From the Institute of Neuro- and Bioinformatics
of the University of Lübeck
Director: Prof. Dr. rer. nat. Thomas Martinetz

Methods for the prediction and guidance of human gaze

Dissertation
for Fulfillment of
Requirements
for the Doctoral Degree
of the University of Lübeck

from the Department of Computer Sciences/Engineering

Submitted by

Eleonora Vig
from Cluj-Napoca

Lübeck 2011

First referee: Prof. Dr.-Ing. Erhardt Barth
Second referee: Prof. Dr. rer. nat. Ulrich G. Hofmann

Date of oral examination: 15.07.2011

Approved for printing. Lübeck, 15.08.2011

Contents

Acknowledgements

The quest to learn ultimate transformation rules that are applied to elements of multidimensional spaces made up a substantial part of my PhD research. In a broader sense, the work towards a PhD also is a “multidimensional” endeavour, a personal transformation process that is not limited to scientific progress only. Erhardt Barth, my advisor, has been the most influential person in shaping and supporting this change. I sincerely thank him for being a constant source of ideas and guidance.

Without Michael Dorr, my colleague in the GazeCom project, this journey would have never been the same. Thank you for the fruitful discussions, numerous readings of my drafts, continuous support, and much more!

With Laura Pomârjanschi, who is a fellow countryman and good friend, our research paths often intersected. She taught me never to be afraid to venture into unexplored subspaces.

I thank Thomas Martinetz for offering me the opportunity to pursue a PhD at the Institute for Neuro- and Bioinformatics, and I am very grateful for the generous financial support the Institute provided for my German classes.

Ein besonderer Dank geht auch an Ebba-Maria Dudde für ihren unermüdlichen Einsatz und ihre Begeisterung, mit der sie mir all diese Jahre Deutsch beigebracht hat.

Many thanks for everyone at the INB for making graduate school so enjoyable. I am also extremely grateful for the continuous love and encouragement of my family and friends.

During my PhD, I have received financial support from the European Commission within the project GazeCom (contract no. IST-C-033816) of the 6th Framework Programme.

Abstract

Our brain's capacities to process visual information are limited, and we therefore make two to four eye movements per second to direct the only high-resolution area of the retina, the fovea, to a few relevant, so-called *salient* locations in the visual input. The mechanism that permits the selection of behaviourally relevant parts of the visual field is called *visual attention*. Apart from its importance in biological vision, attention is a fundamental tool also for artificial vision systems. Several biologically-inspired machine vision systems rely already on the selective processing of visual information to achieve higher efficiency. Recently, the redirection of visual attention to certain goal-relevant areas in the visual field has been recognized as a promising new strategy to integrate into future visual and communication systems.

This dissertation explores techniques and algorithms that allow to "guide" the human gaze, by embedding subtle image-based changes in the visual stimulus that result in an alteration of the gaze patterns. The aim of gaze guidance is to augment human vision with computer vision technology in a least-obtrusive way. Gaze guidance is realized by gaze-contingent interactive displays that use an eye tracker to monitor the viewer's gaze. Based on the visual input and the gaze position of the viewer, first, a limited set of salient, candidate locations is predicted that would attract the user's gaze. Then, based on the desired scanpath, one candidate is selected and its saliency is increased, while all other candidates are simultaneously decreased in saliency. With respect to the outlined strategy, the contribution of this thesis is threefold. In a gaze guiding scenario, the right timing of the so-called gaze-capturing events is critical for achieving the desired guiding effect. Therefore, first, we characterize various visual stimuli with respect to the typical saccadic response lags to salient events. Second, we develop a powerful saliency model to predict potential salient candidates in videos. Finally, we put forth a generic saliency modification framework to alter the saliency distribution of the scene.

In particular, we start by quantifying the typical response time of attention shifts in truly natural movies, which is revealed to differ significantly from that in quasi-realistic scenes such as videos games and TV clips with cuts and camera motion. To this end, we temporally align analytical spatiotemporal saliency maps (encoding salient events in the videos) with an "empirical" saliency measure encoding saccadic reaction to the salient events. To determine the attentional response lag, analytical and empirical saliency are cross-correlated when shifted against each other in time. The time lag at which the cross-correlation function reaches its maximum denotes the average saccadic response delay. The near-zero average lag measured in truly natural scenes is attributable to an adaptation of the human visual system to the (often predictable) dynamics of the environment.

In the main part of this thesis, a novel and generic model of bottom-up saliency is put forth derived from the simple assumption that the degree of local signal variation is related to informativeness (and thus, salience) of an image or video region. The concept of intrinsic dimensionality measures this degree and yields a basic description of how a multidimensional signal may change. Machine learning techniques act on simple image representations of videos derived from efficient coding principles to distill the properties that distinguish "interesting" video regions. Thanks to its generic nature, our model offers a unified framework for incorporating space-, time-, and colour information, which usually are treated separately. The proposed model predicts eye movements on a diverse collection of videos with high accuracy, and because it does not suffer from overfitting as many more complex models do, it also outperforms several existing models in the prediction of saccade landing points.

Finally, a generic saliency modification scheme is proposed, in which, first, the structural differences between attended and non-attended video locations are learnt. The information on the class boundary that separates the two classes is then used to derive the desired image transformations that lead to an alteration in saliency. Transformations performed in the low-dimensional space of the spectral energy are implemented as local contrast manipulation rules on a spatiotemporal Laplacian pyramid. We show with empirical measurements that this scheme is successful in both changing the saliency distribution of scenes and in guiding the eyes, whereby we deliver a proof of concept for gaze-guiding systems.

Zusammenfassung

Die Bandbreite und Informationsverarbeitungskapazität des menschlichen visuellen Systems sind begrenzt. Aus diesem Grund bewegen wir unsere Augen typischerweise drei- bis viermal pro Sekunde, um das hochaufgelöste Zentrum der Netzhaut, die Fovea, auf wenige wichtige, sogenannte *saliente* Regionen der visuellen Szene zu lenken. Die Auswahl dieser verhaltensrelevanten Regionen geschieht dabei durch den Prozess der *visuellen Aufmerksamkeit*, der nicht nur für biologische Systeme wichtig ist, sondern auch in technischen Systemen Anwendung findet. Zum einen gibt es bereits biologieinspirierte Algorithmen, die eine Bandbreitenreduktion durch Beschränkung auf nur saliente Bildteile erreichen, zum anderen können zukünftige Informations- und Kommunikationssysteme davon profitieren, ein Modell der Aufmerksamkeit des Benutzers zu erstellen.

In dieser Arbeit entwickeln wir Methoden, um den Blickpfad und damit die Aufmerksamkeit eines Betrachters zu lenken. Ein Ziel ist dabei, menschliches und technisches Sehen möglichst effizient zu kombinieren. Blicksteuerung wird mithilfe von Systemen zur Blickrichtungsmessung und durch blickrichtungsabhängige Displays realisiert: Auf Grundlage der jeweils aktuellen Blickposition und des Stimulus wird in einem ersten Schritt eine Liste von Kandidatenpunkten vorhergesagt, die wahrscheinlich als nächstes fixiert werden. Dann wird das Video in Echtzeit so verändert, dass der der gewünschten Blickrichtung nächste Kandidatenpunkt in seiner Salienz verstärkt wird, wohingegen die übrigen, ablenkenden Kandidatenpunkte abgeschwächt werden. Zur Verwirklichung dieser Strategie trägt die vorliegende Arbeit drei wesentliche Elemente bei. Ein Element ist dabei die optimale zeitliche Platzierung der Videomodifikationen zur Blicklenkung. Hierzu analysieren wir verschiedene Kategorien visueller Stimuli im Hinblick auf den typischen zeitlichen Versatz zwischen dynamischen salienten Ereignissen und den darauffolgenden Augenbewegungen. Ein weiterer Aspekt ist die Entwicklung eines leistungsfähigen Salienzmodells, um mit Methoden des maschinellen Lernens Augenbewegungen auf natürlichen Videos vorherzusagen.

Schließlich resultiert diese Arbeit in einem generischen neuartigen Ansatz, um Transformationen der Salienzverteilung für natürliche Szenen abzuleiten.

Im Einzelnen beginnen wir mit der Untersuchung der typischen Latenzen von Aufmerksamkeitssprüngen in natürlichen Videos, wobei sich zeigt, dass sich das Blickverhalten auf solchen Videos qualitativ von dem auf weniger natürlichen Szenen wie Computerspielen oder TV-Clips mit Kamerabewegungen unterscheidet. Für diese Untersuchung berechnen wir die Kreuzkorrelation von analytischen Salienzkarten, die orts-zeitlich saliente Ereignisse kodieren, mit "empirischen" Salienzkarten, die aus Augenbewegungsdaten gewonnen wurden. Das Maximum der Kreuzkorrelationsfunktion bezeichnet dabei die mittlere Latenz von Sakkaden und liegt für natürliche Videos ungefähr bei 0 Millisekunden. Dieses überraschende Ergebnis lässt sich damit erklären, dass die Dynamik der natürlichen Umgebung für das (evolutionär optimal angepasste) visuelle System oftmals vorhersagbar ist.

Den Hauptteil der vorliegenden Arbeit macht die Entwicklung und Analyse eines neuen und generischen Modells zur Salienz aus. Die Grundannahme ist dabei, dass das Ausmaß an lokaler Signalveränderung den lokalen Informationsgehalt und damit auch die Salienz bestimmt. Die *intrinsische Dimensionalität* kann zur Berechnung dieses Maßes herangezogen werden und beschreibt ein Alphabet möglicher Veränderungen für multidimensionale Signale. Auf diesen relativ einfachen Signalrepräsentationen, die natürliche Videos effizient kodieren, extrahieren wir mithilfe von Methoden des maschinellen Lernens diejenigen Strukturen, die saliente Bildregionen auszeichnen. In diesem generischen Ansatz können wir Information über Ort, Zeit und verschiedene spektrale Kanäle sowie Skalen integriert verarbeiten, was in konkurrierenden Modellen nur getrennt passiert. Bei einer Auswertung auf einem großen Datensatz von über 50 Probanden und 18 Videos kann unser Modell Augenbewegungen mit hoher Genauigkeit und besser als existierende Modelle aus der Literatur vorhersagen.

Schließlich erarbeiten wir einen neuartigen Ansatz, um aus den gelernten Unterschieden von salienten und nicht-salienten Bildregionen Transformationen abzuleiten, um Regionen algorithmisch von einer Klasse in die andere zu verschieben. Punkte in hochdimensionalen Merkmalsräumen werden nach bestimmten Regeln innerhalb der Mannigfaltigkeit natürlicher Bilder senkrecht zu der separierenden Hyperebene verschoben. Für die praktische Anwendung geeignet zeigen sich Modifikationen der lokalen spektralen Energie auf einer orts-zeitlichen Laplace-Pyramide, die zu Veränderungen des orts-zeitlichen Kontrasts führen. In Experimenten, bei denen die spektrale Energie in Echtzeit auf einem blickrichtungsabhängigen Display verändert wird, können wir einen Effekt auf den Blickpfad von Betrachtern nachweisen.

1

Introduction

1.1 Motivation

Of all the senses, vision is the most dominant, the most developed, and the most complex. It is fundamental to our perception of the world and for interaction with it. In many everyday tasks, such as driving, the eyes provide much of the sensory input. Even though they may seem so natural to us, our perceptual skills are remarkable. We recognize faces and objects with an incredible ease even in highly cluttered scenes and are able to navigate busy highways, find lost keys, and enjoy a pretty landscape. This may suggest that we succeed in these ubiquitous tasks due to their simplicity. However, artificial intelligence has shown that, apart from some highly domain-specific scenarios, to this day, machines and robots have difficulties even remotely approaching human perceptual abilities. The tremendous complexity of the problem is well reflected in the functional organization of our brains. A massive amount of the human brain power is devoted to visual processing and visual perception involves neural computations in many brain areas. Surprisingly, much of this processing focuses on information gathered from only a tiny part of the visual scene that occurs at the centre of our gaze. Only this small high-resolution spot on the retina, called the fovea, provides sharp vision, and visual acuity decreases rapidly towards the periphery. This non-uniform sampling of the visual scene drastically reduces the amount of visual information that must be processed by high-level functions, such as object recognition. Hence, to build up a coherent and detailed representation of the world around us, we move our eyes at frequent intervals. Despite the rapid succession of variable-resolution "still shots" of the visual scene, our visual experience is nevertheless smooth and seamless.

What greatly contributes to this “illusion” is the rapid and appropriate selection of the most relevant scene items to be sampled. To this end, our brains make use of complex *attentional mechanisms* in deciding which parts of the scene deserve further detailed — and hence more resource-consuming — processing. The impressive performance of the human visual system is partly due to such selective processing, and deciphering the underlying mechanisms will help in understanding how the brain accomplishes vision.

In contrast, although the last decades have seen much progress towards building robust and generic *machine vision systems*, existing computer vision approaches still fail to match human visual performance in unconstrained scenarios. Ultimately, the goal of artificial vision systems is to mimic human vision, and thus, inspiration from the neurobiology and from mechanisms involved in vision can prove beneficial not only for practical computer vision, but also for understanding the neural mechanisms underlying perception. As such, the ability to rapidly sort out irrelevant information and restrict the computationally expensive image processing to the potentially relevant scene locations has already proven invaluable for many computer vision applications.

In this dissertation, we shall attempt to unravel some of these visual mechanisms. It was written in the context of the European project “Gaze-based Communication” (or simply GazeCom), which aims at developing state-of-the-art algorithms that enable the unobtrusive *guidance* of human visual attention. GazeCom is an interdisciplinary project on the border between human and computer vision that seeks, on the one hand, to broaden the theoretical understanding of attention and human oculomotor behaviour, and, on the other hand, to augment human vision with computer vision technology in a least-obtrusive way.

As argued above, visual attention is vital to our perception and interaction with the world. However, the way in which humans deploy their attention, e.g. explore a new object or search for specific information, can vary significantly from person to person and depending on the cognitive load of the performed activity. One key factor for this variation is expertise, and it is well accepted that in several domains the viewing behaviour of experts differs considerably from that of novices. For example, experienced drivers' and pilots' gaze patterns are much better defined, and expert radiologists and geologists are more efficient in finding specific patterns in medical or geological images. In other words, experts may possess a better “internal model” of the often-performed activity that enables them to direct their attention more efficiently.

