown in Figure 9.2. During image display, gaze positions were recorded at 50 Hz with an SMI iView eye-tracker using a pupil/cornea reflex system to track the eyes. Subjects were naive in the sense that they had no prior knowledge of either the content of the stimuli or the purpose of the test. Prior to each eye-tracking session, subjects did a nine-point calibration and were instructed to 'free-view' the sequences ('watch the videos as you naturally would do at home').

During the initial eye-movement data collection, we had 17 subjects watching the original image sequence. The collected eye-movement data from 14[1] of these subjects were used for the purpose of foveation.

In the second data collection (the evaluation phase), 18 new subjects watched the foveated and compressed image sequence under the same conditions as in the first test. Again, data from 14 subjects were used.

9.2.2 Off-line foveation and coding – Implementation details

Each image from the sequences is foveated and coded separately. First we exploit the fact that humans are less sensitive to chromatic than to luminance information by a RGB-to-YUV conversion, where the U and V components are subsampled by a factor two. The YUV components

[1] Eye-tracking data from the three test-subjects with the most deviant (from other test-subjects) eye-movement patterns were omitted.

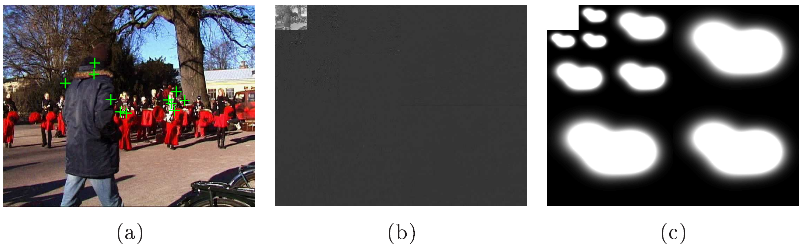


Figure 9.3: (a) Image with overlaid gaze positions. Each marker represents the gaze position from one viewer. (b) Three level wavelet decomposition of the Y-component. (c) Subband weighting masks for a three level wavelet decomposition.

are each wavelet decomposed (c.f. Appendix A) using a Daubechies 4-tap filter with periodic border extension. Each component is decomposed with three levels as depicted in Figure 9.3(b). Foveation is implemented in the wavelet domain by weighting (multiplying) the coefficients in each subband B_λ at decomposition level $\lambda = \{1, 2, 3\}$ with a Gaussian-like function $W_\lambda(m, n)$, whose shape is determined by the distribution of gaze positions. This way, high frequency information is attenuated in regions largely unattended by viewers' gazes. More precisely, if P denotes the number of viewers, (m_i, n_i) denotes the position gazed at by viewer i, and $M \times N$ define the image dimensions, then

$$G_\lambda(m,n) = \sum_{i=1}^{P} - \exp\Big(\frac{(m-m_i)^2 + (m-m_i)^2}{2\sigma^2}\Big) \tag{9.1}$$

$$m = 1, 2, \ldots, \frac{M}{2^\lambda}, n = 1, 2, \ldots, \frac{N}{2^\lambda}$$

defines a *gaze density function* (GDF). $W_\lambda(m, n)$ relates to the GDF as follows

$$W_\lambda(m,n) = \min_{m,n}\big(1, G_\lambda(m,n)\big) \tag{9.2}$$

Consequently, values larger than one are truncated so that $W_\lambda(m, n)$ consists only of values on the interval (0 1]. Figure 9.3(c) gives an example of how the weighting function $W_\lambda(m, n), \lambda = \{1, 2, 3\}$ is composed in the wavelet domain. Notice how coefficients from the lowest frequency band, LL_3, are unaffected by the weighting to ensure a crude background quality in the decoded image. The parameter σ in Eq. (9.1) controls how fast the display quality is degraded away from regions with high gaze density. In our experiments, we use $\sigma = 0.10M$. This lets the 'full width at half max' (used by, e.g., Rajashekar, Cormack, & Bovik, 2004) of each Gaussian function centered at a gaze position cover the foveal span of an observer, and also accounts for the uncertainty introduced by allowing new viewers to watch the off-line foveated video.

Wavelet coefficient weighting is followed by quantization. We use a very simple quantization strategy where coefficients at levels $\lambda = \{1, 2, 3\}$ are quantized with respectively $\{1, 3, 4\}$ bits using a scalar uniform quantizer with the step size optimized for a Laplacian distribution (see e.g., Table 8.3 on p. 225 in Sayood, 2000). The lowest frequency band LL_3, however, is quantized with the step size optimized for a uniform distribution, using 8 bits. Quantized wavelet coefficients are as a last step entropy coded with a Huffman coder. Decoding is straightforward as shown in Figure 9.1.

9.3 Data Evaluation

Image compression algorithms struggle with the trade-off between maintaining a good perceptual quality and at the same time obtaining low bitrates. Unfortunately, there exist currently no objective methods for quality evaluation that produce results indistinguishable from those obtained by human observers. In particular, standard methods for objective quality evaluation would fail miserably if applied to off-line foveated video. For that matter, it is not even clear if standard methods for subjective quality assessment would yield reliable results. We address these concerns by collecting eye-movements from a new group of viewers watching the off-line foveated image sequence, and compare their gaze positions against those collected from the original image sequence. If gaze positions coincide across the two conditions, foveated compression does not change where people look. Consequently, the new viewers look at regions where the quality is high. This is an obvious prerequisite for off-line foveated compression.

To quantify whether off-line foveated compression changes where people look, we define two measures based on the collected gaze data: *between-group (BG) difference* and *within-group (WG) similarity*. The BG difference measures the degree of similarity across any two sets of gaze positions $\mathcal{A}$ and $\mathcal{B}$. We use a modified version of the Kullback-Leibler distance (KLD) (Cover & Thomas, 1991) to define this similarity mathematically. In its standard from, the KLD is expressed as

$$\mathfrak{D}(p \parallel q) = \sum_{\mathfrak{X}} p(x) \log_2 . \frac{p(x)}{q(x)} \tag{9.3}$$

$p(x)$ and $q(x)$ are probability density functions (PDFs) of a discrete random variable X with alphabet $\mathfrak{X}$. The KLD, also known as the *relative entropy*, is a known information theoretic measure and can be thought of as a distance, alas non-symmetric, between two PDFs; it equals zero if and only if the distributions are identical. The more the distributions differ, the larger this distance will be. To address the non-symmetric properties of the KLD, we define the BG difference, $S_{\mathrm{BG}}(\hat{G}^{\mathcal{A}}, \hat{G}^{\mathcal{B}})$ as the harmonic KLD (hKLD) (used by, e.g., Rajashekar et al., 2004) between the normalized GDFs $\hat{G}^{\mathcal{A}}$ and $\hat{G}^{\mathcal{B}}$

$$S_{\mathrm{BG}}(\hat{G}^{\mathcal{A}}, \hat{G}^{\mathcal{B}}) = \left(\frac{1}{\mathfrak{D}(\hat{G}^{\mathcal{A}} \parallel \hat{G}^{\mathcal{B}})} + \frac{1}{\mathfrak{D}(\hat{G}^{\mathcal{B}} \parallel \hat{G}^{\mathcal{A}})} \right)^{-1} \tag{9.4}$$

where $\hat{G}(m,n) = G(m,n)/(\sum_m \sum_n G(m,n))$, and $G(m,n)$ is defined as in Eq. (9.1) with $\lambda = 0$.

The WG similarity quantifies the degree to which subjects' gaze positions are spread out over the screen area. Obviously, in order to achieve large bitrate savings, gaze density must be constrained to limited regions, considerably smaller than the whole display area. The WG similarity, S_{WG} across gaze positions for any set $\mathcal{A}$ is found by computing

$$S_{\mathrm{WG}} = S_{\mathrm{BG}}(\hat{G}^{\mathcal{A}}, U(\Omega)) \tag{9.5}$$

where $U(\Omega)$ denotes the uniform distribution spanned by the image area Ω.

In this chapter, a set ($\mathcal{A}$ or $\mathcal{B}$) will comprise either gaze positions collected during the display of one image from a sequence, or positions drawn from an underlaying distribution (e.g., Gaussian or uniform).

9.4 Results

9.4.1 Compression due to off-line foveation

Off-line foveation prior to quantization and coding reduces the bitrate with, on average, 17.8% in our tested image sequences, despite using a simple coding method not in any way optimized to encode foveated images. Figure 9.4 illustrates an image-by-image comparison in bitrate between the original and off-line foveated image sequences. Notice the increased variability in bitrate due to off-line foveation, which is a result of the constantly varying size of the weighting function used to control the bit allocation. Of course, the potential for improved compression due to foveation reaches its peak when all tested subjects gaze toward exactly the same position. On the other hand, if the gaze density is evenly distributed over an image, off-line foveation may yield no or very little bitrate gain. Figure 9.5 shows two images from the tested sequences where off-line foveation had the largest (Figure 9.5(a)) and smallest (Figure 9.5(b)) impact on compression. As can be seen, a compact gaze density across viewers is an important aspect for improved compression. Just as importantly, however, is the frequency content of the unattended regions;

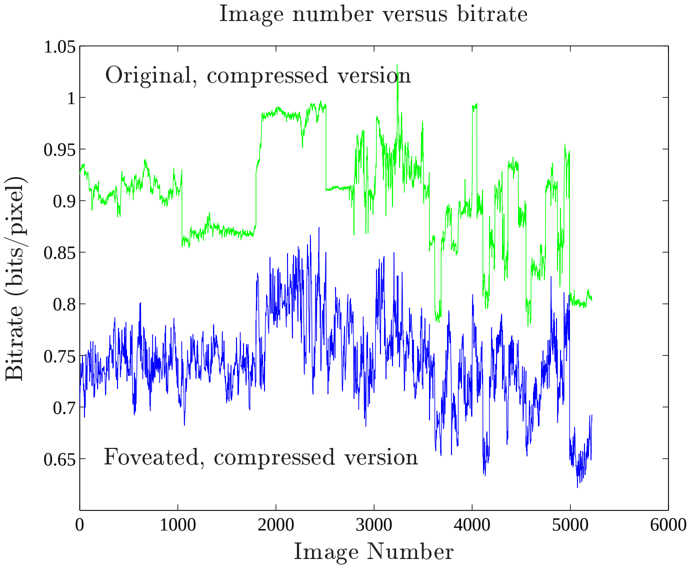
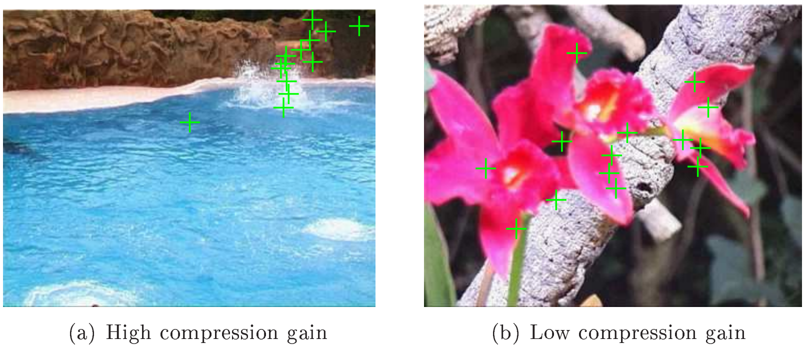


Figure 9.4: Bitrates of the original and off-line foveated image sequences after compression.

removing much high-frequency information due to foveation greatly increases the degree of compression. If on the other hand the unattended regions already are of lowpass nature, little additional gain in compression is won by foveation.