The *gaze-guidance systems* [Barth, 2001, Barth et al., 2006] proposed within the GazeCom project promise to aid the information search of the viewer who does not yet possess such an internal model, by steering the observer’s gaze through a new visual scene in order to enforce a predetermined, optimal viewing pattern. Gaze guidance is realized by *gaze-contingent interactive displays* that use an eye tracker to continuously monitor the viewer’s gaze. In order to achieve an alteration of the gaze patterns, the visual scene is modified in real time by subtle local changes to the visual input. This manipulation assumes the following three steps: i) based on the visual input and the eye position of the viewer, first, a limited set of candidate points is predicted that would attract the user’s attention; ii) using real-time video processing, the probability of being attended is increased for one selected candidate location, and iii) simultaneously decreased for all other candidates. That such modifications are not perceived consciously is assured by the fact that they are embedded gaze-contingently in the periphery.

While perhaps not readily apparent, a fully functional gaze-guidance system cannot be expected as an end result of this thesis. Instead, one must recognize the interdisciplinary dimensions and the complex nature of the problem. Nevertheless, this dissertation will contribute significantly to the realization of such systems by exploring techniques and algorithms that allow the *prediction* and *guidance* of human gaze in naturalistic videos.

1.2 Contributions

With respect to the above outlined gaze-guiding strategy, the contributions of this dissertation are threefold.

First, for step i), we shall develop a generic yet powerful *computational model of attention* to predict potential candidate locations, so-called *salient* points, in natural dynamic scenes. We here attempt to answer the question: what is it about a certain scene area that it automatically draws attention while others do not? Developing *saliency models* has been a longstanding challenge for neuroscientists, although most work has focused on static images. The neuroscience approach of representing biological processes as faithfully as possible, however, has rendered these models highly complex. We shall demonstrate that our model, which is much simpler, predicts eye movements on a diverse collection of naturalistic videos with high accuracy, and because it does not suffer from overfitting as many more complex models do, it also outperforms state-of-the-art models in the prediction of salient candidates.

Second, concerning steps ii) and iii) of the outlined strategy, we shall venture into uncharted territory, by putting forth a generic *saliency modification framework* to manipulate the interestingness of a visual scene. Efficacy in altering the viewing behaviour through appropriate saliency transformations is demonstrated both conceptually and empirically in psychophysical experiments.

Finally, in a gaze guiding scenario, the "right timing" of the embedded image or video manipulations is critical for achieving the desired unconscious guiding effect. Therefore, in this work we shall characterize various visual stimuli with respect to the typical oculomotor response times to attention-capturing scene events.

Apart from their importance for gaze guidance, insights into these questions will advance our theoretical understanding of the underpinnings of human oculomotor behaviour.

1.3 Outline

The organization of the thesis is as follows. Chapter 2 is concerned with the neurophysiological background of human visual perception, emphasizing the selective nature of visual processing. The relevant concepts in the context of visual attention are introduced, and a review of the state of the art in computational modelling of attention is provided. We conclude this chapter with a demonstration of the utility of such models in computer vision and active vision scenarios.

In Chapter 3, we shall outline a few basics on the efficient representation of visual data in biological and artificial systems. In particular, we shall review the concept of intrinsic dimension, which allows a geometric characterization of typical image and video structures. Such an efficient representation of natural stimuli will prove extremely useful throughout the thesis for eye movement modelling. The second part of this chapter is concerned with multiresolution representations, which not only faithfully simulate the variable-resolution processing of early vision, but will also allow us throughout this thesis to efficiently analyse and manipulate high-resolution videos, operations that are indispensable for achieving the desired goals of this thesis.

The following chapters each deal with one of the main contributions of this thesis.

Chapter 4 quantifies the typical response time of attention shifts in truly natural movies, which is revealed to differ significantly from that in quasi-realistic scenes such as video games and TV clips with frequent cuts and camera motion. The temporal component of gaze allocation in naturalistic videos is a less-studied aspect of attentional orienting, and our findings will shed light on the degree of anticipation observed during the free-viewing of real-world videos. As a simple measure of spatiotemporal salient events we shall employ the geometrical framework presented in Chapter 3.

Chapter 5 introduces the computational saliency model for the prediction of eye movements in dynamic scenes. Particular emphasis is placed on the simple and generic nature of the proposed model, and therefore, its complexity is increased gradually. Machine learning techniques act on simple image representations of videos derived from efficient coding principles reviewed in Chapter 3 to distill the properties that distinguish “interesting” video regions. The model is evaluated against several state-of-the-art approaches, and two extensions of the basic model, i) to predict eye movements even on transparently overlaid videos, and ii) for a faster saliency computation for resource-limited systems, are proposed.

Chapter 6 elaborates on the proposed generic saliency modification scheme that is built upon the saliency-learning framework detailed in Chapter 5. Once the structural differences between attention-capturing and non-interesting video regions have been distilled, transformation rules can be derived that manipulate some saliency-relevant properties of video regions. The proposed generic scheme is implemented in practice by considering spatiotemporal contrast manipulations, and is evaluated in terms of its effect in influencing gaze patterns both conceptually and empirically, in a psychophysical study.

Finally, Chapter 7 concludes the work presented in this dissertation and summarizes the contributions. Recommendations for future work are also suggested here.

Some parts of this work are the result of a group effort. In particular, Michael Dorr signs responsible for the efficient software implementation of the pyramidal and geometric video representations reviewed in Chapter 3. Eye movement collection on our natural videos was performed in the lab of Karl Gegenfurtner at the Dept. of Psychology of Giessen University. In Chapter 5, gaze data collection and analysis of eye movement predictability on multiple transparent videos were run by Laura Pomârjanschi.

2

Visual attention in natural and artificial systems

Biological vision is a highly active process. Our brains continuously analyze the visual environment, select relevant, so-called salient areas, and direct our eyes accordingly. Such filtering of the sensory input is necessary because the brain's capacities to process visual information are limited. The mechanism that permits the selection of behaviourally relevant parts of the visual field is called *visual attention*. Understanding and modelling attentional mechanisms has been the topic of extensive research in neurology and psychology, but apart from its importance in biological vision, attention is a fundamental tool also for artificial systems. Biologically-inspired machine vision systems rely on the selective processing of visual information to achieve more efficiency.

In the first part of this chapter, we shall review basic facts about human visual perception, with an emphasis on the *selective nature* of perception. After a short anatomical introduction to the eye, we shall discuss relevant concepts in visual attention, such as the distinction between overt and covert, as well as bottom-up and top-down attention. In the second part, we shall turn to the computational modelling of visual attention and review some of the most influential models that exist in the literature. We conclude this chapter with a brief presentation of successful applications of saliency models in computer vision and robotics.

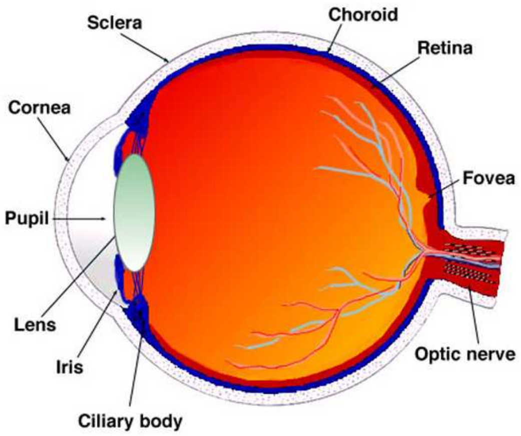


Figure 2.1: A cross-section of the human eye. As light enters the eye, it is refracted by the cornea and the lens, which focus the light onto the retina. Here, light-sensitive receptors convert the light into electrical signals that are carried to the brain via the optic nerve. From [Kolb et al., 2010].

2.1 The human visual system

First, we will give a brief overview of the main components of the human visual system. Far from being an exhaustive account, the aim of this summary is to convince the reader why selective visual attention is crucial in our perception of the world. For a more detailed description of the physiology and neurology of the visual system, we refer to e.g. [Palmer, 1999], [Itti et al., 2005], and [Findlay and Gilchrist, 2003].

2.1.1 Anatomy of the eye

The human eye, depicted schematically in Figure 2.1, is a complex optical system. The *pupil*, the eye's black-looking aperture, allows light reflected from an object to enter the eye through the *cornea* and the *lens*. The lens refracts the incoming light, and focuses it to the back of the eye, projecting an upside-down image onto the *retina*. A light-sensitive tissue, the retina is composed of several different cell layers. The outer nuclear layer consists of *photoreceptors*, i.e. photosensitive cells that convert light into electrical signals. In the inner layer, nerve fibres from the photoreceptors are bundled together at the back of the eye to form the *optic nerve*, which transmits

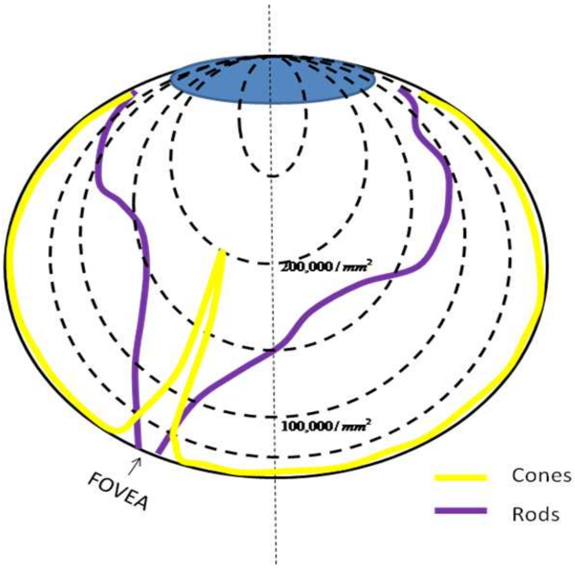


Figure 2.2: Distribution of rods and cones on the retina. Cones, which are distributed towards the centre of the visual field and almost exclusively form the fovea, are used for colour and daylight vision, whereas rods, which are more plentiful in the periphery, serve for brightness and motion perception, and night vision. From Wikimedia Commons (`http://commons.wikimedia.org/wiki/File:Density_rods_n_cones.png`).

the electrical signals to the brain. Photoreceptor cells in the retina can be of two types: *rods* and *cones*. Rods, whose number amounts to about 90 million [Curcio et al., 1990], are extremely sensitive to light, but because they cannot distinguish between wavelengths, they are responsible for achromatic low-light, e.g. nocturnal, vision. Cones, on the other hand, are less numerous (about 5 million [Curcio et al., 1990]), much less sensitive to light, and they provide the eye's colour sensitivity. Cones come in three subtypes, which respond to short (blue), medium (green), and long (red) wavelengths of light and thus allow a distinction between colours. In addition, cones differ from rods in that they can faithfully sample details, whereas rods have low spatial acuity.

As Figure 2.2 shows, rods and cones are not evenly distributed across the retina. Roughly at the centre of the retina lies a small circular region of about two degrees diameter, called the *fovea*, that is densely packed with cone cells and lacks rods almost completely. Despite its small size, it is here that both spatial and colour vision are most accurate. Towards the periphery of the retina, cone cells gradually become sparser. Rods, on the other hand, are concentrated in the outer parts of the retina. As a consequence of this non-uniform cell layout, visual acuity also varies across the retina, i.e. we do not perceive every part of the visual scene with equal sharpness. Thus, the human visual field, which spans almost 200 degrees horizontally, can be divided in three main regions. *Foveal vision* refers to the high-resolution, detailed vision that occurs at the centre of the retina, the fovea. This region constitutes less than one percent of the visual field, but about 50 percent of visual cortex are dedicated to the processing of foveal information. *Parafoveal vision* starts outside the foveal region and spans about 10 degrees of visual angle. Here, spatial resolution drops gradually: the farther away from the fovea, the “blurrier” the scene gets. Beyond the parafovea, visual acuity decreases sharply, i.e. we are sensitive only to coarse visual cues (e.g. no object recognition is possible); however, *peripheral vision* is tuned to detecting changes in the visual scene (e.g. sensing movement). Due to the inhomogeneity of the retina (i.e. the fovea is the only small region of high acuity), the eye must be frequently moving to place the fovea onto the objects of interest and, if the target is moving, track it.

The point on the retina from where the optic nerve emerges is called the *optic disc*. It is characterized by a complete lack of photoreceptors (i.e. it is insensitive to light), therefore, it is also known as the blind spot.

2.1.2 Neural pathways

In the last processing layer of the retina, the long axons of the ganglion cells form the optic nerve, which sends the visual signal (in form of action potentials) towards the processing units of the brain. The optic nerves of the two eyes come together at the *optic chiasm*, where the information from the left half of the visual field is sent to the right half of the brain, while the information from the right visual field is directed to the left side of the brain. This crossed mapping is important for depth perception.

Behind the optic chiasm, the optic nerve is called the *optic tract*. The optic tract carries the visual signal on two separate pathways into subcortical areas. The smaller pathway goes to the *superior colliculus*, a brain area involved in the control of eye movements. The principal projection pathway leads through the *lateral geniculate nucleus* (or LGN) of the thalamus to the higher brain areas, such as the *primary visual cortex* V1. Most of the high-level visual processing, such as object recognition and motion estimation, takes place in the visual cortex.

There are multiple types of nerve cells in the visual pathways. The distinction between *magnocellular* (M) and *parvocellular* (P) cells is important because of their functional differences in visual processing. M cells have a fast response and high contrast gains; hence, they signal the existence of a sudden change. P cells, on the other hand, with their small receptive fields are more suited for signalling details of objects. In the LGN, the two cell types separate in two distinct layers, and it is suggested that they remain separated as two processing streams within the cortex. The *dorsal stream* (or magnocellular pathway) is also known as the "where" stream, as it is involved in recognizing the spatial relationships of objects and in guiding actions. The *ventral stream* (or parvocellular pathway) is concerned with object recognition and form representation, hence the alternative name: "what" pathway. Besides these two principal routes, there are multiple interconnecting pathways between cortical areas in the brain.

With this very brief overview of some basic facts about the anatomy of the human visual system, we will now proceed to a more detailed description of eye movements and attention, which both are critical aspects to this thesis.

2.2 Eye movements and visual attention

2.2.1 Eye movements

As we have seen in the previous section, the parallel processing of visual information in the brain gives rise to an immediate conscious perception. However, this is, in itself, not enough to leave us with the subjective impression of a fully detailed view of the world around us. This illusion is created by our ability to effortlessly move our eyes about two to four times per second to successively sample the scene with the high-resolution fovea, and to integrate this transsaccadic information into one coherent percept.

To explore the environment by selectively processing information at locations of interest, the human visual system has at its disposal a set of oculomotor control processes. The four main types of eye movements and their properties are listed below.

Saccade Saccades are rapid, abrupt movements of the eye that bring new targets of interest to the fovea. Such ballistic movements take about 150–200 ms to plan (time referred to as *saccadic latency*), and have a duration of only 20–80 ms [Becker, 1991]. Depending on the amplitude, saccades can reach speeds up to 900 deg per second. Orienting movements of more than about 30 deg are achieved by a combination of both eye and head movements. Saccades are said to be ballistic in that, once initiated, the trajectory cannot be altered. This has to do with their short duration: 20–80 ms is less than it takes an optical signal to reach the brain regions where eye movements are evoked [Becker, 1991]. Saccades often under- or overshoot an intended target, in which case short corrective saccades are required. During a saccade, the image projected on the retina moves with high velocities. Yet, we are unaware of the motion blur of the image, due to a phenomenon called *saccadic suppression* [Matin, 1974]. To prevent us from being aware of the blurred images, the update of visual information is actively suppressed shortly before, during, and shortly after saccades.

Fixation Between saccades, the eye is held (almost) stationary: it fixates the target of interest for about 150–600 ms so that the visual information at the particular location can be processed by the human visual system, and the next saccade is planned Irwin [1992]. During fixation, the eyes are actually still making small, involuntary movements, called *microsaccades*. They are believed to exist in order to prevent the retinal image from fading, which is provoked by neural adaptation in the retina [Ditchburn and Ginsborg, 1952, Martinez-Conde et al., 2004].

Smooth pursuit The function of pursuit movements is to maintain a moving target stabilized on the fovea, that is: track an object. Unlike saccades, such movements are characterized by a smooth motion of the eye with no abrupt on- and offsets. Pursuit movements are hard to be made in the absence of a moving target. They can reach a maximum velocity of 100 deg per second, but have a low latency of about 100 ms. Contrary to saccadic suppression, visual sensitivity is actually increased during smooth pursuit movements [Schütz et al., 2009].

Vergence Vergence eye movements are used to align both eyes on the same object. The eyes move in opposite directions, either towards or away from each other, which allows viewing objects at different distances (or depth). These are slow-velocity eye movements, rarely exceeding 10 deg per second.

The succession of saccades and fixations carried out while examining a scene is called the *scanpath* [Noton and Stark, 1971]. Techniques for measuring eye movements and recording scanpaths have existed for many years. The first real eye tracking devices were built in 1935 by Buswell [Buswell, 1935], who used photographic gaze monitoring to measure where subjects directed their eyes when viewing art. Later, [Yarbus, 1967] performed important eye tracking research (with the invasive scleral coil method), where he demonstrated the strong influence of task on eye movements. Today's most commonly used eye trackers determine in real time the current focus of the eye, e.g. by recording reflections of projected infrared light from the cornea, and the pupil.

2.2.2 Visual attention

In our daily life, we are constantly faced with a vast amount of visual information that the human visual system cannot simultaneously process. As argued above, despite the illusion that we perceive the entire visual field in full detail, only a small fraction of this information — which falls on the fovea — can be handled at any one time. The selected locations in the visual field are brought to the fovea and processed by the succession of saccades and fixations. The set of mechanisms through which relevant information is selected is called *visual attention*. Attention is therefore an important component of natural vision. It allows to allocate the brain's limited cognitive resources effectively to behaviourally relevant scene locations.

Covert versus overt attention

Although intensively researched, the neurophysiological aspects of attention are still poorly understood. Several brain areas seem to take part in guiding attention, but the exact role of each area is still an unsolved question. An important misconception is that attentional deployment cannot occur without an accompanying eye movement. The ability to direct attention to parts of the visual scene without moving the eyes is called *covert attention* [James et al., 1981]. However, under natural viewing conditions, shifting attention is usually associated with a gaze shift; this is referred to as *overt attention* [von Helmholtz, 1866]. The relationship between these two attention types has been the topic of extensive debate. Some physiological and behavioural evidence suggests that covert selection is closely related to the overt fixational orientation: fast covert attention shifts are made to subsequent scene targets prior to the saccade initiation [Deubel, 2008], and thus they play a key role in the programming of eye movements. Others propose that the two processes arise from the action of a single motor command: covert attention is a "by-product" of the saccade generating mechanism [Rizzolatti et al., 1987].

While eye movements (i.e. overt attention) can be easily measured with eye trackers, investigating covert attention (especially under natural viewing conditions) is difficult. Overt visual attention is also the focus of this thesis.

Visual attention is thus a mechanism used to focus the limited cognitive capacities of the brain on selections of the visual input. But how do we decide which particular locations to pay attention to? What guides our eyes from one fixation to the next? Why are certain parts of a complex scene attended and others not? Such questions were first addressed by the seminal work of Yarbus [Yarbus, 1967], who provided evidence that the location and sequence of eye movements is far from random. For example, during the free inspection of a face, most saccades actually fall on the facial features, such as eyes, nose, and mouth. Consequently, the patterns of fixations of subjects viewing the same scene are highly similar. However, he also pointed out that the sequence of saccades and fixations (i.e. the scanpath) cannot always be predicted from the stimulus itself. This sequence changes by asking the viewers to report on different properties of the scene (e.g. the age or financial situation of the scene characters). This suggests that also higher cognitive processes determine how a scene is explored. His findings were later corroborated by similar studies that examined eye movements during everyday activities, such as tea- and sandwich-making, and driving [Land and Hayhoe, 2001, Ballard and Hayhoe, 2009]. Their results reinforced the idea that visual attention is a function of the continuous interaction between two different mechanisms: on the one hand, *top-down* or goal-driven, and *bottom-up* or stimulus-driven on the other [James, 1890, Treisman and Gelade, 1980, Bergen and Julez, 1983]. In the following, we will shortly explain the two processes.

2.2.3 Bottom-up versus top-down attention

Top-down attention (also called *endogenous* attention) [Desimone and Duncan, 1995, Yarbus, 1967] is a voluntary, conscious form of attention control, where the task at hand and the observer's intentions, motivations, and emotions determine the locations to be fixated. As it involves the voluntary intent to attend to some portion of the visual field, it is a rather slow process.

Bottom-up attention, on the other hand, refers to a set of much faster mechanisms by which eye movements are driven involuntarily, influenced by low-level visual features, such as contrast, colour, and motion, i.e. stimulus "salience" [Yarbus, 1967]. In this case, attention is grabbed involuntarily by an external stimulation (e.g. a bee flying by), therefore this kind of attention is also called *exogenous* or reflexive attention.

Due to the complexity of high-level cognitive functions, much research has focused on bottom-up, so-called data-driven factors, investigating the relationship between eye movements and low-level image features at fixations. This dissertation also focuses on the bottom-up aspects of attentional selection. The characteristics of eye movement patterns have been studied intensively in everyday activities, such as orienting, reading [Huey, 1898], and visual search [Wolfe, 1998]. Although different situations, all three pose constraints on the visual activity required, and stimulus-driven attention interferes with top-down mechanisms. Therefore, to examine what role bottom-up factors play in attentional selection, it has been proposed to analyze eye movements in a less constrained situation: during the scanning of naturalistic images and image sequences. For instance, it has been found that spatial contrast tends to be higher at the centre of fixation than at random control locations [Reinagel and Zador, 1999, Tatler et al., 2005]. Also, there are regularities in the higher-order image statistics at fixations as well [Zetzsche et al., 1998]. When viewing image sequences, eyes are often directed at regions with temporal change.

Attentional selection has also been studied by looking at neuropathological diseases. For instance, some patients, following a right hemispheric stroke, show a unilateral neglect syndrome: unless stimuli are highly salient in the impaired hemifield, they are ignored. Better knowledge of what low-level stimulus properties are needed to guide these patients' attention into the neglected hemifield holds great potential for future therapeutic and assistive interventions, and is currently under investigation by a group of GazeCom researchers in collaboration with the Lübeck Neurology Department.

To exploit the accumulated knowledge on both the neurophysiology and the psychophysical properties (see above) of attention, theoretical models of attention have been proposed. They aim at a better understanding and modelling of visual perception by means of simulating behavioural data, e.g. predicting eye movements. Several competing psychological models have been proposed in the literature but, due to the difficulties involved in modelling higher mental states (motivation, emotions), only few have considered top-down mechanisms. Some notable models of attention are the spotlight model [Posner, 1980], the zoom lens model [Eriksen and James, 1986], the Guided Search model of Wolfe [Wolfe, 1998], and the Feature Integration Theory of Treisman [Treisman and Gelade, 1980]. A more recent approach, called the *triadic architecture* and proposed by [Rensink, 2000] has shown promising results in integrating higher scene knowledge (the so-called *gist*) in the model.

Guided by such psychophysical theories, more recently, *computational models of attention* have been proposed that aim not only at replicating the physiological and psychophysical properties of attention, but also at improving machine vision algorithms in computer vision and robotics. By identifying so-called *points of interest* within a scene, computational models of attention enable the often time- and resource-consuming image processing to focus only on these potentially relevant scene locations. Therefore, the selective processing of visual information has become an important component of biologically-inspired machine vision systems. Note that these bio-inspired models differ from the purely computational interest-point detectors [Harris and Stephens, 1988] that are ubiquitous in the computer vision community (for an overview see Schmid et al. [2000]). Although the latter are inspired by the saliency mechanisms in natural vision, they do not strive to grant biological plausibility or to simulate human gaze behaviour.

Computational saliency models centre on the concept of a *saliency map*, which assigns to each pixel of an image or video a saliency value indicating how likely it is that the viewer of the image or video fixates that location due to its (relative) conspicuity. Although the various models differ in their underlying assumptions concerning the model architecture and the formal definition of saliency, they share some properties that make them biologically plausible. Numerous methods [Meur et al., 2006, Bruce and Tsotsos, 2006, Gao and Vasconcelos, 2009, Gao et al., 2009], including the perhaps most well-known models for bottom-up saliency of Itti and Koch [Itti et al., 1998, 2003, Navalpakkam and Itti, 2005], follow the Feature Integration Theory introduced by Treisman [Treisman and Gelade, 1980], which we will now describe briefly.

2.2.4 Feature Integration Theory

According to the Feature Integration Theory (depicted schematically in Figure 2.3), in the preattentive step of attention, basic visual features (orientation, colour, contrast, etc.) are extracted in parallel on multiple scales, and stored in separate low-level feature maps. Normalized centre-surround difference maps are then computed for individual features and later combined by a weighting scheme to form a master saliency map. In the next, sequential step, attention is guided to peaks (i.e. locations with highest salience) in this map in a winner-take-all fashion. An inhibition-of-return mechanism prevents attention from returning to an already attended location.

The first computational formulation of a model based on the Feature Integration Theory was that of Koch and Ullman [Koch and Ullman, 1985]. Their model served as algorithmic foundation for subsequent implementations (e.g. Itti et al. [1998, 2003]), and other computational models of saliency. Koch and Ullman's initial model has undergone several modifications and extensions since the original description. It has been, for instance, extended to the temporal domain [Itti et al., 2003], and also top-down priors have been incorporated to model phenomena beyond bottom-up attention [Navalpakkam and Itti, 2005]. In [Siagian and Itti, 2007], for example, a low-dimensional signature vector, called the gist of the scene and acquired at multiple scales from basic visual features, was used to perform scene classification. In this thesis, the model of Itti et al. [1998, 2003] serves as a baseline for comparison with our work, and hence, in Appendix B.2 we briefly present the model architecture and the main computational steps.

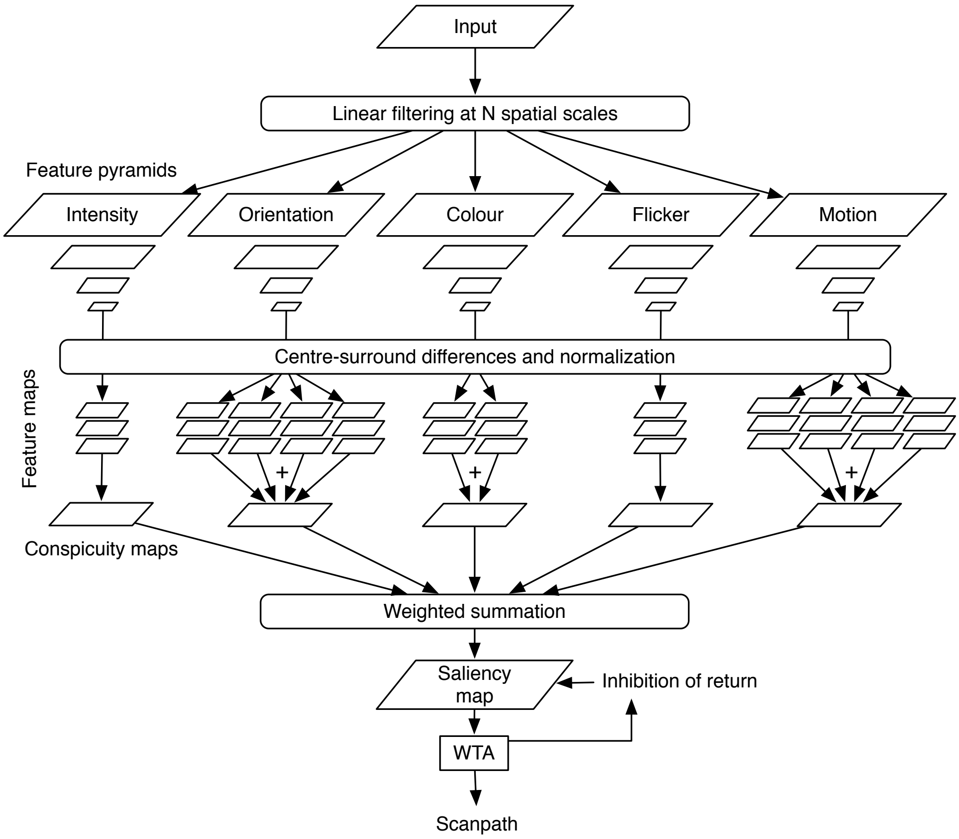


Figure 2.3: The architecture of the Koch and Ullman saliency model (as implemented by Itti et al. [1998, 2003]). Various low-level features (such as colour, luminance, and orientation for static images, and additionally flicker and motion features for videos) are extracted on multiple scales and stored in separate feature maps. A unique saliency map is generated through the combination of centre-surround feature maps (conspicuity maps). On this map, biological mechanisms, such as winner-take-all (WTA) competition and inhibition-of-return, are used to shift attention among the salient regions, thus generating a scanpath for an input scene. For a more formal description of the main computational steps see Appendix B.2.