9.4.2 Evaluation

Figure 9.6(a) shows the within-group (WG) similarity across gaze positions collected in the initial data collection (first column), referred to as 'Original', and those collected during the evaluation phase (second column), named 'Foveated'. Each value reflects the WG similarity among gaze positions collected from one image. A large value on the y-axis indicates a high similarity. For comparison, the WG similarity for random viewers is shown (third column), where 14 gaze positions were drawn from a uniform distribution for each image. The similarities are visualized with box plots. Each box has lines at the lower quartile, median, and upper quartile values. The whiskers extend to 1.5 times the inter-quartile range and values outside this interval are considered as outliers and represented by plus signs. The notch in each box reflects the uncertainty in median in a box-to-box comparison. If the notches between two boxes do not overlap, they have different medians with 95% significance. As can be seen from Figure 9.6(a), the WG similarity is significantly larger ($p < 0.05$) across different human viewers than across positions drawn at random.



(a) High compression gain

(b) Low compression gain

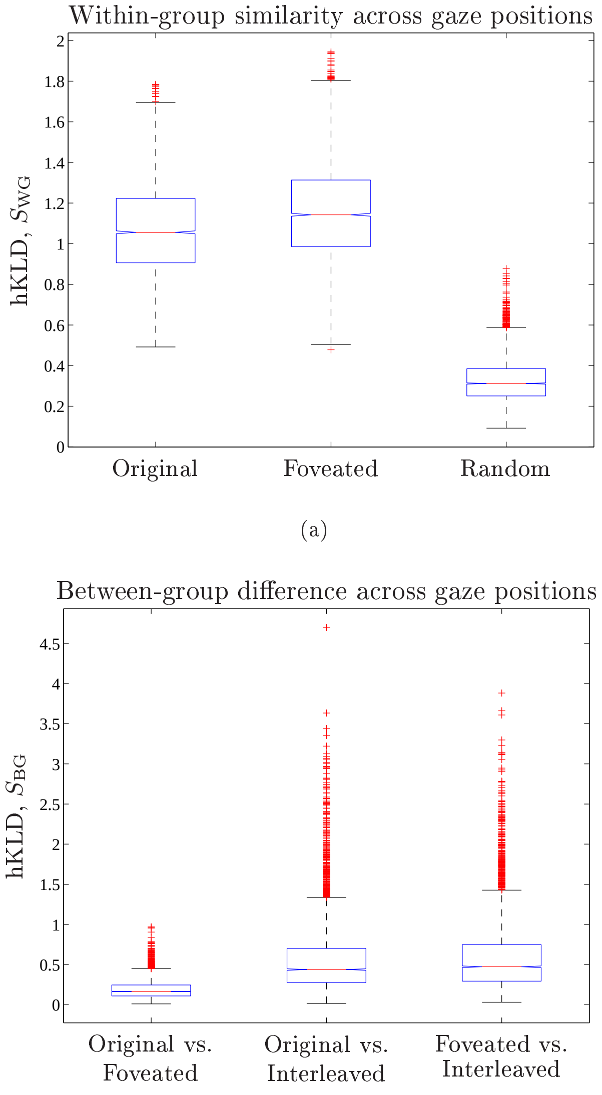
Figure 9.5: Images with the highest (28.2%) and lowest (10.1%) additional bitrate reduction due to off-line foveation.

This is true for both the 'Original' and 'Foveated' data. Clearly, viewers look toward limited parts of the display, and their viewing behavior is not of 'random' nature. Moreover, it can been seen that 'Foveated' gaze positions are more compact than the 'Original'. This could imply that the peripheral blur introduced by off-line foveation repels new viewers' gazes, which instead are attracted to regions with high quality.

While the higher-than-random WG similarity reveals that the distributions of gaze positions are compact, it does not tell us whether two distributions of gaze positions coincide spatially. Therefore, we compute between-group (BG) differences, which are illustrated in Figure 9.6(b). The first column plots the difference between 'Original' and the 'Foveated' gaze positions whereas the second and third columns illustrate the difference between 'Original vs. Interleaved' and 'Foveated vs. Interleaved', respectively. When interleaving gaze positions, we assign each image with gaze positions taken from a different, non-contiguous image in the sequence. Interleaving is done to compare the collected data against 'random' viewing behavior, which includes the central bias inherent in typical gaze data. Figure 9.6(b) shows that different viewers' gaze positions largely coincide when watching the same video before and after off-line foveated compression.

9.4.3 Viewer ratings and comments

Directly after the eye-movement recording in the evaluation phase, subjects were asked to name one or many scene(s) that were of better or worse quality than the others. To reduce the potential top-down bias on eye-movements that a quality evaluation task could give, subjects were not informed of this quality assessment in advance. Although giving only a crude estimate of the quality of the off-line foveated image sequences,



(a)

(b)

Figure 9.6: Within group similarity (a) and between group difference (b) across gaze positions recorded from viewers watching the original image sequence ('Original') and those watching the off-line foveated and compressed video ('Foveated'). 'Interleaved' gaze positions originate from the 'Original' image sequence, after the image order has been randomly shuffled.

the assessment showed some interesting tendencies. Figure 9.7 illustrates these. The numbering is according to Figure 9.2 (top left to bottom right). In particular, the quality of the first and the second image sequences mirrored each other. The first sequence contained many objects with vivid colors whereas the second contained a few main objects (the dolphins) performing acrobatic tricks. This leads us to assume that foveated compression works better when there is one or a few main objects reliably attracting viewers' gazes. Subjects were also able to freely provide feedback on the viewed sequences. One of the most frequent comments was that single, isolated high quality regions seemed to 'float around'. The same effect was reported by Duchowski and McCormick (1998). Such artifacts derive from individual viewers whose deviant gaze positions were nevertheless used to foveate the images.

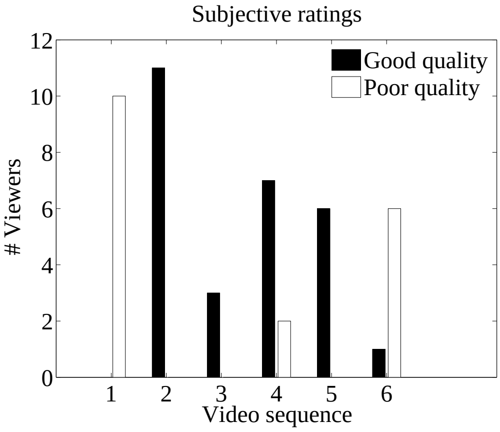


Figure 9.7: Subjective rating of the six image sequences. Subjects were asked to name the image sequences that stood out from the others in terms of better or worse quality.

9.5 Summary

In this first look at off-line foveated compression, we found that: 1) Off-line foveation prior to compression yields an additional 17.8% bitrate reduction. These results are obtained without exploiting temporal reduncancies, as used in coding of video. 2) Viewers largely look toward similar regions when watching image sequences. 3) Off-line foveation affects subjects' viewing behavior only a little, with a slight shift toward regions kept in high quality. Since viewers look toward regions kept in high quality, it is likely that their subjective quality remains high, given that peripheral regions do not contain easily identifiable artifacts.

Chapter 10

Using Volumes of Interest in Off-Line Foveated Video Compression

IN contrast to what previously has been reported in the literature, the results from last chapter support that off-line foveation can benefit image sequence compression. At the same time, some of the subjects viewing the off-line foveated sequences reported that rapidly appearing and disappearing high quality regions were disturbing, hence decreasing the subjective quality. To maintain a high subjective quality, it seems crucial to transform collected gaze data into a function smoothly controlling the spatio-temporal amount of blur introduced by off-line foveation. Using the observations and design issues from (Stelmach & Tam, 1994; Duchowski & McCormick, 1998), discussed in Chapter 8, combined with those from the previous chapter, we will in the current chapter design, implement and evaluate an improved and more elaborate system for off-line foveated compression. The remainder of this chapter is structured as follows. Section 10.1 describes how gaze positions are used to define smooth volumes of interests (VOIs), which are used in Section 10.2 to implement off-line foveation through wavelet domain filtering. We use the state-of-the-art video codec H.264 to encode the off-line foveated sequences, and compute the bitrate gain due to off-line foveation prior to compression. Finally, evaluations are performed in Section 10.3 to answer how off-line foveation affects subjective quality and viewing behavior.

10.1 Creating Volumes of Interests (VOIs) From Gaze Positions

Volumes of interests (VOIs) are derived from gaze data in the following steps:

Gaze positions $\overset{(A)}{\rightarrow}$ GDF $\overset{(B)}{\rightarrow}$ Intra-frame ROI $\overset{(C)}{\rightarrow}$ Inter-frame ROI $\overset{(D)}{\rightarrow}$ VOI

Each of these steps will now be described in detail.

Step (A)

Initially, gaze coordinates are processed per frame and represented by *gaze density functions* (GDFs), denoted $G(m,n)$ (See Eq. (9.1) for a definition). The widths of the Gaussian functions composing the GDF are motivated by setting the parameter σ such that when a Gaussian function is cropped at half its maximum height, the slice plane or *active area* (Wooding, 2002) spans the foveal region of an observer viewing the video at a distance d. If α denotes the visual angle, then σ is easily found as

$$\sigma = \sqrt{\frac{-\big(d\tan(\pi\alpha/360)\big)^2}{2\log_e(1/2)}} \tag{10.1}$$

A GDF reflects the likelihood of where future viewers will direct their gazes and contains valuable information about where the ROIs are located.

Step (B)

Using GDFs to predict ROIs, we address two heuristic design criteria. First, ROIs should be representative for viewers of the off-line foveated video and take into account the uncertainty of where new viewers will look relative to those originally recorded from. Obviously, there is a trade-off between keeping the ROIs as small as possible (and thus maximizing the bitrate gain due to off-line foveation), but large enough to encapsulate the gazes of as many new viewers as possible. Second, besides the global peak of a GDF, local peaks in gaze density may indicate potentially interesting regions and must therefore have the chance to be fully recognized as ROIs.

To resolve the first issue, we compute the inter-subject gaze point dispersion across P viewers as

$$S = \frac{1}{P}\sum_{i=1,2,\ldots,P}\frac{G^{i'}_{\max} - G^{i'}(m_i,n_i)}{G^{i'}_{max} - G^{i'}_{avg}} \tag{10.2}$$

and use this as a measure of the uncertainty of where new viewers will look. $G^{i'}(m,n)$ denotes a GDF that has been generated by all gaze points except that for viewer i; $G^{i'}_{max}$ and $G^{i'}_{avg}$ denote the maximum and average of $G^{i'}(m,n)$, respectively. Consequently, S equals zero when all viewers gaze toward exactly the same position. In this case the likelihood that a new viewer will look elsewhere is low. The opposite is true when S approaches one; then it is difficult to make qualified predictions of where new viewers will look, and foveation may have to be omitted to ensure a reliable, high subjective quality. The uncertainty is accounted for by computing a scaled σ_s

$$\sigma_s = f(\sigma, S), \sigma_s \geq \sigma \tag{10.3}$$

and use this parameter to generate a new, scaled GDF $G^s(m,n)$. Notice that σ_s does not directly shape Gaussian functions contingent on the viewing setup (visual angle, etc.), but instead reflects and compensates for the uncertainty in ROI location.

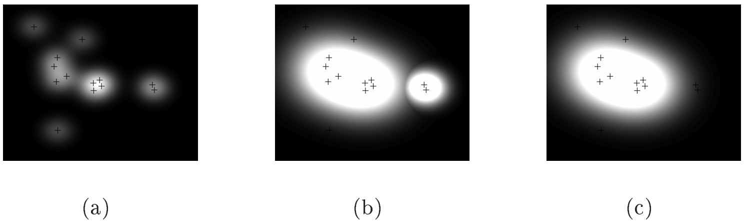
Using the scaled GDFs, positions and shapes of the ROIs are defined for each frame. We present a hierarchical approach to ROI selection, which finds ROIs in order of decreasing saliency and prioritizes regions with high gaze density. Below we describe the mapping from a set of gaze positions $\mathcal{X}$ to the function $G^s(m,n)$; it represents ROI pixels by unit values, and non-ROI pixels are represented by values less than unity with Gaussian-type fall-off toward the ROI edges. To emphasize that gaze points are processed frame-wise, we borrow terminology from video compression by referring to $G^s(m,n)$ as the *intraframe ROI function*.