2.3 Computational modelling of visual attention

In the following, we shall review some of the most important computational saliency models in the literature. Existing bottom-up saliency models, be they purely computational or biologically inspired, differ in the underlying *computational principles* they use to formally define the concept of saliency and motivate the model architecture (i.e. the choice of optimal features and major computational steps). A number of recent approaches turn to information theory to define “distinctiveness”, i.e. conspicuity. The model of Bruce and Tsotsos [Bruce and Tsotsos, 2006] aims at maximizing Shannon’s self-information to find the most informative locations in the image. Gao et al. [Gao and Vasconcelos, 2009, Gao et al., 2009] introduced the concept of “discriminant saliency”, which based on the definition of the target and null hypotheses (e.g. centre vs. surround, object class of interest vs. all other object classes) can act both as a bottom-up saliency predictor or top-down object detector. In this context, salient locations are those where the discrimination between target and non-target (in terms of some selected optimal features) can be made with minimum probability of error. Discrimination and classification confidence are here defined with respect to a number of existing computational principles for perceptual organization (e.g. infomax or Barlow’s inference by detection of suspicious coincidences).

The authors in [Avraham and Lindenbaum, 2010] present a region-based bottom-up model for images, which uses roughly segmented regions as candidates for salient objects. The most salient segment is found through graphical model approximation. The proposed stochastic model here, too, quantifies several intuitive observations, such as the likelihood of correspondence between visually similar image regions, and the assumption that the number of interesting objects in the scene is small.

Often, the problem of predicting eye movements on complex scenes is formulated in a Bayesian framework. This kind of approach provides an elegant way to, again, incorporate prior knowledge, e.g. about the statistics of visual attributes in specific scene types or descriptions and layout of the scene. Itti and Baldi [2009], for instance, proposed a Bayesian notion of surprise measured in “wows”, by calculating the mismatch (or Kullback-Leibler divergence) between expectations of the observer, i.e. priors, and the perceived reality, i.e. posteriors. The models SUN (for static scenes) and SUNDAy (for videos) of [Zhang et al., 2008, 2009], also use a Bayesian framework to analyze fixations. Similarly to [Bruce and Tsotsos, 2006], novelty is defined as self-information of the visual features, but the feature statistics used to detect outliers are learned from previous examples, and are not based only on the current image or video. For comparison purposes in the later chapters, in Appendix B.1 we formally describe the model architecture of SUNDAy.

While most approaches described above strive to address biological plausibility, the resulting models tend to be complex, having a large number of free parameters that need to be tuned by hand. Learning techniques are increasingly being employed as a practical solution to the parameter tuning problem (e.g. as above in [Zhang et al., 2008]). Such models even allow to infer the model structure from the data, without the need to quantify several assumptions about perceptual processes. Still, the usefulness of learning in visual saliency modelling has been recognized only recently. Kienzle et al. were the first to derive saliency-based interest operators from human eye movement data using machine learning techniques that operated directly on the pixel intensities of static scenes [Kienzle et al., 2007b] and Hollywood movies [Kienzle et al., 2007a]. They showed that the learned discriminative features have a centre-surround pattern. Due to constraints imposed by the reduced ability of learning algorithms to operate in high-dimensional (pixel) spaces given a limited number of training samples, the algorithms in [Kienzle et al., 2007b,a] were limited to a single spatial scale. A data-driven approach is used in [Judd et al., 2009], too, where optimal parameters are learned (from fixation data on static scenes) for an attention model that is based on low-, mid- and high-level features calculated by several existing saliency methods. In [Liu et al., 2010], another supervised approach aims at learning to detect salient objects from manually labelled examples. Here, a set of novel features, such as multi-scale contrast, centre-surround histogram, and colour spatial distribution, are combined through conditional random field learning.

While several models exist for saliency prediction on still images, only recently has the number of studies dealing with scene sequences increased. Although some of the static approaches have been generalized to videos [Kienzle et al., 2007a, Itti and Baldi, 2009, Zhang et al., 2009], these models often lack a unified framework for the static (spatial) and space-time saliency domains. Traditional ways to incorporate temporal information have often simply complemented the feature set with dynamic features, e.g. the optical flow information. In [Liu et al., 2010], for instance, the same set of novel features proposed for still images are defined on the motion field to capture spatiotemporal cues. Mahadevan and Vasconcelos [2010] extended the bottom-up discriminant centre surround saliency model of Gao and Vasconcelos [2009] to background subtraction in highly dynamic scenes. In a saliency prediction framework, background regions are those classified as non-salient by comparison of centre and surround appearance and dynamics (the video patches being modelled as dynamic textures).

As it has been shown in [Böhme et al., 2006], simple spatiotemporal saliency measures based on intrinsic dimensionality can generate a small set of salient locations that is likely to contain the next saccade target. Throughout this thesis, we shall make extensive use of such geometrical features (combined with machine learning) to investigate perceptual phenomena and to predict gaze in natural dynamic scenes.

Incorporating temporal information is also not straightforward in a learning context, where the task of eye movement prediction is further complicated by the increased number of (pixel-) dimensions.

Since most saliency models for videos are sensitive to dynamic content, camera motion and film-editing (e.g. jump cuts and gradual transitions) pose difficulties — even for the most advanced predictors — by causing false alarms in the salient features. Such a shortcoming is typically corrected with compensation of camera motion and shot boundary elimination. Shot boundary detection, too, can be tackled with an attentional paradigm. In Boccignone et al. [2005], for example, saliency maps of nearby frames are compared for consistency and shot boundaries are detected when the similarity is below a given threshold.

2.4 Applications in computer vision

Until recently, most computer vision algorithms performed an in-depth processing of the input data, e.g. by scanning the image exhaustively to locate objects of interest. However, the need to process — often in real-time and with restricted computing resources, e.g. in the case of mobile robots — a vast amount of continuously inflowing high-resolution data, had turned such brute-force approaches computationally intractable. Therefore, as a way to control the combinatorial explosion, the ability to restrict the processing to the salient, i.e. potentially relevant scene locations in the scene has proven invaluable for computer vision applications. Saliency is therefore justified as a preprocessing step for image analysis that not only saves computation but can also improve performance.

The relevance of visual attention is probably the most evident in object recognition, detection, and tracking, e.g. [Rutishauser et al., 2004, Serre et al., 2007, Liu et al., 2010], tasks that assume a two-stage processing — (1) attentional selection and (2) recognition through a classifier — that is often adapted to human perception [Neisser, 1967]. The interest points detected in the first, attentional filtering phase are utilized either for pattern matching or to extract local descriptors that serve in later steps for object and scene representation and discrimination. The HMAX system [Riesenhuber and Poggio, 1999], one of the early biologically motivated object recognizers, followed this architecture, and was capable of simulating processes (related to object recognition) in the human cortex.

Other applications of saliency models include image and video compression [Geisler and Perry, 1998, Ouerhani et al., 2001, Itti, 2004b]. Here, a saliency-based non-uniform compression algorithm allocates more bits for salient regions, whereas the rest is encoded with lower quality. Thus, relevant scene regions have a higher reconstruction quality as compared to the rest of the image.

Attention-based algorithms have been proposed also for automatic image cropping [Santella et al., 2006] (e.g. for centred display of images on small portable screens), image and video quality assessment [Ninassi et al., 2007], non-photorealisic rendering [DeCarlo and Santella, 2002], and video event detection and summarization [Evangelopoulos et al., 2008].

In robotics, the field of active vision [Aloimonos et al., 1988], i.e. the active redirection of a camera to semantically relevant scene regions, has largely benefitted from the use of the visual attention paradigm. Here, too, the aim is to focus the computationally expensive processing only on the relevant scene locations. Active vision is an important component of solutions in robotics. Thus, attention systems are used to guide the "gaze" (i.e. cameras) of robots to aid navigation, self-localization, object manipulation, and human-robot interaction [Frintrop et al., 2006, Siagian and Itti, 2007].

To summarize this chapter, we have seen that the capabilities of the human visual system are highly space-variant. Tasks that require high visual fidelity, such as object recognition or reading, can only be performed foveally, and attention usually is deployed simultaneously at the centre of fixation. In order to obtain detailed information from the whole visual field, humans make several eye movements per second to bring the fovea onto different parts of the scene, and the periphery then is mainly used for navigation and to determine where to direct the fovea next.

Several models have been put forward to establish the exact mechanism by which eye movements are controlled. A common approach is based on the Feature Integration Theory and models the relevance of a location by the statistical irregularity of a set of low-level image features, such as a local deviation from surrounding orientation statistics. More recently, however, attempts have also been made to automatically extract the relevant structure of salient locations based on a set of human eye movements. In Chapter 5, we will use this technique and show that our model outperforms other state-of-the-art models on naturalistic videos. A successful predictor of relevance in natural environments can also be useful in computer vision and active vision scenarios, such as video compression and robotics.

3

Efficient coding of natural image sequences

In the following chapter, we will discuss some basic aspects of the efficient representation of visual data both in artificial and in biological systems.

3.1 Introduction

Digital images and image sequences are stored as numerical arrays of pixel intensities. As Figure 3.1 illustrates, making sense of such numerical data proves extremely difficult. Even though the photo-receptors of our retina receive visual information in essentially the same manner, the human visual system manages to effortlessly decipher information from this (seemingly nonsensical) data, e.g. we recognize the position and identity of objects in the scene in no time at all (see Figure 3.2). Understanding how the human visual system solves this extremely difficult computational task is one of the fundamental challenges in human vision and neuroscience. Beyond its biological importance, the gained knowledge could also guide computer vision research towards building efficient machine vision systems that mimic human vision.

Some insight into the perceptual processes going on in our brain can be obtained by considering the nature of the visual input the human visual system needs to decipher. For this, let us consider images of a fixed size, say 128 by 128 pixels. Each image can be represented as a sample (or point) of an $128 \times 128 = 16,384$ dimensional pixel intensity space, in which each axis stands for the brightness value of one image pixel (for simplicity, we will discuss grayscale images only). One important observation relates to the fact that in this high-dimensional space "natural" images, i.e. the kind of sensory input that biological visual systems got adapted to during evolu-

240	241	248	254	254	252	253	252	252	250	245	242	240	237	209	171	170	186	188	171	156	149	139	134	132	144	142	143	142	151
255	255	255	253	254	254	252	250	248	244	237	230	222	214	207	192	180	159	142	122	101	89	91	99	97	109	123	139	141	150
253	250	255	251	252	248	245	242	239	232	222	210	195	182	169	145	125	99	74	54	43	32	44	62	70	88	111	123	138	146
254	253	253	248	246	242	240	234	225	212	197	179	158	141	118	90	73	68	72	83	82	54	41	51	65	82	99	116	140	153
251	252	248	243	235	230	220	210	194	174	152	130	105	85	78	89	103	107	103	97	99	86	56	47	59	77	86	113	121	132
249	245	239	228	220	211	196	178	157	133	104	78	74	87	103	117	127	123	108	97	88	77	48	34	52	62	70	94	85	144
241	234	220	209	200	186	167	139	110	85	69	76	102	124	129	110	115	110	90	63	51	37	27	21	21	26	33	39	96	127
231	219	200	187	175	152	122	92	76	82	95	114	133	139	121	88	81	66	44	20	19	17	11	11	8	10	26	44	67	109
207	193	176	160	135	105	78	78	92	120	140	144	134	118	76	46	35	27	18	14	28	31	13	8	7	25	53	83	135	134
183	168	149	120	88	70	84	113	135	147	152	137	104	71	45	25	22	20	20	21	47	50	22	17	23	48	70	87	116	139
160	137	107	81	79	99	128	149	147	136	125	101	70	45	27	31	35	28	42	38	46	41	32	29	43	53	88	126	146	153
128	101	83	91	121	154	165	153	120	98	82	64	53	45	44	43	49	43	47	59	54	46	41	49	63	84	97	141	150	158
103	93	108	145	167	168	150	116	74	67	82	91	90	74	56	48	85	102	54	66	58	48	51	60	75	104	148	175	191	187
90	109	142	168	157	137	108	76	60	97	138	159	172	127	59	35	87	171	79	45	58	60	72	83	95	126	172	200	208	178
120	140	141	137	109	88	83	89	120	176	207	216	214	145	66	33	67	129	62	49	62	83	107	116	121	155	203	220	218	197
145	132	108	94	69	75	113	155	188	220	240	236	220	149	69	47	40	38	31	47	77	109	145	157	155	183	228	235	228	216
118	101	90	89	82	116	172	202	222	235	238	238	231	159	73	57	35	12	36	51	95	137	187	208	202	211	239	244	239	228
107	113	96	83	92	141	193	217	229	236	238	239	234	173	84	63	54	45	51	69	109	150	210	240	238	236	244	248	247	234
159	163	138	104	89	102	164	206	226	223	229	233	229	186	100	68	68	53	63	89	117	156	214	247	251	250	249	249	247	235
210	211	211	185	142	93	97	146	206	228	226	227	231	199	119	66	63	57	76	80	96	147	208	241	243	248	252	251	246	235
235	241	243	233	218	164	128	106	112	158	219	239	228	199	135	70	53	60	57	43	55	109	185	218	218	227	243	250	244	235
247	250	252	246	244	225	205	177	132	105	95	139	182	190	151	91	51	40	42	39	45	48	108	173	206	207	225	244	245	233
251	255	253	254	252	244	243	230	198	168	125	97	81	80	85	75	53	41	44	50	37	35	90	138	165	201	210	233	237	237
254	254	250	250	247	247	246	243	232	216	190	171	137	111	97	79	54	49	61	46	18	45	87	124	132	180	207	228	239	238
255	252	254	250	247	245	244	241	243	236	215	204	190	185	171	137	127	74	52	63	43	36	123	158	153	154	195	214	232	246
251	251	254	252	251	247	250	247	243	241	237	222	206	213	200	172	168	167	137	114	138	156	126	155	201	193	190	216	240	245

Figure 3.1: A natural image displayed using pixel-intensity values. Numbers correspond to intensity values ranging from 0 (black) to 255 (white).

tion, are not uniformly distributed. An arbitrary element of this space, e.g. an image whose pixel intensities are chosen at random, will most certainly be noise-like and will hardly ever resemble natural images. It actually turns out that the distribution of natural images is not uniform; such images lie on an unknown lower-dimensional manifold in the "space" of all possible image patches. This equates to the fact that, in an information-theoretic sense, natural images contain a significant amount of *redundancies*. The seminal paper of Attneave [1954] was the first to demonstrate the redundant nature of natural visual stimuli. Exploiting these redundancies, i.e. by recoding the input signal in a more efficient way, is crucial for biological visual systems. Furthermore, characterizing the statistical properties of natural images is essential also for engineering applications, where images and image sequences ought to be stored and transmitted in possibly the most compact digital format. The work of Attneave [1954] and Barlow [1961] has led to the "efficient coding hypothesis", according to which the brain uses efficient representations to encode the structural regularities observed in an organism's natural visual environment. In support of this theory, Olshausen and Field [1996, 1997] found filters resembling the receptive fields of simple-cells in cortical area V1, when these filters — so-called overcomplete sparse codes/bases — were optimized to encode natural stimuli.

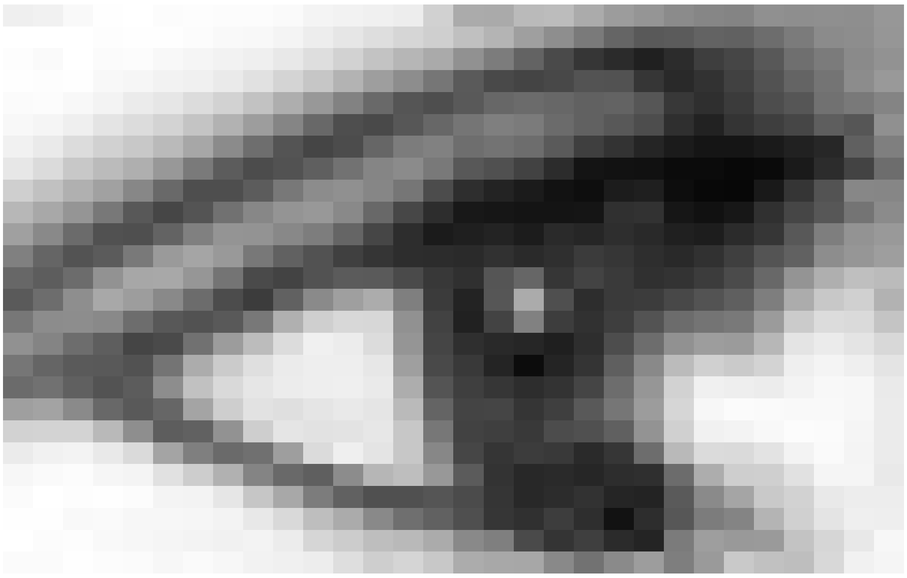


Figure 3.2: The image of Figure 3.1. It is now easy to recognize the content of the image. If you still have difficulties, squint your eyes.

The concept of *intrinsic dimensionality*, which we shall introduce shortly, provides a basic description of how a multidimensional (e.g. spatiotemporal) signal may change. Within this framework, typical image and video structures can be characterized and categorized. Those parts of the visual input where the image/video does not change in a particular direction (or set of directions) contain redundancies and may not need to be encoded in an efficient representation. In this thesis, we shall provide evidence from eye movement studies that the human visual system avoids such redundant parts and therefore indeed employs efficient coding to sense the visual world.

In the second part of this chapter, we shall discuss coding schemes for an efficient representation of images and image sequences at multiple scales. *Pyramidal multiresolution data structures* encode each scale at optimal resolution, i.e. with the fewest bits necessary, and therefore will allow us throughout this thesis to efficiently analyse and manipulate high-resolution videos. Again, we can make a strong link between efficient techniques from computer vision and human vision because early visual processing also uses a bandpass representation of the visual information [Marr, 1982].

3.2 Geometry of time-varying images

We will start by looking at the geometry of time-varying images and ways to characterize (and categorize) various types of video intensity changes. For this, we will first introduce the concept of intrinsic dimension and discuss a technique to estimate it, following [Mota et al., 2006].

3.2.1 Intrinsic dimension

The *intrinsic dimension* (iD) [Zetzsche and Barth, 1990] quantifies the information content of a signal. It describes the number of degrees of freedom needed to locally represent the observed signal. Thus, static and homogeneous video locations are intrinsically zero dimensional ($i0D$); stationary edges and uniform regions that change in time have an intrinsic dimension of one ($i1D$); stationary corners and edges that change in time are $i2D$, while transient corners and non-uniform motion are intrinsically three dimensional ($i3D$). For an illustration with a synthetic image see Figure 3.3. An example of the intrinsic dimensionality of natural movies is shown in Figure 3.4. Since, in natural scenes, regions with high intrinsic dimension are less common than regions with low intrinsic dimension [Zetzsche et al., 1993], the concept of intrinsic dimension is particularly relevant for image and video coding. Moreover, it has been shown that an image or video can be fully reconstructed from only those regions where the iD is greater than one, i.e. $i0D$ and $i1D$ regions are redundant [Barth et al., 1993, Mota and Barth, 2000].

We consider a grayscale video represented by the function $f(\mathbf{p}) : \mathbb{R}^3 \to \mathbb{R}$, $\mathbf{p} = (x, y, t)$. Following Mota et al. [2004b], to estimate the intrinsic dimension of a given video region Ω, a linear subspace $E \subset \mathbb{R}^3$ of highest dimension is chosen, such that

$$f(\mathbf{p} + \mathbf{v}) = f(\mathbf{p}) \text{ for all } \mathbf{p}, \mathbf{v} \text{ such that } \mathbf{p}, \mathbf{p} + \mathbf{v} \in \Omega, \mathbf{v} \in E. \tag{3.1}$$

The intrinsic dimension of Ω is $3 - \dim(E)$ for videos and $n - \dim(E)$ for n-dimensional signals.

The intrinsic dimension can be estimated with different differential methods; here, we will use the method based on the *structure tensor* [Jähne et al., 1999].

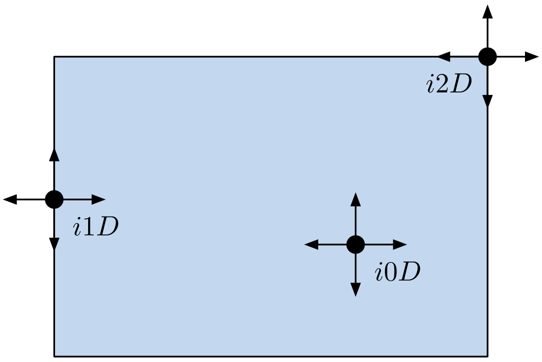


Figure 3.3: Illustration of the intrinsic dimensionality of images. Using this measure, one can distinguish between homogeneous ($i0D$), edge-like ($i1D$) and junction-like structures ($i2D$). Spatiotemporal corners in videos are $i3D$.

3.2.2 Invariants of the structure tensor

As shown in [Mota et al., 2004b], Equation 3.1 is equivalent to writing

$$\frac{\partial f(\mathbf{p})}{\partial \mathbf{v}} = 0 \quad \text{for all } \mathbf{v} \in E, \mathbf{p} \in \Omega \,. \tag{3.2}$$

which, based on [Mota et al., 2004b], is in turn equivalent to minimizing the energy function

$$\varepsilon(\mathbf{v}) = \int_{\Omega} \left| \frac{\partial f}{\partial \mathbf{v}} \right|^2 \mathrm{d}\Omega = 0 \,. \tag{3.3}$$

At any point $\mathbf{p} \in \Omega$, $\frac{\partial f}{\partial \mathbf{v}}$ is in fact the directional derivative of f along $\mathbf{v}$ and can be written as

$$\frac{\partial f}{\partial \mathbf{v}} = v_x f_x + v_y f_y + v_z f_z = \sum_{i \in \{x,y,t\}} v_i f_i \,, \tag{3.4}$$

where $\mathbf{v} = (v_x, v_y, v_t)$ and f_x, f_y, and f_t stand for the first-order partial derivatives of f. Thus, Equation 3.2 is equivalent to writing

$$\begin{bmatrix} v_x \; v_y \; v_t \end{bmatrix} \cdot \begin{bmatrix} f_x \\ f_y \\ f_t \end{bmatrix} = 0 \tag{3.5}$$

and

$$\left|\frac{\partial f}{\partial \mathbf{v}}\right|^2 = \sum_{i,j\in\{x,y,t\}} v_i f_i v_j f_j = \tag{3.6}$$

$$= \begin{bmatrix} v_x \, v_y \, v_t \end{bmatrix} \cdot \begin{bmatrix} f_x \\ f_y \\ f_t \end{bmatrix} \cdot \begin{bmatrix} v_x \, v_y \, v_t \end{bmatrix} \cdot \begin{bmatrix} f_x \\ f_y \\ f_t \end{bmatrix} = \tag{3.7}$$

$$= \begin{bmatrix} v_x \, v_y \, v_t \end{bmatrix} \cdot \underbrace{\begin{bmatrix} f_x \\ f_y \\ f_t \end{bmatrix} \cdot \begin{bmatrix} f_x \, f_y \, f_t \end{bmatrix}} \cdot \begin{bmatrix} v_x \\ v_y \\ v_t \end{bmatrix} = \tag{3.8}$$

$$= \begin{bmatrix} v_x \, v_y \, v_t \end{bmatrix} \cdot \underbrace{\begin{bmatrix} f_x^2 & f_x f_y & f_x f_t \\ f_x f_y & f_y^2 & f_y f_t \\ f_x f_t & f_y f_t & f_t^2 \end{bmatrix}}_{\mathbf{J}'} \cdot \begin{bmatrix} v_x \\ v_y \\ v_t \end{bmatrix} . \tag{3.9}$$

Hence, within a spatiotemporal neighbourhood Ω, Equation 3.2 can be expressed as

$$\varepsilon(\mathbf{v}) = \int_\Omega \left|\frac{\partial f}{\partial \mathbf{v}}\right|^2 \mathrm{d}\Omega = \mathbf{v}^T \mathbf{J} \mathbf{v} \ , \tag{3.10}$$

where $\mathbf{J}$ is the *structure tensor* [Jähne et al., 1999]:

$$\mathbf{J} = \int_\Omega \mathbf{J}' \ \mathrm{d}\Omega = \int_\Omega \nabla f \otimes \nabla f \ \mathrm{d}\Omega = \int_\Omega \begin{bmatrix} f_x^2 & f_x f_y & f_x f_t \\ f_x f_y & f_y^2 & f_y f_t \\ f_x f_t & f_y f_t & f_t^2 \end{bmatrix} \mathrm{d}\Omega \ . \tag{3.11}$$

In the above formula, $\otimes$ denotes the tensor product. In practice, the integral over Ω is implemented as smoothing with a spatiotemporal Gaussian filter function. The linear subspace E is estimated as the eigenspace associated with the smallest eigenvalue of $\mathbf{J}$ [Mota et al., 2004b], and the intrinsic dimension of f within the neighbourhood Ω corresponds to the rank of $\mathbf{J}$. Alternatively, the intrinsic dimension can be computed from $\mathbf{J}$'s *geometrical invariants* H, S, and K [Mota et al., 2001]:

$$\begin{aligned} H &= 1/3 \ \mathrm{trace}(\mathbf{J}) &&= \lambda_1 + \lambda_2 + \lambda_3 \\ S &= M_{11} + M_{22} + M_{33} &&= \lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_1\lambda_3 \\ K &= |\mathbf{J}| &&= \lambda_1\lambda_2\lambda_3 \end{aligned} \tag{3.12}$$

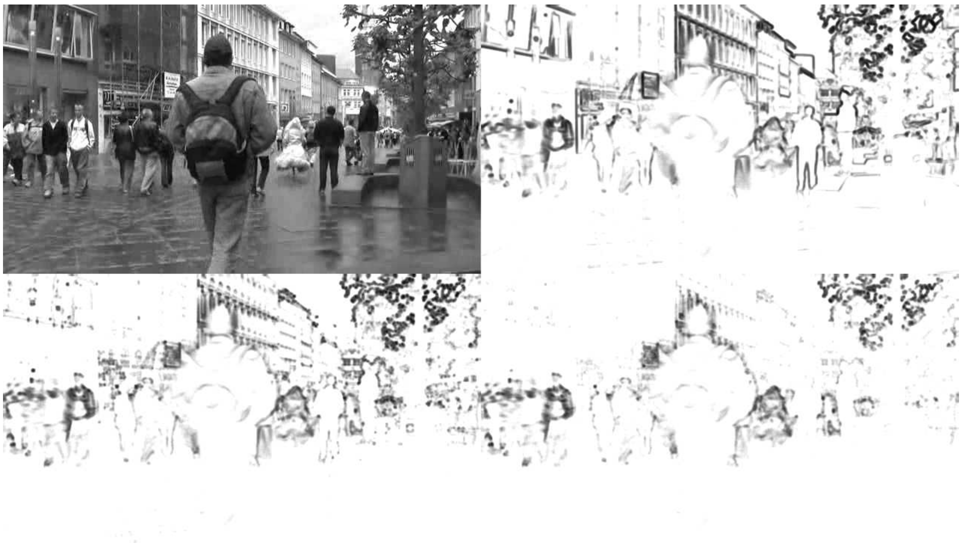


Figure 3.4: Still shot from a video (top left quadrant) and the corresponding geometrical invariants. For invariant K (bottom right quadrant), non-white locations change in all three spatiotemporal directions, whereas for S (bottom left), the video signal changes in at least two directions. Additionally, invariant H (top right) also responds to edges (i.e. one dimensional changes). The (small) response even of K at the corners of the windows is due to small camera vibrations and noise. For the invariants, the brightness has been thresholded and inverted for better legibility.

where λ_i are eigenvalues and $M_{i,j}$ are the minors of $\mathbf{J}$ (i.e. determinants of submatrices of $\mathbf{J}$ obtained by removing row i and column j). The geometrical invariants correspond to the minimum intrinsic dimension of a region, i.e. if $K \neq 0$, the intrinsic dimension is 3 ($i3D$); if $S \neq 0$ it is at least $i2D$; and if $H \neq 0$ it is at least $i1D$. In Figure 3.4 an example image is shown for the invariants on a natural image.

The structure tensor and its eigenvalue analysis are widely used in image processing to estimate orientation and motion [Granlund and Knutsson, 1995, Jähne et al., 1999]. The method has later been extended to multispectral images [Mota et al., 2006], and to multiple orientations and multiple motions [Mota et al., 2001, 2004a], and we will briefly outline these extensions in the following.

3.2.3 Multispectral invariants

The above formalization may only be used to estimate the intrinsic dimension of grayscale videos. To investigate colour saliency, the concept of intrinsic dimension has been extended to multispectral signals [Mota et al., 2006].