At the first hierarchical level ℓ_1, a GDF generated from all gaze points $\mathcal{X}$ in a frame is cropped at half its maximum height[1]. Each gaze point is classified as significant or insignificant depending on whether it is located within or outside an active area, and also labeled according to which active area it belongs. For example, if n active areas are found, the gaze points in $\mathcal{X}$ are divided into the subsets $\{\mathcal{X}^{(1)}_{\ell_1}, \mathcal{X}^{(2)}_{\ell_1}, \ldots, \mathcal{X}^{(n)}_{\ell_1}, \mathcal{X}_{\ell_2}\}$, where the subset $\mathcal{X}_{\ell_2}$ contains all gaze points outside of the active areas. Additionally, the subsets are sorted in order of decreasing saliency, where saliency is defined by the number of gaze points contained in a subset. Classification into significant and insignificant gaze points continues in the same manner at the next hierarchical level ℓ_2, but now with $\mathcal{X} \leftarrow \mathcal{X}_{\ell_2}$. The classification algorithm can run until all gaze points are allocated to different hierarchical subsets,

$$\{\mathcal{X}^{(1)}_{\ell_1}, \mathcal{X}^{(2)}_{\ell_1}, \ldots, \mathcal{X}^{(n)}_{\ell_1}, \mathcal{X}^{(1)}_{\ell_2}, \mathcal{X}^{(2)}_{\ell_2}, \ldots, \mathcal{X}^{(m)}_{\ell_2}, \mathcal{X}^{(1)}_{\ell_3}, \ldots\}$$

or until gaze points no longer indicate interesting frame regions. The stop criterion can ultimately be left as a user option.

[1] From this point on, we assume that all GDFs are generated with a scaled sigma, σ_s, as defined by Eq. (10.3)



(a) (b) (c)

Figure 10.1: A GDF (a) and the corresponding intraframe ROI functions before (b) and after (c) removal of temporal outliers. Gaze positions (one for each tested subject) are represented by crosshairs.

Once the significant clusters of gaze points have been identified, each subset ($\mathcal{Y}$) of gaze points is used to generate a new GDF $G_{\mathcal{Y}}(x,y)$, which is cut off at half its maximum height and normalized to unit height. All such cropped and normalized GDF are then combined into the intraframe ROI function

$$\begin{aligned} G^s(m,n) \quad &= \max_{m,n} \quad \{G_{\mathcal{X}_{\ell_1}^{(1)}}(m,n), G_{\mathcal{X}_{\ell_1}^{(2)}}(m,n), \ldots, G_{\mathcal{X}_{\ell_1}^{(n)}}(m,n), \\ & \qquad G_{\mathcal{X}_{\ell_2}^{(1)}}(m,n), G_{\mathcal{X}_{\ell_2}^{(2)}}(m,n), \ldots, G_{\mathcal{X}_{\ell_2}^{(m)}}(m,n), \\ & \qquad G_{\mathcal{X}_{\ell_3}^{(1)}}(m,n), \ldots\} \qquad (10.4) \\ & \qquad m = \{1,2,\ldots,M\}, n = \{1,2,\ldots,N\} \end{aligned}$$

Simulations with our data have shown that four or more ROIs rarely emerge in $G^s(m,n)$. Instead, mostly one and sometimes two and three ROIs account for viewers' visual interest. Figure 10.1(a-b) show the relationship between a GDF and the corresponding intraframe ROI function, which was generated assuming a viewing distance $d = 0.75$m and $\alpha = 5$ degrees. The function $f(\cdot)$ was empirically defined as $f(\sigma, S) = \sigma \cdot (1+2S)$ in order to fulfill the criteria that ROIs should cover the whole display area in case of a spread out gaze point distribution. Only active areas containing two or more gaze points were considered as "interesting". In the coming sections of this chapter, we will use these parameters in our simulations.

The method for clustering gaze-points into hierarchical subsets described above differs from most other clustering techniques. First, it makes no assumptions about the number of clusters (ROIs). Second, the cluster formation is driven by GDFs, naturally taking into account the spatial coherence between different points by modeling the resolution falloff by Gaussian functions. Moreover, it takes the uncertainty of where interesting frame regions reside into account by introducing a measure of gaze-position dispersion. Finally, the shapes of the ROIs are decided automatically.

The hierarchical search for cluster formations (and ROIs) can be applied to other types of data. However, unless the data comprises point of gaze coordinates, it is unclear how to choose and motivate σ.

Step (C)

Even though the detected intraframe ROIs make perfect sense when looking at the gaze point distributions frame-wise, an ROI can be temporally extraneous if it lacks neighboring ROIs adjacent in time. What appears to be a distinct formation of gaze positions in one frame can instead be eye-movements from different subjects briefly overlapping each other in time. This must be accounted for when extending the ROIs into 3-D *volumes of interest* (VOIs).

We let a new VOI appear only if it remains long enough for a viewer of the off-line foveated video to plan and execute a saccade (∼200 ms) to that particular region and to dwell for a typical fixation duration (∼300 ms). Therefore, only VOIs emerging and remaining for more than 500 ms are considered. In practice, this is implemented by finding the centroid of each ROI in the current frame and making sure that temporally adjacent ROIs exist at the same spatial location(s) for $\geq$ 500 ms. Temporally extraneous gaze points are identified as those contained inside of an ROI not fulfilling the above criteria. Remaining gaze points are used to generate a new intraframe ROI function $\tilde{G}^s(m,n)$, which is depicted in Figure 10.1(c). It is generated from the same distribution of gaze points as the GDF in Figure 10.1(a). Notice how the rightmost active area is excluded since it does not fulfill the temporal criteria above.

Step (D)

In the final step, we define an *interframe ROI function*, $G^t_j(m,n)$ for frame j by convolving a number of temporally adjacent intraframe ROI functions by a one-dimensional Gaussian kernel $\boldsymbol{\phi}$:

$$G^t_j(m,n) = \sum_k \phi_k \tilde{G}^s_{j-k}(m,n) \tag{10.5}$$

where $\sum_k \phi_k = 1$.

Temporal smoothing varies contingent on the length and variance of the convolution kernel. Figure 10.2 illustrates 29 adjacent interframe ROIs with a kernel length of 29 pixels and the variance set to 20 pixels. We define a volume of interest (VOI) as a collection of interframe ROIs.

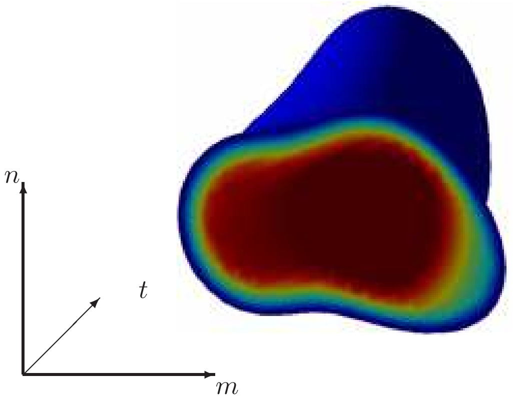


Figure 10.2: A VOI visualization.

10.2 Using VOIs in Video Compression

To maintain a pleasant viewing experience, previous work on off-line foveation and its effects on subjective quality and eye-movements emphasized the importance of implementing smooth variations in quality, both spatially and temporally (Stelmach & Tam, 1994; Duchowski & McCormick, 1998; Nyström et al., 2004). We approached this recommendation by deriving volumes of interest (VOIs) from gaze positions collected by previewers. In this section, we will use the VOIs to manipulate and compress video frames such that quality changes contingent on the VOI-shapes. An overview of the proposed system is schematically depicted in Figure 10.3. As can be seen from the figure, the video is processed such that each frame is off-line foveated in the wavelet domain before being fed to an H.264 encoder. At the decoder side, the bit stream is directly decoded. Since off-line foveation is generated independently of the video coder, no modifications of the H.264 implementation are required. In fact, H.264 can be replaced by any other video coder.

10.2.1 Implementing wavelet foveation

Early techniques for *real-time* degradation (foveation) of the image quality away from the position of gaze either increased the pixel-size in the periphery (Kortum & Geisler, 1996) or used multi-resolution pyramids (Geisler & Perry, 1998). The shape of the foveation mask was derived from experimental measurements of contrast sensitivity. More recently, wavelets have become popular to implement image foveation (Chang & Yap, 1997; Duchowski & McCormick, 1998; Wang & Bovik, 2001). If an observer's position of gaze and viewing distance from the screen are known, wavelet subbands can be weighted such that visually redundant (high-frequency)

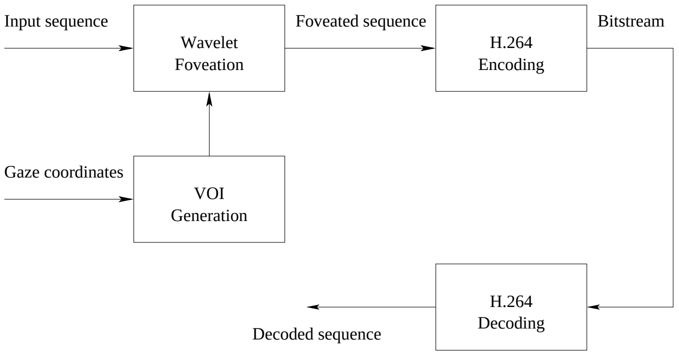


Figure 10.3: Overview of the proposed compression system for off-line foveated video coding.

information is removed from the peripheral regions in the reconstructed image (Wang & Bovik, 2001). Implementing *off-line* foveation requires different strategies for a number of reasons. Most importantly, gaze positions of viewers watching the off-line foveated images/videos are not known exactly. Furthermore, the viewing distances and screen parameters (size, resolution) are not known and can differ between observers. Therefore, there are no straightforward methods either to find the shape of the ROI function or the mapping from an ROI function in the spatial domain to the wavelet domain. Below, we address these issues.

Interframe ROI functions define the visual saliency for different frame regions in the spatial domain. To generate similar interframe ROI functions in the wavelet domain we need a slightly different strategy, both in order to smoothly degrade the display resolution away from the ROIs and also to preserve the low frequency subbands in the wavelet decomposition where most of the energy resides. For the first level ($\lambda = 1$) in the wavelet decomposition, we use the intraframe ROI function $\tilde{G}_j^s(m,n)$, generated from gaze positions where temporal outliers have been removed. At each of the subsequent levels in the wavelet decomposition, σ_j in Eq. (10.3) is increased as $\sigma_j \leftarrow \sigma_j \lambda \beta$ when creating the intraframe ROI function at level λ. β denotes a scaling factor controlling the amount of peripheral blurring. As with the intraframe ROI functions in the spatial domain, their wavelet adjusted counterparts are as a last step smoothed with the same kernel as in Eq. (10.5). Figure 10.4(d) shows an interframe ROI function adjusted to the wavelet domain when four levels of decomposition are used and Figure 10.4(e) illustrates a frame that has been foveated by multiplying its wavelet decomposition with the mask in Figure 10.4(d). For wavelet filtering, we used the bi-orthogonal 9/7 filter (Cohen, Daubechies, & Feauveau, 1992) and periodic border extension. When using color images, each color component (R,G and B) is foveated

	Quality factor					
Video	Lowest	Low	Medium	High	Highest	M±SD
Alte	0.13	0.29	0.52	0.45	0.33	0.34±0.15
Dolphin	0.19	0.27	0.34	0.35	0.29	0.29±0.06
Fish	0.14	0.24	0.32	0.29	0.21	0.24±0.07
Aikyo	0.02	0.01	0.01	0.06	0.16	0.05±0.06
Football	0.13	0.15	0.16	0.16	0.19	0.16±0.02
Hall	0.03	0.03	0.06	0.20	0.18	0.10±0.08
all	0.11	0.16	0.24	0.25	0.23	**0.20±0.13**

Table 10.1: Bitrate gain due to off-line foveation before video encoding with H.264 for different quality factors. Results are presented for the six video clips in Figure 10.5.