We consider a multispectral image sequence $\mathbf{f}$ with q colour channels ($\mathbf{f} : \mathbb{R}^3 \to \mathbb{R}^q$), and define the scalar product between two vectors $\mathbf{y} = (y_1, \dots, y_q)$ and $\mathbf{z} = (z_1, \dots, z_q)$ as $\mathbf{y} \cdot \mathbf{z} = \sum_{k=1}^{q} a_k y_k z_k$. The weights $a_k > 0$ are here meant to emphasize different colour channels, if needed.

As above, the intrinsic dimension of $\mathbf{f}$ within a small region Ω can be estimated by minimizing the energy function

$$\varepsilon(\mathbf{v}) = \int_\Omega \left\| \frac{\partial \mathbf{f}}{\partial \mathbf{v}} \right\|^2 \mathrm{d}\Omega, \tag{3.13}$$

where the directional derivative $\frac{\partial \mathbf{f}}{\partial \mathbf{v}}$ of $\mathbf{f}$ has a similar form, as above:

$$\frac{\partial \mathbf{f}}{\partial \mathbf{v}} = v_x \mathbf{f}_x + v_y \mathbf{f}_y + v_t \mathbf{f}_t, \quad \mathbf{v} \in E, E \subset \mathbb{R}^3. \tag{3.14}$$

With a similar derivation, the energy function can be expressed as

$$\varepsilon(\mathbf{v}) = \mathbf{v}^T \mathbf{J} \mathbf{v}, \tag{3.15}$$

where $\mathbf{J}$ is the *multispectral structure tensor*:

$$\mathbf{J} = \int_\Omega \begin{bmatrix} \|\mathbf{f_x}\|^2 & \mathbf{f_x} \cdot \mathbf{f_y} & \mathbf{f_x} \cdot \mathbf{f_t} \\ \mathbf{f_x} \cdot \mathbf{f_y} & \|\mathbf{f_y}\|^2 & \mathbf{f_y} \cdot \mathbf{f_t} \\ \mathbf{f_x} \cdot \mathbf{f_t} & \mathbf{f_y} \cdot \mathbf{f_t} & \|\mathbf{f_t}\|^2 \end{bmatrix} \mathrm{d}\Omega \, . \tag{3.16}$$

Note that the above formulation does not assume any particular colour space. However, videos are often represented in the $Y'C_bC_r$ colour space (instead of RGB, for instance) because the luma (Y') and the two chroma (C_b, C_r) channels are less correlated and the chroma channels are subsampled to take advantage of the lower colour sensitivity of the human visual system. However, when using $Y'C_bC_r$, the dynamic range of the luma channel is much greater than that of the chroma channels, so that the contribution of colour to $\mathbf{J}_{Y'C_bC_r}$ is small. To compensate for this, one can compute the standard deviation of each channel and use their inverse for the weights

a_y, a_u, and a_v.

3.2.4 Generalized structure tensor for multiple motions

In the following, we briefly touch on the extension of the structure tensor to the *generalized structure tensor* [Mota et al., 2001], with which the characterization of multiple superimposed motions becomes possible. In Chapter 5, we will use such generic representations to predict eye movements on multiple overlaid videos.

Let us consider the image sequence f that consists of the superposition of two transparent image layers that are moving with different constant velocities $\mathbf{u} = (u_x, u_y, 1)$ and $\mathbf{v} = (v_x, v_y, 1)$. The video signal f can thus be written as

$$f(\mathbf{p}, t) = g_1(\mathbf{p} - t\mathbf{u}) + g_2(\mathbf{p} - t\mathbf{v}), \tag{3.17}$$

The generalized structure tensor of f

$$\mathbf{J_G} = \int_\Omega \begin{bmatrix} f_{xx}^2 & f_{xx}f_{xy} & \cdots & f_{xx}f_{tt} \\ f_{xx}f_{xy} & f_{xy}^2 & \cdots & f_{xy}f_{tt} \\ \vdots & \vdots & & \vdots \\ f_{xx}f_{tt} & f_{xy}f_{tt} & \cdots & f_{tt}^2 \end{bmatrix} \mathrm{d}\Omega \tag{3.18}$$

and its corresponding invariants are used to characterize (and categorize) different combinations of multiple motions, such as transient dots, stationary and moving gratings, etc. For a definition of the generalized structure tensor for N (rather than only two) motions we refer to [Mota et al., 2001].

3.3 Multiscale representations

In the remainder of this chapter, we will review methods for an efficient representation of image and video signals at multiple spatiotemporal scales. Multiscale analysis is a well-established technique in signal and image processing. Whereas Fourier analysis tells us about what frequencies are present in the image, Fourier coefficients contain no spatial information. Multiresolution processing (e.g. pyramidal coding and wavelets [Mallat, 1989]), on the other hand, provide information on both the frequency and spatial domain simultaneously. Therefore, such an approach is able to model the function of the human visual system, which has been shown to represent visual information on several spatiotemporal scales [Marr, 1982]. To mimic this, in image processing, efficient representations, so-called *multiresolution image pyramids* have been developed. Pyramids correspond to a decomposition of the image (or video) into spatial (and temporal) frequency bands. They store information on each pyramid level in a compact format, with fewest bits possible. Here, we will review two basic pyramidal structures: (1) the Gaussian pyramid, which corresponds to a low-pass representation, and (2) the Laplacian pyramid, which performs a bandpass decomposition of the image.

3.3.1 Gaussian pyramid

The Gaussian multiresolution pyramid consists of a series of images or signals at different resolutions or sampling rates (for a review see [Jähne and Haußecker, 2000]). The resolution of the original signal is reduced iteratively by a factor of two, and the size of the signal decreases correspondingly; hence, the resulting low-pass filtered signal requires less storage space. For this property, multiresolution pyramids are useful for data compression, texture analysis, and scale-invariant pattern recognition (e.g. target tracking [Anderson et al., 1985]).

The Gaussian pyramid is constructed by progressively low-pass filtering and downsampling the input (see Figure 3.5). When subsampling by a factor of two, one must consider Shannon's *sampling theorem* [Shannon, 1949], i.e. the signal should not contain frequencies above the (halved) Nyquist rate of the new sampling rate. Therefore, to avoid aliasing, the low-pass filtering of the image is necessary before the downsampling.

More formally, we use $I(x, y)$ to denote the original image, which has W columns and H rows of pixels. By combining the two operations of filtering and subsampling into one step, the Gaussian pyramid representation of I is defined recursively as

$$\begin{array}{ll} G_0(x,y) & = I(x,y) \\ G_{k+1}(x,y) & = \sum_{i=-c}^{c} w_i \sum_{j=-c}^{c} w_j \cdot G_k(2x+i, 2y+j) \end{array}, \tag{3.19}$$

where each pyramid level G_k has a resolution of $W/2^k$ by $H/2^k$ pixels. w is a filtering kernel of length $2c + 1$ that has the following properties:

1. symmetric — so that no phase shift occurs during the filtering,

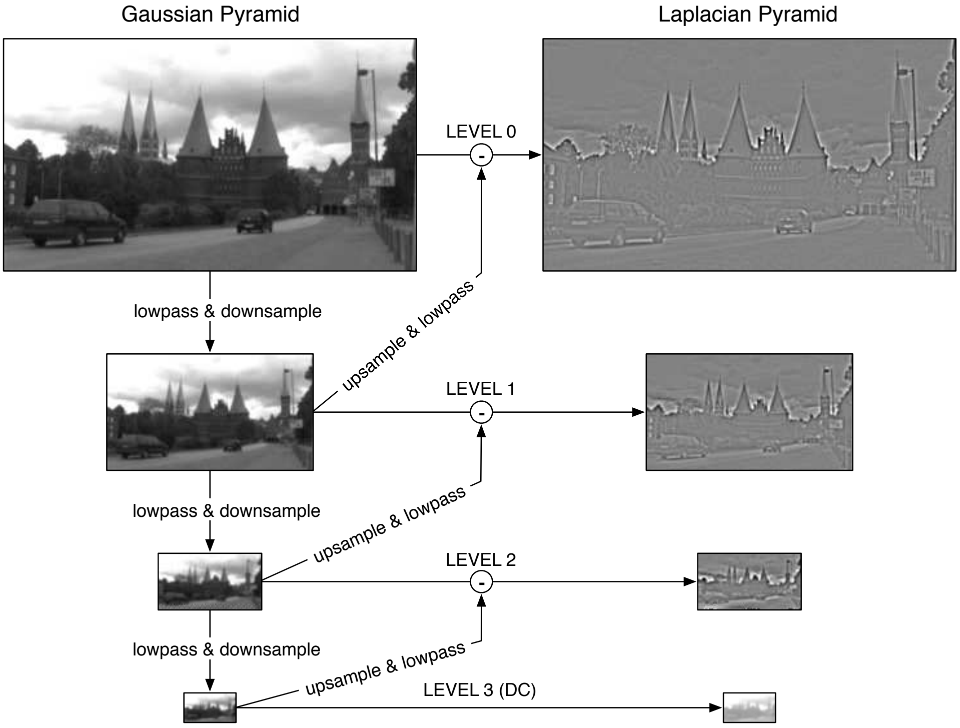


Figure 3.5: Construction of the Gaussian and Laplacian image pyramids. Depicted are four spatial pyramid levels. The Gaussian pyramid is constructed by repeatedly convolving the image with a low-pass kernel and downsampling the result (left). The Laplacian is then obtained by subtracting every Gaussian level (after upsampling and low-pass filtering it) from the next lower Gaussian level (right).

2. separable — a non-separable kernel $w_{i,j}$ is also possible, but less efficient,

3. the filter coefficients $w_{-c}, \ldots, w_c$ sum to one,

4. adheres to the equal contribution principle: each pixel should have equal contribution in the low-resolution version (otherwise, artefacts occur). A possible five-tap filter kernel with this property has the coefficients $w_0 = p$, $w_{-1} = w_1 = q$, $w_{-2} = w_2 = r$, where $p + 2r = 2q$.

A common choice for such a kernel is the 5-tap binomial $\frac{1}{16}[1\ 4\ 6\ 4\ 1]$. Such a filter closely approximates the Gaussian function, hence the name of the image pyramid. Note that while the lowest pyramid level contains the original image, the highest possible level (with a size of one pixels) consists of the *DC*, i.e. the mean luminance of the image.

The above recursive steps are equivalent to convolving I with a set of smoothing filters where the filters double in size from level to level. However, the effectiveness of the presented approach lies in the fact that the *same* (small) kernel is applied (locally) to all levels of the pyramid. From level to level, the band-limit is reduced by an octave, thus, for a finer-grained partition, alternative methods must be considered.

Extending such a multiresolution representation to the temporal domain is, in principle, straightforward. Instead of considering only every second pixel during downsampling, in the case of a temporal Gaussian pyramid every second frame must be discarded [Böhme et al., 2008]. Thus, from level to level, the temporal resolution of the video is halved. For a spatial Gaussian pyramid, the complete stack of pyramid levels (images of different resolution) can fit into memory at the same time. This is, however, not feasible for videos, so for an implementation an appropriate buffering of the required pyramid levels is necessary [Böhme et al., 2008].

Also note that because the filter w used in the computation of the different pyramid levels is non-causal, the method, in the current formulation, requires video frames (so-called *lookahead* frames) from the "future".

If the image sequence is progressively filtered and subsampled both in space and time, a spatiotemporal pyramid representation of the video is created. As opposed to an *isotropic pyramid*, where space and time are subsampled simultaneously, in case of an *anisotropic spatiotemporal pyramid* each level of a spatial pyramid is decomposed further into its temporal bands. Such a pyramid has the advantage of providing a finer partition of the spectrum, but is also computationally intensive.

3.3.2 Laplacian pyramid

The Laplacian multiresolution pyramid is an extension of the Gaussian pyramid and consist of a sequence of bandpass-filtered signals or images. A computationally efficient pyramid structure, similar to that seen above, makes this representation suitable for a wide range of computer vision applications from image fusion [Blum and Liu, 2005] and mosaicing [Burt and Adelson, 1983], to image enhancement [Trifas et al., 2006] and compression [Adelson and Burt, 1981]. Since the image is dissected into distinct frequency bands, the correlation of neighbouring pixels is reduced and the resulting images consist mostly of zeros, i.e. can be encoded with fewer bits.

As shown schematically in Figure 3.5, the Laplacian pyramid computes differences of successive levels of a Gaussian pyramid. Thus, each Laplacian level corresponds to a bandpass filtered version of the original image. Because two adjacent Gaussian levels differ in sampling density, it is necessary to upsample the lower level (before the subtraction), by inserting zeros between neighbouring pixels and interpolating with a Gaussian lowpass filter. The filter is often the same 5-tap binomial used in the creation of the Gaussian pyramid. This operation is often referred to as *expansion* as it doubles the image size at each iteration.

The pyramid construction can be formally summarized as

$$\begin{array}{ll} L_N(x,y) & = G_N(x,y) \\ L_k(x,y) & = G_k(x,y) - \text{Expand}(G_{k+1}(x,y)), \quad k = 0, \dots, N-1 \end{array}, \tag{3.20}$$

where the Expand operation is defined as

$$\text{Expand}(G_k(x,y)) = \sum_{i=-c}^{c} w_i \sum_{j=-c}^{c} w_j \cdot G_k\left(\frac{x-i}{2}, \frac{y-j}{2}\right). \tag{3.21}$$

Here, $(x-i)/2$ and $(y-j)/2$ contribute to the sum only when they are integers. Note that since the lowest Gaussian level G_N has no lower pair G_{N+1} anymore, L_N is set to the lowest level of the Gaussian, i.e. the DC component of the image.

Similarly to the Gaussian case, the pyramid construction is now equivalent to repetitively filtering the image with different kernels that are the difference of two Gaussian kernels with varying width. Again, with such a pyramid scheme, filtering operations also with large kernels can be performed efficiently.

An important advantage of the approach is that the construction scheme of the Laplacian pyramid can be easily inverted, which allows a perfect reconstruction of the original image. To *synthesize* the image from its bandpass decomposition, the Laplacian levels (starting with the lowest) are iteratively upsampled and added to the next higher level:

$$\begin{aligned} G_N(x,y) &= L_N(x,y) \\ G_k(x,y) &= L_k(x,y) + \text{Expand}(L_{k+1}(x,y)), \quad k = 0, \ldots, N-1 \end{aligned} \quad . \tag{3.22}$$

The reconstructed image is contained in the highest resolution G_0.