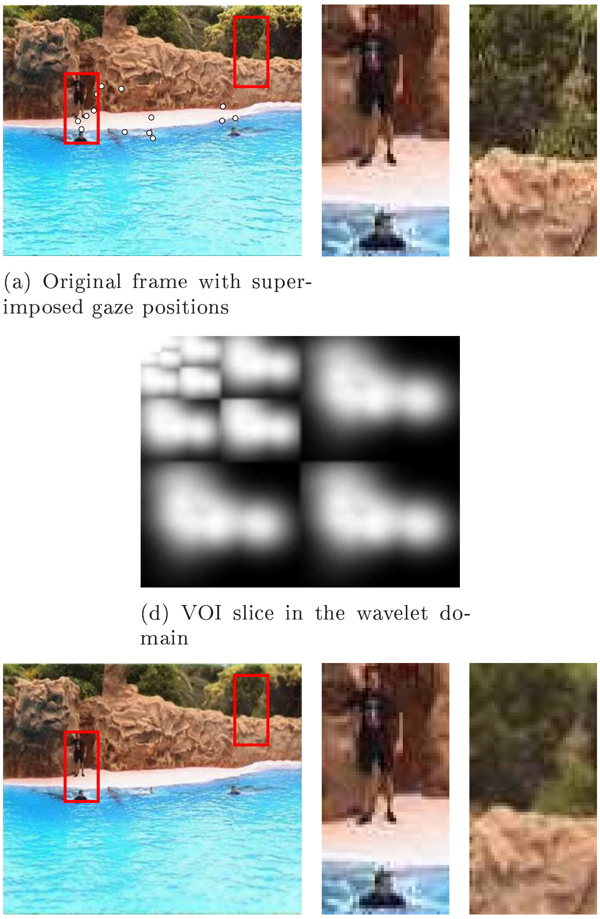
using the same method. Through pilot testing we found that $\beta = 2.3$ introduced a level of peripheral blurring that, when looking at regions of high gaze density, was very hard to notice.

10.2.2 Compression gain due to off-line foveation

Using the above method, we computed the compression gain due to off-line foveation on six video clips. A representative frame from each video is shown in Figure 10.5. The three videos in the upper row in Figure 10.5 were eight seconds long with resolution 720×576 and those in the bottom row 352×288 pixels (CIF format) and of durations five, three and four seconds, counting from the left. Eye-movements had been collected from these videos as described in (Nyström et al., 2004; Johannesson, 2005). Each of the videos was encoded before and after off-line foveation using H.264 (Quicktime 7.3 Pro. implementation) at five different quality settings: Lowest, Low, Medium, High and Highest. Table 10.1 summarizes the results where the bitrate gain due to foveation is defined as

$$\text{Gain} = \frac{\text{FileSize}_{\text{Unfoveated}} - \text{FileSize}_{\text{Off-line foveated}}}{\text{FileSize}_{\text{Unfoveated}}} \qquad (10.6)$$

The table reveals that off-line foveation decreases the file size by 20% on average. However, the variations are large. Videos containing much high frequency content in regions where people do not look can be reduced by as much as 52%. In contrast, off-line foveation barely contributes to additional compression when the background is static and out of focus, as in 'Aikyo'.



(a) Original frame with superimposed gaze positions

(d) VOI slice in the wavelet domain

(e) Off-line foveated frame

Figure 10.4: *Implementing off-line foveation.* The wavelet representation of each frame is multiplied by a VOI slice such as the one illustrated in Figure 10.4(d). Figure 10.4(e) shows the same frame after off-line foveation. To more clearly visualize the difference in quality between the attended and unattended regions, boxed parts of the original and foveated frames are zoomed in.

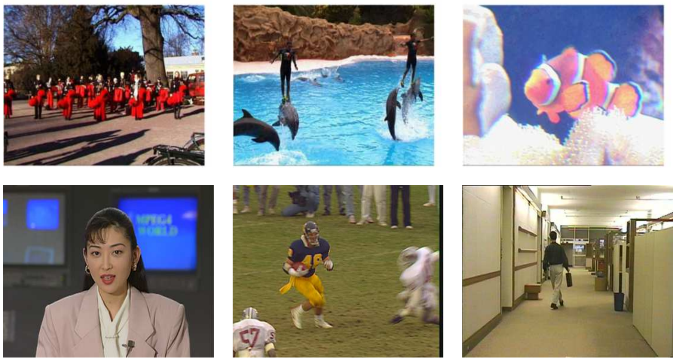


Figure 10.5: Representative frame from each of the tested video clips Alte, Dolphin, Fish, Aikyo, Football and Hall.

10.3 Subjective Evaluations

Off-line foveation clearly reduces the number of bits needed to represent a video digitally. Of course, the reductions are of no value unless the subjective quality remains high. In the remainder of this chapter, we will present a number of new methods to assess how off-line foveation affects subjective quality and gazing behavior. Results from three subjective evaluations are presented. In *Evaluation I*, we let subjects compare the quality of unfoveated and off-line foveated videos compressed with the same quality factor. *Evaluations II and III* extend the methodology used in the first evaluation; we use, for example, eye-tracking data collected during different task instructions and over repeated viewings to obtain direct and indirect measurements of how viewers perceive the off-line foveated videos.

10.3.1 Evaluation I

Subjects and Video material

To investigate how viewers experience the quality of off-line foveated video clips, we let 12 subjects (five women, 28.4±6.3 (M±SD)) watch one unfoveated and one off-line foveated version of three different, eight second video clips. All subjects had normal or corrected-to-normal vision.

As stimuli, we use the videos depicted on the first row in Figure 10.5. These videos are all shorter parts of the videos used in last chapter, and were chosen to depict different types of scenes; one with several people moving around in the display, another with a few main objects of interest, and the last containing one main object of interest. Before being presented to the subjects, both versions of all three clips were compressed with the

+	3	A *much better* than B
+	2	A *better* than B
+	1	A *slightly better* than B
	0	The same
-	1	A *slightly worse* than B
-	2	A *worse* than B
-	3	A *much worse* than B

Table 10.2: Scale for quality ratings.

Quicktime 7.3 Pro H.264 encoder with the quality factor set to 'medium'. Since the objective video quality of the unfoveated and off-line foveated videos is the same within the VOIs before encoding, it is essentially the same also after compression. However, variations can occur along the VOI boundaries.

Procedure

Subjects were instructed that they would be watching three different eight second video clips, each compressed by two different algorithms in an AB trial. A and B denoted either the *unfoveated and compressed* or the *off-line foveated and compressed* version of the same video clip, and were presented one by one in full screen.

After each viewing, subjects were asked to evaluate the video quality of A relative to B according to the quality ratings in Table 10.2.

In order to see the effects of multiple viewings on off-line foveated video quality, subjects were presented to each of the three video clips another two times (ABAB). After all three viewings, most subjects felt that they had a clear picture of the difference in video quality between A and B. If not, they could watch the clips again until they felt confident of giving an accurate vote. Only two of the subjects used this option. Subjects were not informed in advance about the possibility to assess the videos additional times.

The reason for allowing multiple viewing was twofold. First, as in traditional methods for quality evaluations, subjects are given additional viewings to get a clearer picture of the difference in quality. Second, since the videos are stored in subjects' memory after the initial viewing, it is likely that increasing top-down knowledge affects viewing behavior such that gaze positions between the first and later viewings are located at slightly different video regions. To get an indication of whether this occurred, subjects were asked to estimate their viewing behavior during the trials on a scale reaching from 5 (*I was actively searching for quality impairments*) down to 1 (*Just like I watch video at home; my natural viewing behavior*). Although people may be quite poor at estimating

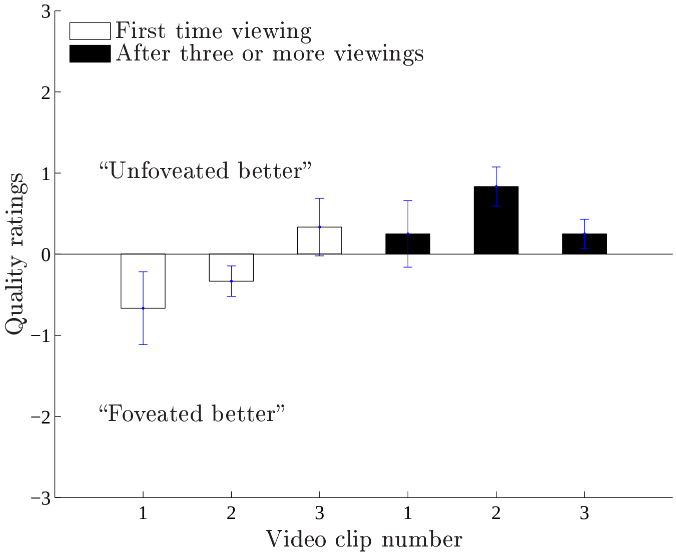


Figure 10.6: Results of the subjective quality evaluation. The x-axis shows which of the three video clips that was tested. Given on the y-axis is the difference (according to Table 10.2) in subjective quality between the unfoveated video and the off-line foveated video. A value larger than zero means that subjects preferred the unfoveated video quality. Bars span one standard error around the mean.

where they look relative to where they actually look as measured by an eye-tracker, we believe that this will give some valuable insight regarding the connection between subjective quality and viewing behavior.

Before a session started, subjects were informed about the quality rating scales, carefully introduced to the testing methodology and also guided through a test session. Subjects were free to ask questions if anything was unclear. None of the subjects was familiar with off-line foveated compression. All data was gathered, to the extent it was possible, under the same conditions as when the eye-movements were collected. The presentation order of the different video clips and the order of the unfoveated and off-line foveated versions were randomized. Hence, six different constellations of the video clips were used.

Results

Figure 10.6 shows the results from the subjective quality evaluation where subjects were asked to compare an unfoveated and an off-line foveated version of the same video after compression with H.264 (medium quality factor). The first three columns show the average quality votes for each of the three tested video clips after one AB viewing, whereas the three rightmost columns show similar votes after two (or more) additional AB viewings required for the subject to feel confident about the judgment. A positive value of the quality vote means that the subject preferred the quality of the unfoveated over the off-line foveated version, while a negative vote means the opposite. Error bars span one standard error around the mean.

For the first two tested clips, we see the rather surprising effect that subjects judge the off-line foveated video quality as better. Similar findings have been reported for real-time gaze-contingent, multi-resolution still images (L. Loschky, McConkie, Yang, & Miller, 2001). However, as in this paper, the effects were not significant. Overall, no significant effects on the difference in video quality were found except for the second video after multiple viewings, where the unfoveated version received slightly better ratings.

After completing the evaluations, subjects were asked to estimate their viewing behavior during the quality evaluations on a scale $\{5, 4, 3, 2, 1\}$, where 5 implied that a viewer actively was searching for video quality impairments while a 1 reflected a viewer's natural viewing pattern while watching video. The average value for the answers was 3.17 with a standard deviation of 0.71. Noticeable, however, was that most of the subjects mentioned that during the first AB trial, their viewing pattern was close to a 1, whereas later in the tests more toward a 5. This suggests that a quality evaluation task does not alter the viewing pattern of subjects from their normal, task neutral viewing pattern, at least not during first time viewing of previously unknown video material. This argument is further strengthened by the observation that peripheral degradations in the off-line foveated videos were difficult to detect during first time viewing as shown by the quality votes. This indicates that viewers indeed looked at the regions of high resolution.

10.3.2 Evaluation II

In order to investigate how off-line foveation changes the gazing behavior during free-viewing, we measure how eye-movements are affected in terms of spatial and temporal distribution in addition to repeated viewings. Without explicitly asking subjects for their subjective opinion about the video quality, the collected gaze data will help us understand how off-line foveated videos are perceived during task neutral, "normal" viewing conditions. The measures we compute in this evaluation will then be compared to those from Evaluation III, where subjects view the same videos while evaluating the subjective quality.

Subjects and video material

15 naive subjects (nine women) of ages 30.2±16.1 (M ± SD) years volunteered to take part in the experiment. They all had normal or corrected-to-normal vision. Stimuli consisted of six original video clips shown in Figure 10.5 and six off-line foveated versions of these, thus 12 videos in total. The three videos in the upper row in Figure 10.5 were eight seconds long with resolution 720×576 and those in the bottom row 352×288 (CIF format) pixels and of durations five, three and four seconds, counting from the left. All videos were displayed in color at 25 fps and compressed with H.264 (in Quicktime 7.3 Pro.) at high bit rates (quality factor 'High') such to no compression artifacts were visible to the bare eye. No sound was used.

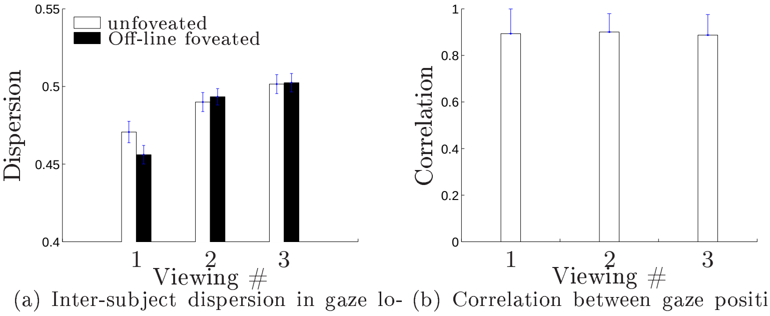
Procedure

Subjects were asked to view the stimuli as they normally would. To prevent subjects from trying to guess the purpose of the experiment, they were told that the study would investigate mental workload by measuring the pupil size. This way, attention was drawn away from the fact that gaze positions were recorded. Subjects were further informed that the same video clip could occur more than once during one presentation.

Each subject was placed at a viewing distance of 76.5 cm in front of a 19 inch computer screen with resolution 1280×1024 and update rate 75 Hz. The active screen area subtended a visual angle of 28 degrees horizontally and 23 degrees vertically. A chin rest was used to restrict head movements.

Prior to each recording, a 13-point spatial calibration was performed. During data recording, all 12 videos were presented one after the other on the screen, separated in time by a mid gray image displayed for one second. Videos were displayed in full screen while maintaining their aspect ratio. No prefixation cross was used to restrict subjects' initial gaze position. The order was randomized with the restriction that two versions (unfoveated and off-line foveated) of the same video could not be displayed directly after each other. To see how repeated viewing affects eye-movement behavior, all 12 videos were presented twice more in the same manner. In total, each video was viewed three times by each subject.

Eye-movements were recorded monocularly with an SMI iViewX Hi-Speed eye-tracker, sampling gaze positions at 240 Hz with position accuracy 0.2°. On average, 9.6 gaze coordinates were recorded for each displayed frame. A Matlab script was developed to collect data about the subjects, communicate with the eye-tracker, display the videos in Quicktime player and control the accuracy in timing during the experiments.



(a) Inter-subject dispersion in gaze locations. Error bars span on standard error in each direction around the mean.

(b) Correlation between gaze positions collected from the unfoveated and off-line foveated versions of the same videos. Error bars span one standard deviation.

Figure 10.7: Eye-movement behavior during free-viewing before and after off-line foveation.

Analysis and results

The perceptual effects of off-line foveation toward video are assessed by comparing gaze positions of viewers watching the tested videos before and after off-line foveation. More precisely, we measure how off-line foveation influences inter-subject dispersion, i.e., how well (or poorly) viewers' gaze positions coincide. This is done both for the initial and later viewings. The inter-subject dispersion, S_t at time t is calculated according to Eq. (10.2). When generating the GDFs in this equation, σ equals 10% of the horizontal video dimension, i.e., $\sigma = 0.10M$ pixels. We tested slightly different parameter values, and the all gave largely similar results.

Figure 10.7(a) illustrates the inter-subject dispersion after one, two and tree viewings of the unfoveated (white bars) and off-line foveated (black bars) videos. It can be seen that off-line foveation has no or little effect on the inter-subject dispersion. However, during first time viewing, there is a tendency ($p = 0.10$, two-sample t-test) that the dispersion decreases due to off-line foveation. Arguably, this effect is present since subjects avoid the blurred regions in the off-line foveated videos such that gaze positions cluster in the high quality regions. Another clear effect is that the dispersion increases significantly after repeated viewings, both for unfoveated and off-line foveated videos. This type of behavior is little surprising since additional viewings encourage more individual viewing strategies, which are likely to reflect an increase in top-down control originating, for example, from memory effects.

To estimate the similarity between two sets of gaze positions $\mathcal{A}$ and $\mathcal{B}$

at time t, we compute the correlation coefficient

$$\rho = \frac{\sum_{x,y}(G_t^{\mathcal{A}}(m,n) - G_{t,avg}^{\mathcal{A}})(G_t^{\mathcal{B}}(m,n) - G_{t,avg}^{\mathcal{B}})}{\sqrt{(\sum_{x,y}(G_t^{\mathcal{A}}(m,n) - G_{t,avg}^{\mathcal{A}})^2)(\sum\sum(G_t^{\mathcal{B}}(m,n) - G_{t,avg}^{\mathcal{B}})^2)}} \qquad (10.7)$$

between the GDFs $G_t^{\mathcal{A}}(m,n)$ and $G_t^{\mathcal{B}}(m,n)$ generated from $\mathcal{A}$ and $\mathcal{B}$, respectively. Figure 10.7(b) shows how gaze positions recorded from viewers watching the unfoveated video correlate with those watching the off-line foveated video after the first, second and third viewing. It can be seen that the correlation is high in all three cases, indicating that subjects' gaze positions have similar distributions.

10.3.3 Evaluation III

It is well known that a task instruction may change where people look (Yarbus, 1967). In off-line foveated video coding, a task that changes viewers' gazing behavior from their 'normal' behavior may have a strong effect on the perceived quality. One that, for example, directs peoples' gazes toward regions unattended by previewers will most certainly decrease the subjective quality. In this section, we will perform subjective quality assessments of off-line foveated video and investigate the effect a *quality evaluation task* has on eye-movements. Moreover, we will quantify how subjects' viewing behavior correlates with their perception of quality. The stimuli and experimental setup are the same as in Evaluation II; procedural changes are explained below.

17 naive subjects (six women) of ages 23.8±4.2 (M ± SD) years were asked to estimate the difference in quality between two versions, A and B, of the same video in an AB trial. They were told that the two versions resulted from different compression algorithms being applied to the original video. To encourage subjects to do their best and maintain focus during the evaluation, they were told that quality assessment is a difficult task and the differences in quality would sometimes be hard to notice. As in Evaluation II, subjects were informed that the study would investigate mental workload during quality assessment by measuring the pupil size. The videos were assessed as follows. Each AB trial started by displaying a uniform mid-gray image with a large, centered black capital A, followed by version A of the stimulus. Directly after A had been shown followed the same procedure for version B. Then a pop-up window containing a slider bar and a button appeared on the screen (see Figure 10.8). On the slider bar, three different levels of quality were given: *A better than B*, *A equal to B*, *B better than A*. Subjects could freely adjust the slider to a position reflecting their experienced quality, and then press the button to continue with the next AB trial. For subsequent data analysis, the slider position was quantized to an integer value between -5 and +5. The pre-

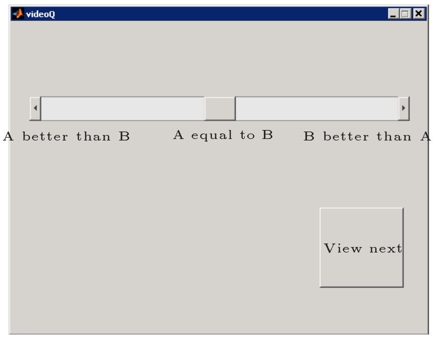


Figure 10.8: Pop-up window for quality assessment.

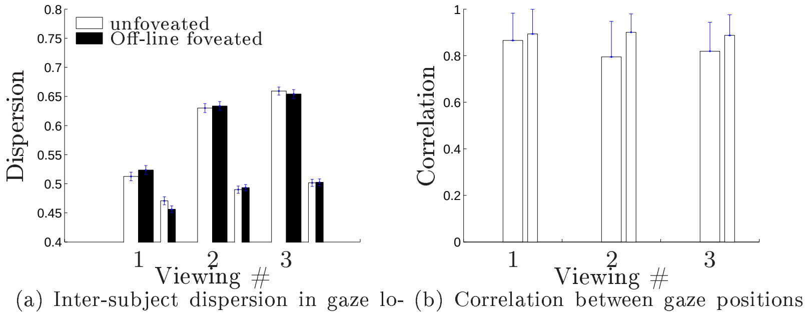
sentation order of the video (AB) pairs was randomized. A and B denoted the original and off-line foveated versions of a video clip.

In standardized methods for quality evaluation, subjects are usually allowed to view the videos to be assessed several times before giving the actual judgment. Therefore, to see the effect a quality evaluation task has on repeated viewings, the above video pairs were shown another two times after which a second quality vote was taken. Subjects did not know in advance that further chances to evaluate the quality would be given.

Results

Figure 10.9 compares the dispersion of, and the correlation between gaze positions collected before and after off-line foveation during first, second and third time viewing. For comparison, similar measures from the second evaluation are given as bars with smaller width. As can be seen from the figure, the results are similar to those from the second evaluation with the difference that the dispersions are significantly larger ($p < 0.01$, two-sample t-test) during quality evaluation. Supposedly, the more active task of quality evaluation encourages individual viewing strategies, and explains why subjects' gaze positions spatially are less similar to each other. During first time viewing, the dispersion during quality assessment is rather close, although significantly different ($p < 0.01$, two-sample t-test), to the baseline value (first time free-viewing), and it can be assumed that subjects look within the non-degraded regions in the off-line foveated video. The further pursue this assumption, Figure 10.10 compares the perceived quality of the six tested video clips before and after off-line foveation and how it is affected by repeated viewings.

The white bars in Figure 10.10 show the average subjective quality of the videos after the first viewing. Error bars extend one standard error. A value larger than zero indicates that subjects prefer the quality



(a) Inter-subject dispersion in gaze locations. Error bars span on standard error in each direction around the mean.

(b) Correlation between gaze positions collected from the unfoveated and off-line foveated versions of the same videos. Error bars span one standard deviation.

Figure 10.9: Eye-movement behavior during quality assessment before and after off-line foveation. For comparison, results from Evaluation II are included in the figure, depicted by the thinner bars.