Since the distinct frequency bands can now be easily accessed, a modification (e.g. attenuation or amplification) of the frequency content of individual pyramid levels is achievable. In Chapter 6, we make use of this property to manipulate high-resolution videos with the goal of guiding the gaze, and implicitly the attention, of the viewer.

An extension to the temporal domain, in the same manner as for the Gaussian pyramid, is conceptually straightforward but, technically, hides a great amount of complexity (e.g. the buffering of intermediate results is needed). Similarly, a generalization to the spatiotemporal domain is also possible. Again, such a pyramid can either be *isotropic*, i.e. the spatial and temporal frequences vary together (low spatial with low temporal, high spatial with high temporal), or *anisotropic*, with which a finer-grained decomposition of the spectrum is possible. For example, in such an image pyramid high spatial and low temporal frequencies are also represented.

In this chapter, we have reviewed some basic aspects of image processing and the efficient coding of natural image sequences. In the following chapters, we will describe our own original research that will make use of such techniques.

4
Eye movements on naturalistic videos

In this chapter, we shall investigate a less-studied aspect of gaze allocation on naturalistic videos. Due to the highly predictive nature of salient real-world events, eye movements are often anticipating such events rather than just responding to them. Here we quantify the anticipatory nature of eye movements during the free viewing of real-world videos, and compare the degree of anticipation with that on other, less realistic stimuli, such as edited TV-clips and video games. The work in this chapter is motivated by our research on gaze guidance, where the optimal timing of so-called gaze-capturing events is critical for obtaining the desired effect, i.e. to unobtrusively guide gaze in real time to relevant scene areas. There is also a strong link to our work on the prediction of eye movements. We will here employ methods from the previous chapter (namely, the geometric invariants of the structure tensor) as simple means of identifying spatiotemporally salient events. In the next chapter, we shall refine the procedure and propose a generic approach for the prediction of eye movements on videos.

The work described here has previously been published in [Vig et al., 2011b].

4.1 Introduction

Over the last decades, much research has explored the factors that drive eye movements during the viewing of natural, real-world scenes. In the classical studies by Buswell [1935] and Yarbus [1967], the human gaze was primarily investigated using line drawings and static images. They noted that, although individual differences exist, viewers tend to consistently fixate the *semantically informative* regions when scanning a scene. Also, they found that fixation durations increase with increased viewing time and that viewing patterns are sensitive to the specific task the observer is performing. Surprisingly, colour has been shown to have little effect on viewing patterns. When free-viewing images, the distributions of both saccade amplitude and fixation duration are skewed, with an average amplitude around 2–4 deg and fixations durations of 330 ms on average with a mode at 230 ms [Henderson and Hollingworth, 1998].

While most work on gaze allocation in naturalistic scenes has dealt with static stimuli, the study of Itti [2005] was among the first to confirm on real-world complex videos that humans look at video regions of higher bottom-up salience than would be expected by chance. The authors found that motion and image transients are more predictive for eye movements than static features, such as colour, intensity, or orientation. Moreover, Carmi and Itti [2006] have shown on MTV-style video clips that dynamic visual cues can play an important causal role in drawing attention. ’t Hart et al. [2009] went further and used recordings of a mobile eye tracking setup to replay the visual input (during in- and outdoor exploration) in the laboratory, under head-fixed viewing conditions. The study showed that gaze recorded in the lab can predict reasonably well eye positions in the real world, but the temporal continuity of the scene is of importance. Tatler et al. [2005] were among the first to draw attention to the tendency of human subjects to fixate, in such eye-tracking experiments, more in the central part of the display rather than in the periphery. Tseng et al. [2009] quantified this phenomenon — the so-called *central fixation bias* — and linked it to the bias of the photographer to place the subject of interest in the centre of the image.

Eye movements have been collected and examined on a wide variety of dynamic realistic stimulus types (i.e. video categories). Gaze allocation has been studied while people watched Hollywood movies [Goldstein et al., 2007, Smith and Henderson, 2008], video games, or even driving scenes [Crundall et al., 2003, Underwood et al., 2005]. Dorr et al. [2010a] analyzed and compared the variability of eye movements on a range of different dynamic stimulus categories: natural videos, professionally-cut Hollywood trailers, and so-called “stop-motion” stimuli. Scanpaths were particularly similar across observers on Hollywood trailers, where, for example, frequent scene cuts elicited temporally coherent reorienting eye movements towards the screen centre. Briefly presented static snapshots from natural videos that were shown in their correct chronological order, however, proved not very representative of natural human viewing behaviour. A different series of studies, led by Michael Land, examined eye movements in a variety of everyday active tasks, such as driving [Land and Lee, 1994], food preparation [Land et al., 1999, Land and Hayhoe, 2001], and playing sports [Land and McLeod, 2000, Chajka et al., 2006, Russo et al., 2003].

Contrary to the above studies, in the following we will not focus on the spatial component of gaze allocation in dynamic real-world scenes. Instead, we will investigate the average time lag of eye movements in responding to dynamic attention-capturing events during the free viewing of natural or realistic videos. Despite the vast amount of research on anticipatory gaze behaviour in natural situations, such as action execution and observation, little is known about the predictive nature of eye movements when viewing different types of natural or realistic scene sequences. Here, we quantify this degree of anticipation while subjects freely view dynamic natural scenes. The cross-correlation analysis of image-based saliency maps with an empirical saliency measure derived from eye movement data reveals the existence of predictive mechanisms responsible for a near-zero average lag between dynamic changes of the environment and the responding eye movements. We shall also show that the degree of anticipation is reduced when moving away from natural scenes by introducing camera motion, jump cuts, and film-editing.

4.1.1 Anticipatory gaze behaviour

In Chapter 2, we argued that due to the anatomical structure of the eye, a sophisticated oculomotor system is needed to direct the fovea, the small high-resolution area of the retina, to regions of interest within the periphery. This is achieved by saccades — the rapid eye movements by which we shift our line of sight. However, the required neural processing introduces a certain delay until the oculomotor system reacts to a visual stimulus. In a typical laboratory setup, it takes about 200 ms until a saccade is made towards a spatially and temporally unpredictable target [Becker, 1991, Carpenter, 1981]. This delay can, in principle, obstruct immediate reaction to potentially critical events in everyday life. Yet, we are not hindered in our daily activities by this inherent lag in the visual feedback, most likely due to anticipation of the course of future events.

Early studies have shown the existence of predictive mechanisms if the target's spatial and temporal characteristics, such as amplitude, direction, and onset, are known a priori. For example, anticipatory saccades, with near-zero or even negative latencies, occur when the target systematically moves back and forth between two fixed locations [Findlay, 1981, Smit and Gisbergen, 1989]. A number of experiments investigating eye movements during natural interaction with the environment have found that the human visual system can benefit from expectations and prior knowledge about the surrounding world: eye movement patterns were examined during the performance of well-learned everyday tasks, such as tea- and sandwich-making [Land et al., 1999, Land and Hayhoe, 2001], hand-washing [Pelz and Canosa, 2001], and driving [Land and Lee, 1994]. These studies show that in everyday life eye movements are "proactive, anticipating actions rather than just responding to stimuli" [Land and Furneaux, 1997]. That is, saccades are often made to predicted locations of expected events even in advance of the event. However, these authors stress that eye movement patterns are highly task-specific: they seem to be influenced by some learned internal model of the performed actions [Hayhoe and Ballard, 2005, Land and Furneaux, 1997]. More recent experiments examined gaze patterns in more dynamic environments, during the execution of actions requiring specific physical skills. These studies confirm the proactive nature of eye movement control. For example, in the ball game cricket, experienced batsmen make high-precision anticipatory saccades to predict the ball's trajectory [Land and McLeod, 2000]. Similar results were reported when gaze patterns of elite-shooters [Russo et al., 2003] and experienced squash players [Chajka et al., 2006] were compared to those of novices. The main conclusion of these studies is that these predictive mechanisms may have evolved by learning the dynamic properties of the surrounding world (here, of the ball). These studies present evidence for predictive mechanisms during the execution of different natural tasks.

Furthermore, anticipation is found also during action observation. Experiments have shown that predictions are made also during the viewing of block stacking and model building tasks [Flanagan and Johansson, 2003, Gesierich et al., 2008, Mennie et al., 2007]. When subjects watch a block stacking task, their gaze anticipates the hand movements of the actor, as if they performed the task themselves.

Based on these findings, in this chapter we address the question: to what extent does the human visual system benefit from predictive mechanisms during the free viewing of dynamic natural scenes? Furthermore, how does the visual system adjust to different degrees of predictability? Our interest in these questions arose in connection with our work (described in detail in Chapter 5) on eye movement prediction in dynamic real-world environments. A critical issue, often neglected in the design of computational saliency models for eye movement prediction, is when exactly a salient location is fixated. Depending on the degree of predictability of a salient event, saccades may lag, coincide with, or even anticipate the event. Here, we quantify the average time lag between salient events in the natural scene and the eye movements responding to the events. Insights into these questions may have important implications for the design of computational models of saliency and of gaze guiding systems.

As shown in Chapter 2, most models of visual attention are built around the concept of a saliency map, which topographically encodes stimulus conspicuity [Koch and Ullman, 1985]. In the following, we will refer to these maps as “analytical saliency maps”, as they are computed analytically by means of local low-level image properties.

4.1.2 Outline of the approach

To measure the delay between events in a video and saccades towards these events, we temporally aligned analytical saliency maps with an “empirical” saliency measure based on real gaze data (see Figure 4.1 for a sketch of the analysis). According to our hypothesis, a dynamic event, such as the appearance of an object (e.g. the car on the left), would yield a local spatiotemporal maximum in our dynamic analytical saliency measures (middle row in Figure 4.1). After a certain time, any saccade made towards this dynamic event would, in turn, yield a local spatiotemporal maximum in the empirical saliency map (bottom row). To determine this time lag, analytical and empirical saliency can be cross-correlated, that is, multiplied when shifted against each other in time by varying amounts. Therefore, the time lag at which the cross-correlation function reaches its maximum denotes the average response delay.

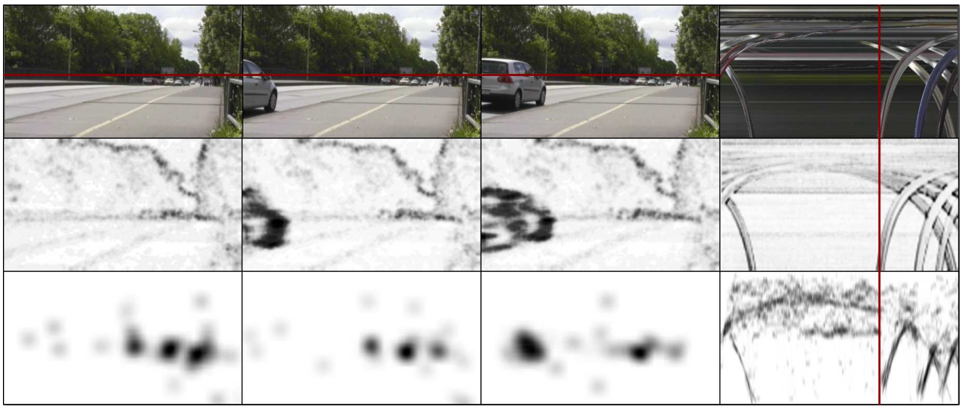


Figure 4.1: Different (x, y) and (x, t) slices of the spatiotemporal volume of a video and corresponding analytical and empirical saliency maps. **Top row:** Three neighbouring (but not consecutive) frames (i.e. (x, y) slices) of a video and a horizontal (x, t) slice of the movie cube at fixed $y = 400$ pixels (red horizontal line in the spatial screenshots). For the (x, t) slice, time axis is along the horizontal direction. Here, the time of the sudden appearance of the car is marked by the red vertical line. **Middle row:** Corresponding frames from the analytical saliency map (invariant K of the structure tensor). The sudden appearance of a car from left yields a strong response in the analytical saliency map. **Bottom row:** Empirical saliency map based on raw gaze samples of all subjects. Attention is drawn to the salient event (appearance of the car in the scene), but the eyes arrive at the target only after a certain time lag. Saccadic responses yield a spatiotemporal maximum in the empirical saliency map. The two saliency maps can be cross-correlated to determine the average time lag between the two maps.

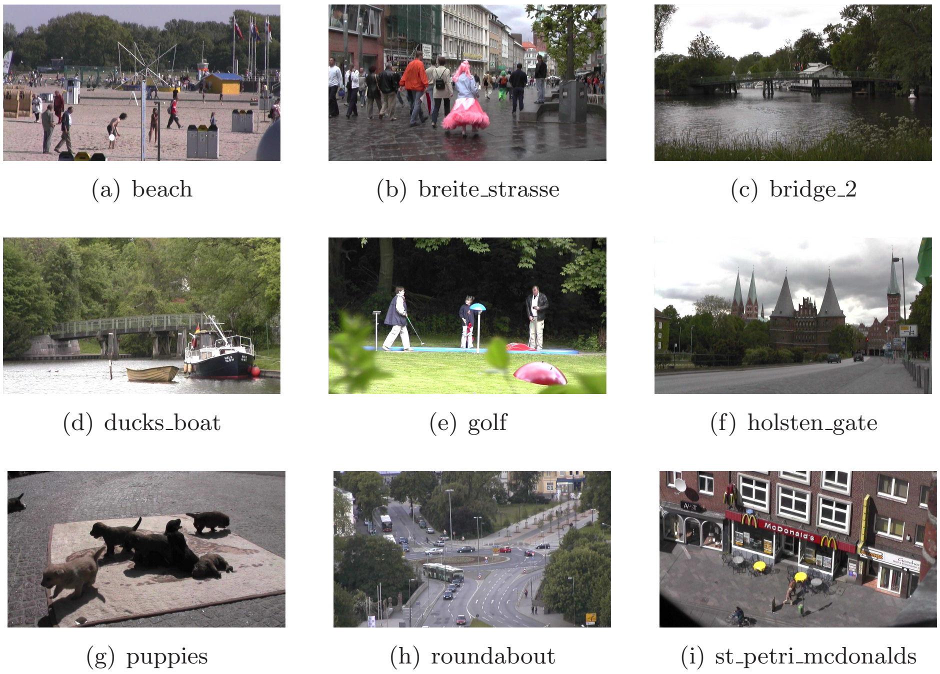
In our analysis, the geometrical invariants of the structure tensor presented in Chapter 3 serve as the analytical saliency measure to predict gaze-capturing events. As we shall see in the next chapter, the invariants (combined with machine learning) will prove to be simple and fast alternatives to state-of-the-art saliency algorithms (e.g. [Itti et al., 1998]). The invariant H, for instance, encodes spatiotemporal contrast, whereas K is (only) sensitive to dynamic content, i.e. encodes dynamic gaze-capturing events.