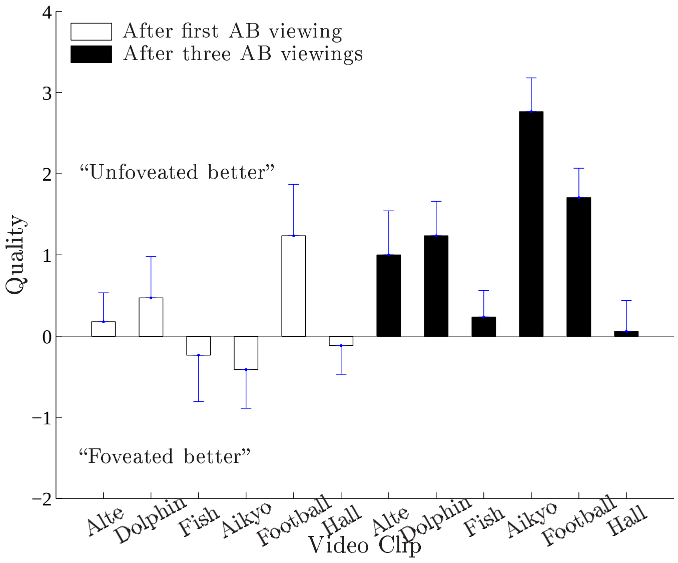


Figure 10.10: Subjective votes reflecting the difference in quality between unfoveated and off-line foveated versions of the same video clip. A value below zeros indicates that the quality of the off-line foveated video was judged as higher whereas the opposite is true for values larger than zeros. Error bars span one standard error.

of the unfoveated video whereas the opposite is true for values below zero. Off-line foveation resulted in decreased quality in one of the tested videos, `Football`. The reason for this is most likely that eye-data used to implement off-line foveation was slightly inaccurate temporally, such that foveation was performed with a slight lag in time. It is therefore no surprise that the video containing the fastest movements gets a lowered subjective quality. The rest of the off-line foveated videos were essentially indistinguishable from the unfoveated videos in terms of subjective quality. However, as a result of repeated viewings subjects changed their viewing pattern and gazed directly at degraded parts of the off-line foveated videos. The consequence of repeated viewings in terms of subjective quality is illustrated by the black bars in Figure 10.10, where subjects strongly prefer the quality of the unfoveated versions. An interesting observation is the large change in subjective quality between the first and later viewings of `Aikyo`. Most likely, the facial region is such a strong visual attractor that it is initially hard to not gaze at. However, when looking outside the facial region, which happens after repeated viewings, it is rather easy to see the introduced blurring effects.

10.4 Summary

The work in this chapter extends our initial approach to off-line foveation and its applicability in compression. We have proposed a mapping from gaze positions into volumes of interest (VOI), which are use to implement off-line foveation in video. VOI based off-line foveation prior to compression decreased the bitrate significantly. In disagreement with previous works, off-line foveation neither decreased the subjective quality nor did it change the eye-movement behavior.

Chapter 11

Discussion of Part II

IN contrast to known methods for real-time foveated video coding, we propose that *off-line foveation* can be used for improved video compression. Using gaze positions collected from a number of previewers, off-line foveation is implemented by reducing the quality in regions where few or none of the previewers look. Such quality reductions can give rise to significant bit rate reductions when combined with traditional methods for compression.

In this part of the thesis we reviewed previous techniques for foveated coding, investigated the rationales behind off-line foveation, and implemented and evaluated systems for off-line foveation. The highlights of our results show that:

- Viewers' gaze positions coincide when looking at video.
- Off-line foveation prior to compression reduces the bitrate with up to 50% compared to compressing the same, unfoveated video.

Contrary to previous work (Stelmach & Tam, 1994; Duchowski & McCormick, 1998), we report that:

- The bitrate gain is achieved without decreasing the subjective quality.
- During initial free-viewing of a video, off-line foveation has little effect on subjects' eye-movement behavior.

We conclude that off-line video foveation combined with compression can indeed be successful to increase the efficiency of today's state-of-the-art methods. On the videos we tested, the average additional compression gain due to off-line foveation was 20%. There are some reasons why this number should be regarded a lower bound. First, the methods we used for compression are in no way optimized to encode off-line foveated videos. Using methods that better take advantage of the properties of an off-line foveated image sequence will yield even better compression gains. For example, improvements can include coding the motion vectors with unequal importance, such that fewer bits are given to motion vectors in non-attended regions. Second, the degree of peripheral blurring is experimentally tuned, and it is therefore not clear how much additional blurring could be introduced without degrading the subjective quality. In real-world situations, the optimal amount of blurring depends on factors that can only be approximated, such as the screen size, screen resolution, and viewing distance of an observer. Third, it can be seen from Table 10.1 that off-line foveation is less beneficial when the video quality is poor. In this case, foveation could probably have been increased further. Finally, the tested videos were assessed in a lab environment and presented without sound. Using a more engaging viewing setup, it is likely that the coherence between subjects' gaze positions would increase even further. In addition to yielding large bitrate gains, off-line foveation allows for complexity reductions where computational resources can be focused toward high quality regions.

Clearly, bitrate reductions due to off-line foveation would be of little interest without conserving subjective quality. We estimated the quality by quantifying changes in gazing behavior between unfoveated and off-line foveated videos and by performing modified versions of standard tests for subjective quality assessment. Moreover, we calculated the effects these measures had over repeated viewings. Results show that off-line foveation had no or very slight effects on both gazing behavior and subjective quality during first time viewing. However, both gaze locations and subjective quality were affected as a result of multiple viewings. As we expected, the results also showed that traditional methods for video quality assessment were not directly applicable to off-line foveated video. In standards outlined in, e.g., (VQEG, 2003), it is advised to show the video to be assessed several times to the subject before a quality vote is taken. However, as seen by our results, repeated viewings make subjects' gazing behavior deviate from normal, first time free-viewing, thus shifting visual attention toward regions where viewers normally would not look. In view of this, traditional methods would all give very poor results in judged quality for off-line foveated videos (as was found in Stelmach & Tam, 1994). To our knowledge, these issues have not been considered in standard quality assessment using 'normal', unfoveated stimuli. It is therefore not clear how multiple viewings affect subjective quality in these cases. Since gazing behavior changes over multiple viewings, and therefore increasingly more deviates from typical 'free-viewing', do standardized methods for video quality assessment produce results that reflect 'typical' viewing? This is indeed an interesting question for future research.

Part III

General Discussion and Conclusions

Chapter 12

Conclusions and Outlook

WE found off-line foveation prior to compression to decrease the video bitrate without neither decreasing the subjective quality nor changing subjects' eye-movement behavior. Investigating the prerequisites for using low-level algorithmic gaze prediction, instead of eye-tracking, for the purpose of off-line foveation gave few promising answers; using contrast manipulated still images, we showed that low-level features such as contrast and edge density can easily be overridden by higher cognitive factors, both early after image onset and later in viewing.

Today, there are some practical issues making it cumbersome to effectively utilize off-line foveated systems for video compression. The one met with most skepticism is that eye-tracking recordings require expensive equipment and are time consuming, and therefore would be a bottleneck in a real-world application. We see two future solutions to this problem. First, it is by many envisioned that eye-trackers will be embedded in web cameras, and that other low cost, simple-to-use eye-tracking equipment will be available for practical use in a near future. Already today, such systems have been suggested and implemented (Hansen, MacKay, Hansen, & Nielsen, 2004; Pedersen & Spivey, 2006). This would make eye-tracking recordings more autonomous and less time consuming, since individual viewers themselves could download videos and record gaze positions through self-paced experiments. Since eye-movements would not have to be measured in real-time, the lack of technical sophistication a webcamera offers compared to a state-of-the-art eye-tracking system can be compensated for by first recording the eye-movements, and then let a high-complexity algorithm calculate gaze positions off-line. One interesting application where webcamera based eye-tracking could have a huge impact is streaming video over the Internet. For example, around 13 hours of video are uploaded to YouTube every minute, and the estimated daily cost of bandwidth utilization for YouTube is approximately $1 million (Wikipedia, 2008c). Consequently, off-line foveated compression could save millions of dollars every year or provide better video quality for the same cost. Second, there is no doubt that off-line foveation greatly would benefit from algorithms that automatically and accurately predict where subjects will look, given only the raw video as input. Such algorithms would increase the practical usability of off-line foveation for video coding since eye-tracking collections with human observers would be unnecessary. Since dynamic features such as motion and flicker seem to attract attention more robustly than static features (Itti, 2005), models including a dynamic feature channel appear even more promising to account for human eye-movements.

To this date, there have been some implementations aiming to predict human gaze positions in dynamic scenes (Osberger & Rohaly, 2001; Böhme, Dorr, Krausea, Martinetz, & Barth, 2006; Le Meur et al., 2007). A few of these directly target foveated video compression applications (Wang, Sheikh, & Bovik, 2003; Itti, 2004; Agrafiotis et al., 2006). Wang, Sheikh, and Bovik (2003) use the heuristic rule of always choosing face regions as foveation points and, to minimize the prediction error, foveation points are also positioned where the residual error is large. Agrafiotis et al. (2006) exploit off-line foveation to optimize the quality of video coded for sign-language; they use eye-tracking to measure where people look during sign-language comprehension, and code the videos according to where the people looked. The only method using a general purpose algorithm (without a specific application in mind) at the gaze prediction stage is the one by Itti (2004), and even though he showed that a substantial amount of compression can be obtained by using this algorithm to foveate an image sequence, it was left as future research to measure whether it changes viewers' subjective quality and eye-movement behavior. A recent abstract offers some empirical support that the subjective quality remains high also after foveated compression (Li & Itti, 2008). Overall, however, there is no doubt that several issues still need to be addressed and empirically investigated regarding bottom-up algorithms for gaze prediction in both static and dynamic scenes.

Using contrast manipulated images we showed some limitations of bottom-up predictors. In particular, the two current state-of-the-art algorithms we tested were far from robust in finding fixations comparable to those found by human viewers. To improve algorithms based on these principles, a trend in current research is to endow purely bottom-up models with top-down knowledge (e.g., Navalpakkam & Itti, 2005; Torralba, Oliva, Castelhano, & Henderson, 2006; Cerf, Harel, Einhäuser, & Koch, 2007). The model by Navalpakkam and Itti (2005) provides keywords describing a search target and uses prior, learned information about the features of this target to bias the search. Torralba et al. (2006) extend a bottom-up algorithm by feeding it with contextual information. An example could be to inform the algorithm searching for pedestrians to look only at the sidewalk, and not in the sky (where cloud edges could introduce peaks in a saliency map). Searching for people in real world photographs, another suggested top-down modification simply adds a face detector to the bottom-up predictor (Cerf et al., 2007). Although this type of additional knowledge can improve the performance of a predictor under certain conditions and well defined tasks, it is still an open question whether (and what type of) top-down knowledge improves the performance during a free-viewing task.

A recent study found the central bias inherent in video viewing to account for eye-movements better than a state-of-the-art model for gaze prediction (Le Meur et al., 2007). Given the strong influence on both top-down factors and systematic tendencies (such as the central bias) in video viewing, it seems very optimistic to believe that bottom-up driven algorithms can completely account for human eye-movements during free-viewing, and therefore be successful for the purpose of off-line foveated video coding. On the good side, we know that semantically informative regions generally coincide with peaks in bottom-up saliency (Henderson et al., 2007), and that saliency often is biased toward the center of the display. As a consequence, a bottom-up algorithm has the potential to find locations fixated by human viewers, even though the raw video features do not causally contribute to gaze selection. From this optimistic point of view, thus, a bottom-up algorithm may at times provide gaze predictions accurate enough to enable successful off-line foveated compression. A severe limitation is that, sooner or later, the prediction a bottom-up algorithm makes will deviate from the positions attended by humans. In terms of subjective video quality this deviation is likely to affect the quality negatively since the frames with the poorest quality dominantly decide the overall video quality (compare with packet losses) (Liu, Wang, Boyce, Wu, & Yang, 2007). However, using a moderate degree of foveation it is possible that these predictive errors may pass unnoticed by the viewers.

When the problems of accurately predicting foveation points or gaze densities are solved, we predict that off-line foveation will be an interesting technology for future applications in video communications. In particular, is would be beneficial in bandwidth constrained applications such as wireless communications in mobile devices, and for video streamed over Internet. For example, prioritizing regions with high gaze density can be useful to facilitate interpretation, recognition and subjective quality of image and video data, especially at low bit rates.

To resolve the question how image features are related to fixation locations we believe that, using a fixed task instruction, a range of features must be manipulated using an experimental paradigm similar the one used in this thesis. By systematically reducing and increasing features like contrast, color, and luminance in a scene, we are more likely to elicit the causal contribution for each of these features. Our results show that when studying gaze control in images, the choice of stimuli is crucial. Obviously, a gaze prediction algorithm trained on images with neutral semantics may perform poorly when tested on images containing objects with high semantic importance, which we know can override bottom-up features cognitively.

Part IV

Appendix

Appendix A

Wavelet Transform

An image is commonly described in either the time or frequency domain, where each representation emphasizes different information about the image. For the purpose of analysis, it would be desirable to have a representation that simultaneously describes the image in both time and frequency; this is where *wavelets* come in. For an introduction to wavelets and their application to image coding, see e.g., Antoni, Barlaud, Mathieu, and Daubechies (1992) and Sayood (2000).

Wavelets are mathematical functions that are generated from scaled and translated versions of a single function ψ

$$\psi^{a,b}(t) = |a|^{-1/2}\psi(\frac{t-b}{a}) \tag{A.1}$$

ψ is usually called the mother wavelet. The wavelet transform $W\{f(t)\}$ of a signal $f(t)$ can then be described by a superposition of wavelets

$$W\{f(t)\} = \int_{-\infty}^{\infty} \overline{\psi^{a,b}(t)} f(t) dt. \tag{A.2}$$

In practical implementations, wavelets are defined by discrete filters, and the *discrete wavelet transform* (DWT) takes an input signal and passes it through these filters to create a wavelet based representation. Figure A.1 illustrates a 1-level wavelet decomposition of an image. The image is initially passed through either of two 1-dimensional filters: h_0 and h_1. The former filter is of lowpass nature and the latter of highpass nature. Initially, the filters operate in the vertical direction, and filtering is followed by downsampling by a factor of two in the same direction as the filter operated. These filtering and downsampling procedures are repeated in the horizontal direction and result in four wavelet subbands: Lowpass-Lowpass (LL), Lowpass-Highpass (LH), Highpass-Lowpass (HL),

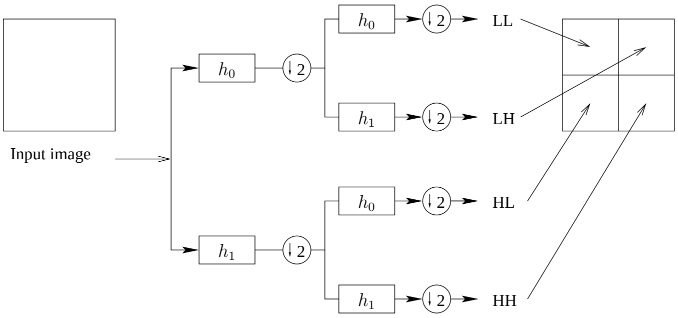


Figure A.1: Analysis filter bank.

and Highpass-Highpass (HH). The LL subband is a downscaled version of the original image, whereas other subbands contain more detailed image information. Together, the subbands can fully reconstruct the original image by reversing the above operations: upsampling is followed by filtering with a new set of filters, $\tilde{h}_0$ and $\tilde{h}_1$. To ensure a perfect reconstruction of the image, the filters used for decomposition and reconstruction must fulfill the requirements of quadrature mirror system.

Figure A.2 depicts a 1-level wavelet decomposition of the `lena` image using the Daubechies 4 tap family of filters.

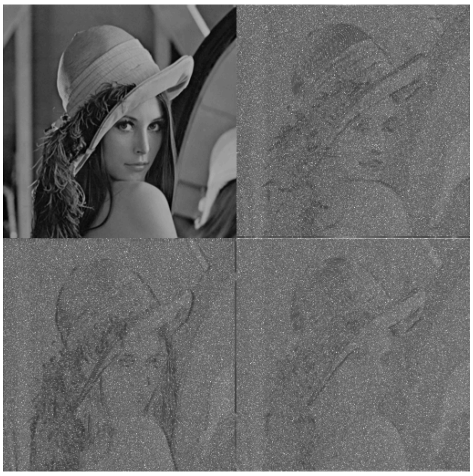


Figure A.2: A one level wavelet decomposition. Wavelet coefficients are logarithmically enhanced for display purpose.

Bibliography

Agrafiotis, D., Canagarajah, N., Bull, D., Kyle, J., Seers, H., & Dye, M. (2006). A perceptually optimised video coding system for sign language communication at low bit rates. *Signal Process. Image Commun.*, *21*(7), 531-549.

Antes, J. (1974). The time course of picture viewing. *J. Exp. Psychol.*, *103*(1), 62-70.

Antoni, M., Barlaud, M., Mathieu, P., & Daubechies, I. (1992). Image coding using wavelet transform. *IEEE Trans. Image Processing*, *1*(2), 205-220.

Baddely, R., & Tatler, B. (2006). High frequency edges (but not contrast) predict where we fixate: A bayesian system identification analysis. *Vision Research*, *46*, 2824-2833.

Bergström, P. (2003). *Eye-movement controlled image coding.* Doctoral dissertation, Linköping University, Sweden.

Böhme, M., Dorr, M., Krausea, K., Martinetz, T., & Barth, E. (2006). Eye movement predictions on natural videos. *Neurocomputing*, *69*, 1996-2004.

Brainard, D. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433-436.

Buswell, G. (1935). *How people look at pictures.* University of Chicago Press, Chicago.

Cerf, M., Harel, J., Einhäuser, W., & Koch, C. (2007). Predicting human gaze using low-level saliency combined with face detection. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems 120* (p. 241-248). MIT Press, Cambridge, MA.

Chang, E., & Yap, C. (1997). A wavelet approach to foveating image. In *13th symposium on computational geometry* (p. 397-399).

Chua, H. F., Boland, J. E., & Nisbett, R. E. (2005). Cultural variation in eye movements during scene perception. *Proc. of the National Academy of Sciences*, *102*(35), 12629-12633.

Cohen, A., Daubechies, I., & Feauveau, J. (1992). Biorthogonal bases of compactly supported wavelets. *Commun. Pure Appl. Math*, *45*(5), 485-560.

Cover, T., & Thomas, J. (1991). *Elements of information theory.* Wiley, New York.

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, *36*(12), 1827-1837.

Dorr, M., Böhme, M., Drewes, J., Gegenfurtner, K. R., & Barth, E. (2005). Variability of eye movements on high-resolution natural videos. In *8th Tübinger perception conference* (p. 162). Tübinger, Germany.

Duchowski, A. (2000). Acuity matching resolution degradation through wavelet coefficient scaling. *IEEE Trans. Image Processing*, *9*(8), 1437-1440.

Duchowski, A. (2002). A breadth-first survey of eye tracking applications. *Behavior Research Methods, Instruments, and Computers*, *34*(4), 455-470.

Duchowski, A. (2003). *Eye tracking methodology: Theory and practice.* Springer-Verlag, New York.

Duchowski, A., & McCormick, B. (1998). Gaze-contingent video resolution degradation. In B. Rogowitz & T. Pappas (Eds.), *Human vision and electronic imaging* (Vol. 3299, p. 318-329).

Einhäuser, W., & König, P. (2003). Does luminance-contrast contribute to a saliency map for overt visual attention? *European J. Neuroscience*, *17*(5), 1089-1097.

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

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision*, *8*(3), 1-15.

Elias, G. S., Sherwin, G. W., & Wise, J. A. (1984). *Eye movements while viewing ntsc format television.* (SMPTE Psychophysics Subcommittee white paper)

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, *7*(1), 1-29.

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision*, *8*(2), 1-17.

Gall, D. L. (1991). Mpeg: a video compression standard for multimedia applications. *Commun. ACM*, *34*(4), 46-58.

Geisler, W., & Perry, J. (1998). A real-time foveated multi-resolution system for low-bandwidth video communication. In B. Rogowitz & T. Pappas (Eds.), *Human vision and electronic imaging* (Vol. 3299, p. 294-305).

Geisler, W., & Perry, J. (1999). Variable-resolution displays for visual communication and simulation. *Society for Information Display*, *30*(1), 420-423.

Girod, B. (1988). Eye movements and coding of video sequences. In *Visual comm. and image processing* (Vol. 1001, p. 398-405). Cambridge, MA: SPIE.

Goldstein, R. B., Peli, E., Lerner, S., & Luo, G. (2004). Eye movements while watching video: comparisons across viewer groups. *Journal of Vision*, *4*(8), 643.

Hanley, J., & McNeil, B. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29-36.

Hansen, D., MacKay, D., Hansen, J., & Nielsen, M. (2004). Eye tracking off the shelf. In *Eye tracking research & application* (p. 58). San Antonio, Texas: ACM New York.

Haskell, B., & Netravali, A. (1995). *Digital pictures - representation, compression and standards* (Second ed.). Plenum, New York.

Henderson, J. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, *7*(11), 498-504.

Henderson, J. (2007). Regarding scenes. *Current Directions in Psychological Science*, *16*(4), 219-222.

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

Henderson, J., & Ferreira, F. (2004). Scene perception for psycholinguists. In J. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (p. 1-58). New York: Psychology Press.

Holmqvist, K. (2009). *Eye-tracking methods (unpublished book).*

Huffman, D. (1952). A method for the construction of minimum-redundancy codes. *Proc. Inst. Radio Eng.*, *40*(9), 1098-1101.

Itti, L. (2004). Automatic foveation for video compression using a neurobiological model for visual attention. *IEEE Trans. Image Processing*, *13*(10), 1304-1318.

Itti, L. (2005). Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, *12*(6), 1093-1123.

Itti, L. (2006). Quantitative modeling of perceptual salience at human eye position. *Visual Cognition*, *14*(4-8), 959-984.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*, 1489-1506.

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

ITU. (2002). *Rec. ITU-R BT.500-11. Methodology for the subjective assessment of the quality of television pictures.* Geneva, Switzerland.

Johannesson, E. (2005). *An eye-tracking based approach to prediction of gaze behavior using low level features.* Master's thesis, Lund University.

Johansson, R., Holsanova, J., & Holmqvist, K. (2006). Pictures and spoken descriptions elicit similar eye movements during mental imagery, both in light and in complete darkness. *Cognitive Science*, *30*(6), 1053-1079.

Juday, R., & Fisher, T. (1989). Geometric transformations for video compression and human teleoperator display. In *Optical pattern recognition* (Vol. 1053, p. 116-123). SPIE.

Kelly, D. (1962). Information capacity of a single retinal channel. *IRE Trans. Information Theory*, *8*(3), 221-226.

Khan, J., & Komogortsev, O. (2006). Hybrid scheme for perceptual object window design with joint scene analysis and eye-gaze tracking for media encoding based on perceptual attention. *J. Electron. Imaging*, *15*(023018).

Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D. (2002). Defining and quantifying the social phenotype in autism. *Am. J. Psychiatry*, *159*(6), 895-908.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Hum. Neurobiol.*, *4*(4), 219-227.

Kortum, P., & Geisler, W. (1996). Implementation of a foveated image coding system for image bandwidth reduction. In B. Rogowitz & J. Allebach (Eds.), *Human vision and electronic imaging* (Vol. 2657, p. 350-360). San Jose, CA: SPIE.

Kuyel, T., Geisler, W., & Ghosh, J. (1999). Retinally reconstructed images: Digital images having a resolution match with the human eye. *IEEE Trans. Systems, Man, and Cybernetics - Part A: Systems and Humans*, *29*(2), 235-243.

Land, M. (2007). Fixation strategies during active behaviour. a brief history. In R. van Gompel, M. Fischer, W. Murray, & R. Hill (Eds.), *Eye movements: a window on mind and brain* (p. 76-98). Oxford: Elsevier.

Land, M. F., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research*, *41*, 3559-3565.

Law, B., M. Atkins, M., Kirkpatrick, A., Lomax, A., & Mackenzie, C.

(2004). Eye gaze patterns differentiate novice and experts in a virtual laparoscopic surgery training environment. In *Eye tracking research & application* (p. 41 - 48). San Antonio, Texas: ACM, New York.

Le Meur, O., Le Callet, P., & Barba, D. (2007). Predicting visual fixations on video based on low-level visual features. *Vision Research*, *46*, 2483-2498.

Levine, M. (2000). *Fundamentals of sensation and perception* (Third ed.). Oxford University Press Inc., New York.

Li, Z., & Itti, L. (2008, May). Visual attention guided video compression. In *Vision science society annual meeting.* Naples, FL.

Lipps, M., & Pelz, J. B. (2004). Yarbus revisited: task-dependent oculomotor behavior. *Journal of Vision*, *4*(8), 115.

Liu, T., Wang, Y., Boyce, J., Wu, Z., & Yang, H. (2007). Subjective quality evaluation of decoded video in the presence of packet losses. In *IEEE conf. acoustics, speech and signal proc.* (Vol. 1, p. 1125-1128). Honolulu, Hawaii.

Loftus, G., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *J. Exp. Psychol: Human Perception and Performance*, *4*(4), 565-572.

Loschky, L., McConkie, G., Yang, J., & Miller, M. (2001). Perceptual effects of a gaze-contingent multi-resolution display based on a model of visual sensitivity. In *Advanced displays and interactive displays fifth annual symposium* (p. 53-58). College Park, MD.

Loschky, L. C., & Wolverton, G. S. (2007). How late can you update gaze-contingent multiresolutional displays without detection? *ACM Trans. Multimedia Comput. Commun. Appl.*, *3*(4), 1-10.

Mannan, S., Ruddock, K., & Wooding, D. (1995). Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-d images. *Spatial Vision*, *9*(3), 363-386.

Mannan, S., Ruddock, K., & Wooding, D. (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision*, *10*(3), 165-188.

Martinez-Conde, S., Macknik, S., & Hubel, D. (2004). The role of fixational eye movements in visual perception. *Nature Reviews Neuroscience*, *5*(3), 229-240.

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research*, *45*(2), 205-231.

Nyström, M., & Holmqvist, K. (2007a). Deriving and evaluating eye-tracking controlled volumes of interest for variable-resolution video compression. *J. Electron. Imaging*, *16*(013006).

Nyström, M., & Holmqvist, K. (2007b). Variable resolution images and their effects on eye-movements. In B. Rogowitz, T. Pappas,

& S. Daly (Eds.), *Human vision and electronic imaging.* San Jose, CA: SPIE.

Nyström, M., & Holmqvist, K. (2008). *Effect of compressed off-line foveated video on viewing behavior and subjective quality.* (Submitted to the ACM Trans. on Multimedia Computing, Communications, and Applications)

Nyström, M., & Holmqvist, K. (2008, in press). Semantic override of low-level features in image viewing – both initially and overall. *Journal of Eye Movement Research.*

Nyström, M., Novak, M., & Holmqvist, K. (2004). A novel approach to image coding using off-line foveation controlled by multiple eye-tracking measurements. In *Picture coding symposium.* San Francisco.

O'Driscoll, G. A., Lenzenweger, M. F., & Holzman, P. (1998). Antisaccades and smooth pursuit eye tracking and schizotypy. *Arch. Gen. Psychiatry*, *55*(9), 837-843.

Oliva, A., & Schyns, P. (1997). Coarse blobs or fine edges? evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology*, *34*(1), 72-107.

Oliva, A., & Schyns, P. (2000). Colored diagnostic blobs mediate scen recognition. *Cognitive Psychology*, *41*, 171-210.

Osberger, W., & Maeder, A. (1998). Automatic identification of perceptually important regions in an image. In *Pattern recognition, fourteenth international conf.* (Vol. 1, p. 701-704). Brisbane, Australia.

Osberger, W., & Rohaly, A. (2001). Automatic detection of regions of interest in complex video sequences. In B. Rogowitz & P. T. (Eds.), *Human vision and electronic imaging* (Vol. 4299, p. 361-372). San Jose, CA: SPIE.

Osterberg, G. (1935). Topography of the layer of rods and cones in the human retina. *Acta Ophthalmologica*, *6*, 1-102.

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

Parkhurst, D., & Niebur, E. (2002). Variable-resolution displays: A theoretical, practical, and behavioral evaluation. *Human Factors*, *44*(4), 611-629.

Parkhurst, D., & Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision*, *16*(2), 125-154.

Parkhurst, D., & Niebur, E. (2004). Texture contrast attracts overt visual attention in natural scenes. *European J. Neuroscience*, *19*(3), 783-789.

Pedersen, B., & Spivey, M. (2006). Offline tracking of eyes and more with a simple webcam. In *28th annual meeting of the cognitive science* *society.* Vancouver, Canada.

Posner, M. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*(1), 3-25.

Privitera, C., & Stark, L. (2000). Algorithms for defining visual regions of interest: Comparison with eye fixations. *IEEE Trans. Pattern Analysis and Machine Intelligence*, *22*(9), 970-982.

Rajashekar, U., Cormack, L., & Bovik, A. (2004). Point of gaze analysis reveals visual search strategies. In *Human vision and electronic imaging* (Vol. 5292, p. 296-306). San Jose, CA.

Rajashekar, U., Linde, I. van der, Bovik, A., & Cormack, L. (2007). Foveated analysis of image features at fixations. *Vision Research*, *47*(25), 3160-3172.

Rajashekar, U., Linde, I. van der, Bovik, A., & Cormack, L. (2008). Gaffe: A gaze-attentive fixation finding engine. *IEEE Trans. Image Processing*, *17*(4), 564-573.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422.

Rayner, K., & Castelhano, M. (2007). Eye-movements. scholarpedia. *Scholarpedia*, *2*(10). (Revision No. 20895)

Reinagel, P., & Zador, A. (1999). Natural scene statistics at the centre of gaze. *Computation in Neural Systems*, *10*(4), 341-350.

Rissanen, J. (1976). Generalized kraft inequality and arithmetic coding. *IBM J. Res. Develop.*, *20*(3), 198-203.

Roepstorff, A., & Frith, C. (2004). What's at the top in the top-down control of action? script-sharing and 'top-top' control of action in cognitive experiments. *Psychological Research*, *68*, 189-198.

Rothkopf, C., Ballard, D., & Hayhoe, M. M. (2007). Task and context determine where you look. *Journal of Vision*, *7*(14), 1-20.

Rupp, H., & Wallen, K. (2007). Sex differences in viewing sexual stimuli: An eye-tracking study in men and women. *Hormones and Behavior*, *51*(4), 524-533.

Sayood, K. (2000). *Introduction to data compression* (Second ed.). New York: Morgan Kaufmann.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.*, *27*, 379-423, 623-656.

Sheikh, H. R., Liu, S., Evans, B. L., & Bovik, A. C. (2001). Real-time foveation techniques for h.263 video encoding in software. In *Proc. IEEE int. conf. acoustics, speech, and signal proc.* (Vol. 3, p. 1781-1784).

Stelmach, L. B., & Tam, W. (1994). Processing image sequences based on eye movements. In *Human vision, visual processing and digital display* (Vol. 2179, p. 90-98). San Jose, CA.

Stelmach, L. B., Tam, W., & Hearty, P. (1991). Static and dynamic spatial resolution in image coding: An investigation of eye-movements.

In *Human vision, visual processing and digital display* (Vol. 1453, p. 147-152). San Jose, CA.

Tatler, B. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*(14), 1-17.

Tatler, B., Baddeley, R., & Gilchrist, I. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Research*, *45*(5), 643 - 659.

Tatler, B., Baddeley, R., & Vincent, B. (2006). The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. *Vision Research*, *46*(12), 1857-1862.

Tatler, B., & Vincent, B. (2008, in press). Systematic tendencies in scene viewing. *Journal of Eye Movement Research.*

Taubman, D. S., & Marcellin, M. W. (2001). *JPEG 2000: Image compression fundamentals, standards and practice.* Norwell, MA: Kluwer Academic Publishers.

Torralba, A., Oliva, A., Castelhano, M., & Henderson, J. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*(4), 766-786.

Tosi, V., Mecacci, L., & Pasquali, E. (1997). Scanning eye movements made when viewing film: preliminary observations. *Intern. J. Neuroscience*, *92*(1-2), 47-52.

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

Tseng, P. H., Carmi, R., Cameron, I. G. M., & Munoz, L., D.P. Itti. (2007). The impact of content-independent mechanisms on guiding attention. *Journal of Vision*, *7*(9), 633.

Underwood, G., Foulsham, T., Loon, E. van, Humphreys, L., & Bloyce, J. (2006). Eye movements during scene inspection: A test of the saliency map hypothesis. *European J. Cognitive Psychology*, *18*(3), 321-342.

VQEG. (2003). *Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II.*

Wallace, G. K. (1992). The JPEG still picture compression standard. *IEEE Trans. on Consumer Electronics*, *38*(1), 30-44.

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*, *19*(9), 1395-1407.

Wang, Z., & Bovik, A. (2001). Embedded foveation image coding. *IEEE Trans. Image Processing*, *10*(10), 1397-1410.

Wang, Z., Lu, L., & Bovik, A. (2003). Foveation scalable video coding with automatic fixation selection. *IEEE Trans. Image Processing*, *12*(2), 243-254.

Wang, Z., Sheikh, H., & Bovik, A. (2003, Sep.). Objective video quality assessment. In B. Furht & O. Marqure (Eds.), *The handbook of video databases: Design and applications* (p. 1041-1078). CRC Press.

Welch, T. (1984). A technique for high-performance data compression. *Computer*, *17*(6), 8-19.

Wiegand, T., Sullivan, G., Bjontegaard, G., & Luthra, A. (2003). Overview of the h. 264/avc video coding standard. *IEEE Trans. Circuits and Systems for Video Technology*, *13*(7), 560- 576.

Wikipedia. (2008a, 09:42 on 20 August). Available from `http://en.wikipedia.org/wiki/Image:Gray722.png`

Wikipedia. (2008b, 09:42 on 20 August). Available from `http://en.wikipedia.org/wiki/Image:Human_eye_cross-sectional_view_grayscale.png`

Wikipedia. (2008c, 09:42 on 20 August; 232962274). Available from `http://en.wikipedia.org/wiki/YouTube`

Wolfe, J. (1998). Visual search: A review. In H. Pashler (Ed.), *Attention.* London: University College London Press.

Wooding, D. (2002). Fixation maps: quantifying eye-movement traces. In *Eye tracking research & applications* (p. 31-36). New Orleans, LA: ACM Press, New York.

Yarbus, A. (1967). *Eye movements and vision.* Plenum Press, New York.