4.2 Methods

4.2.1 Stimuli and data collection

Natural dynamic scenes with static camera

In a free-viewing task, fifty-four participants (eight male, 46 female) watched eighteen high-resolution (HDTV standard, 1280×720 pixels, 29.97 Hz) natural outdoor video sequences with a duration of about 20 s each. During the recordings, the camera was held still; only four movies contained minor pan and tilt movements. The clips depicted real-world outdoor scenes in and around Lübeck: people in a pedestrian area (on the beach, playing in a park), populated streets and roundabouts, animals. Still shots from nine movies are shown in Figure 4.2. The videos were displayed at 45 cm viewing distance and at a visual angle of 48 by 27 degrees, so that the maximum spatial frequency of the display was 13.3 cycles per degree. The commercially available videographic eye tracker EyeLink II was used to record gaze data at 250 Hz. The experiment was conducted using two computers, the first of which was used to display the videos, while the second ran the eye tracking software. Recordings were performed in Karl Gegenfurtner's lab at the Dept. of Psychology of Giessen University. To synchronize gaze recording and video timing, the display of a new movie frame was signalled to the tracking computer with a UDP packet sent over a dedicated Gigabit Ethernet link; there, these packets were stored together with the gaze data using common timestamps by the manufacturer's software. From these recordings, about 40,000 saccades were extracted using a dual-threshold velocity-based procedure [Böhme et al., 2006]: to improve noise resilience, gaze velocity had to exceed a high threshold $\theta_1 = 137.5\,\mathrm{deg/s}$ to initiate saccade detection; saccade onset and offset then were determined by the first samples where



(a) beach

(b) breite_strasse

(c) bridge_2

(d) ducks_boat

(e) golf

(f) holsten_gate

(g) puppies

(h) roundabout

(i) st_petri_mcdonalds

Figure 4.2: Still shots from nine natural movies captured with a static camera.

gaze velocity rose above or fell below a lower threshold $\theta_2 = 17.5\,\mathrm{deg/s}$, respectively. Finally, several checks were performed for biological plausibility: minimal and maximal saccade duration and average and maximal saccade velocity (the reason being that impulse noise might lead to high sample-to-sample velocities). The data set is publicly available at [Dorr et al., 2010a] and `http://www.inb.uni-luebeck.de/tools-demos/gaze`.

Moving camera and edited videos

As a control data set, we use the CRCNS eye-1 database [Itti, 2004a] (available at `http://www.crcns.org/data-sets/eye/eye-1`), a benchmark data set for the analysis of eye movement data on complex video stimuli. The database consists of 100 video clips (640 × 480 pixels, 30 Hz) and the gaze data of eight human subjects freely viewing these videos. For our analysis, we used a subset of 50 clips and their corresponding eye traces called "original" experiment [Itti, 2004a, 2005]. The sequences include indoor and outdoor scenes, television broadcasts (commercials, sports, news, talk shows, etc.), and video games. Example still shots from six movies are shown in

PROFESSIONAL DRIVER. CLOSED COURSE
NFL Coach Hank Stram
HeadlineNews
Kuwait Ambush 6 more arrests made in
Tue attack on US businessmen; 1 killed
NHL

(a) beverly06 (b) gamecube04 (c) standard05

(d) tv-ads03 (e) tv-sports01 (f) tv-news02

Figure 4.3: Still shots from the CRCNS eye-1 video set.

Figure 4.3. In case of all videos, transitions between shots are achieved by camera movements, such as panning, tilting, and zooming. Besides these, transitions are realized in television clips (23 out of 50) through jump cuts and special video editing effects, such as fading, dissolving, and wiping. Text overlays are also common. The total number of saccades extracted from the raw gaze data with the aforementioned saccade detection procedure was about 11,000.

4.2.2 Analytical saliency measures

In search of saccade triggering stimuli, we use a simple measure to detect salient events in the video. It is well known that the visual signal needs to change over space and time to capture attention (e.g. we tend not to like blank walls). Therefore, a simple assumption one can make is that the more the visual signal changes, the more salient it is. As shown in Chapter 3, the degree to which a spatiotemporal signal changes is qualitatively well described by the intrinsic dimension of the signal and we use this concept as a simple measure of saliency. Obviously, such a simple assumption cannot be sufficient to explain the complex nature of eye movements, but it works well enough for our purpose, i.e. to align eye movements with attention-capturing events in the video.

To estimate our analytical saliency measure, the intrinsic dimension, we computed the geometrical invariants H, S, and K of the structure tensor $\mathbf{J}$, which in Chapter 3 were shown to yield information about the minimum intrinsic dimension of a video region. To improve noise resilience, we performed our analysis on a lowpass-filtered video (6.6 cycles/degree) that was created by filtering the video with a 5-tap spatial binomial filter and downsampling it (in space) by a factor of two.

To obtain the structure tensor $\mathbf{J}$ (see Equation 3.11), partial derivatives were calculated by first smoothing the input video with spatiotemporal 3-tap binomial kernels, and then applying $[-1, 0, 1]$ kernels to compute the differences of neighbouring pixel values. The smoothing of the products of derivatives (with Ω) was done with another spatiotemporal 3-tap Gaussian. In principle, pooling these derivatives over a larger spatiotemporal neighbourhood is desirable for a robust computation of the structure tensor $\mathbf{J}$, but for the present analysis, localized responses were more important than robustness against noise.

In addition to being symmetric, the above filter kernels are centred at the detected events, i.e. are non-causal. Note that with a non-causal filter, the output can anticipate the next event. For our purpose, however, a non-causal filter is more appropriate as its output is maximal at the time of the event, whereas the maximum response of a causal filter would be lagging behind the event and, therefore, would decrease the separation between the analytical and empirical saliency measures.

One might argue that for registering the temporal events with eye movements, it would suffice to consider simple temporal differences only. Indeed, no substantial differences are expected when cross-correlating gaze responses either with the original videos or with their saliency maps. Nevertheless, to keep the noise level low, we prefer to register those spatiotemporal events that are characterized by a high degree of predictability.

4.2.3 Empirical saliency measures

In eye movement research, eye tracking data typically comes in the form of gaze coordinates. An important question is how to represent and evaluate such data. One way to begin is to look at simple parameters such as mean fixation duration and average saccade amplitude. However, often a more complex analysis is required. For instance, when comparing eye movement traces, one may want to compensate for possible imprecisions in both the eye tracking and human visual system. Also, eyes are typically directed at particular regions of interest, not at single points. Wooding [2002] suggested a means to convert the raw fixation data into a *fixation density map*, providing a useful tool to e.g. quantify the similarity of eye traces and compare (or correlate) discrete gaze coordinates with feature (or saliency) maps. A fixation density map, also referred to as *empirical saliency map*, is constructed by the superposition of Gaussians (to account for imprecisions) centred at each gaze sample. A subsequent normalization step turns this map into a probability density map in which regions of interest of human observers are represented.

Average scanpaths (fixation density map)

We defined our first type of empirical saliency measure as the *density of the gaze points* averaged over all subjects. These probability maps were computed for each video, by placing two-dimensional spatial Gaussians at each fixation location of all subjects, similarly to the well-known fixation density distribution [Wooding, 2002]. The Gaussian kernels had a spatial support of about 4.8 degrees of visual angle and a standard deviation σ of 0.25. The superposition of these Gaussians resulted in the empirical saliency map. Example still shots from a fixation density map are given in the last row of Figure 4.1. In that example, eye traces of 54 subjects were used to create the map.

Average saccades (saccade density map)

In the standard approach, all raw gaze samples are used for creation of the empirical saliency map, which includes samples throughout or even to the end of fixations, although ultimately we are interested only in fixation onsets. Therefore, we also created much sparser empirical saliency maps with the above parameters but using only the *saccade landing points* of all subjects.

Single saccades

As the traditional empirical saliency map contains gaze data of several viewers, saccadic responses to a certain salient event might arrive at slightly different times within a short time interval. How does this influence our analysis? To gain a deeper understanding of the underlying causes, we also examined the average time lag of individual saccades in responding to changes in the visual scene. For each saccade, we created a sparse response map (similar to the empirical saliency), by placing a single two-dimensional Gaussian at the endpoint of the saccade. Individually, saccade landing points are more prone to noise than the full empirical saliency map. However, they are more localized in space-time and in such a large number of samples (about 40,000 saccades in the first data set of natural outdoor scenes), noise should cancel out.

4.2.4 Normalized cross-correlation

Our analysis is based on the cross-correlation of the above described analytical and empirical saliency maps shifted, relative to each other, in the time domain [Gonzalez and Woods, 2001]. The normalized cross-correlation function (*ncc*) between two spatiotemporal signals f and g is defined as

$$ncc(f,g,\tau) = \frac{\sum_{x,y,t}(f(x,y,t)-\bar{f})\cdot(g(x,y,t+\tau)-\bar{g})}{\sqrt{\sum_{x,y,t}(f(x,y,t)-\bar{f})^2\cdot\sum_{x,y,t}(g(x,y,t+\tau)-\bar{g})^2}}$$

where τ is the temporal offset and $\bar{f}$ and $\bar{g}$ stand for the DC components (means) of the two signals. *ncc* was computed for each analytical and empirical saliency map pair.

To determine the correlation expected by chance, as a control condition, we randomly paired analytical and empirical saliency maps of different movies and proceeded as above. This shuffling of scanpaths and videos among each other is a standard procedure in relating low-level image features to gaze data [Reinagel and Zador, 1999, Tatler et al., 2005].

4.3 Results

In our analysis, we shifted the empirical saliency relative to the analytical saliency one frame per temporal unit (approximately 33.367 ms in the first movie set of natural outdoor scenes with static camera and 33.333 ms in the CRCNS eye-1 movie set), within a range of 61 (± 30) frames. Here, we are less concerned with the absolute values of the correlation coefficient obtained for the different time lags, but with the value of the time shift at which the maximum correlation occurs. We will mainly restrict ourselves to invariant K, which encodes three-dimensional changes and is therefore most informative to the human visual system in an information-theoretic sense. However, we shall show that the observations also hold for the other invariants.

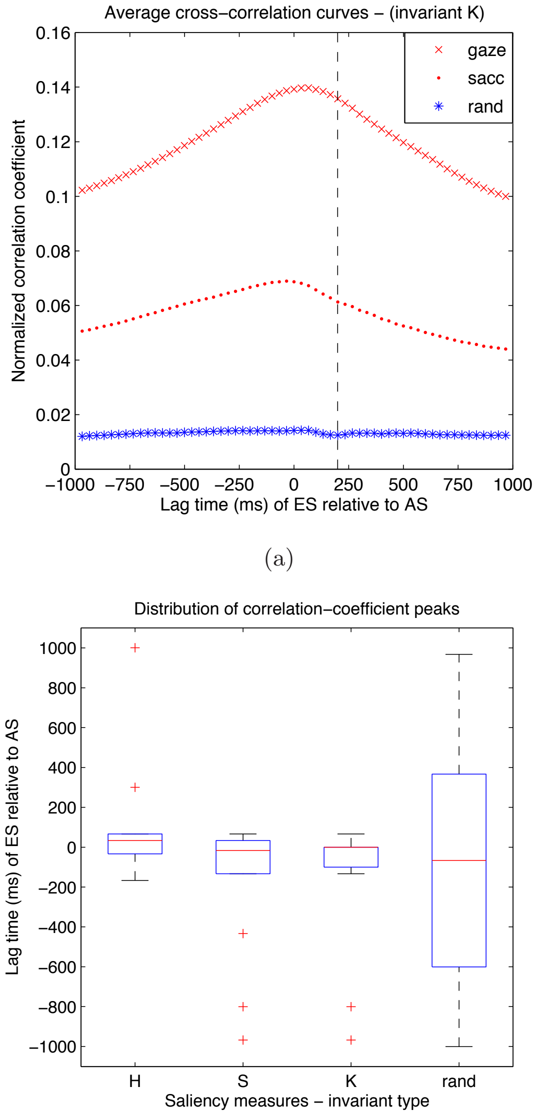
4.3.1 Natural dynamic scenes with static camera

Figure 4.4 summarizes results obtained for cross-correlating the analytical with the average empirical saliency map of all subjects. Mean correlation coefficients for the eighteen movies are plotted against the frame shift in Figure 4.4(a) (red cross curve). A positive lag of t ms indicates that the empirical saliency map follows the invariant movie by t ms. The maximum of the averaged coefficients, for correlating invariant K with the empirical saliency map ("average scanpaths" case), is detected at a lag of 66.73 ms (i.e. two frames).

As can be expected, the maximum is slightly shifted in time to the left when the fixation data is discarded ("average saccades" case – dotted curve in Figure 4.4(a)). In this case, quite surprisingly, highest average correlation is found at -33.36 ms, i.e., on average, the empirical saliency map was ahead of the analytical map by one frame. In both conditions (empirical saliency based on raw gaze samples and on saccade endpoints only), mean correlation curves have a Gaussian-like shape and a pronounced peak, whereas randomly pairing and then correlating analytical and empirical maps of different videos yields a flat curve (blue asterisk curve in Figure 4.4(a)).

In the following, we will restrict our considerations to an empirical map based on saccade endpoints only, because, as results suggest, raw gaze data introduces further undesired shifts in the eye movement response.

The box plot in Figure 4.4(b) shows the distribution (over the eighteen movies) of time shifts at which maximum correlation was measured. Here, we compare the distributions of correlation peaks obtained for the three invariants, H, S, and K, and for the random analytical–empirical pairing case. As already expected, for the invariants, the peaks are all centred around 0 ms with only few exceptions (red crosses in Figure 4.4(b)). For example, for invariant K, the correlation peaks of two movies are identified at very large negative offsets, meaning that the response in the empirical saliency preceded the signal by an unrealistically large amount of time. An inspection



(a)

(b)

Figure 4.4: The empirical saliency map (ES) is offset (with respect to the analytical saliency map – AS) along the time dimension by one frame (33.367 ms) per temporal unit within a predetermined range (± 30 frames). A correlation coefficient is calculated for each individual frame shift. **(a)** Average correlation coefficients over all movies are plotted against the frame shift (red cross: ES based on average scanpaths, red dot: ES based on average saccades, blue asterisk: random AS – ES pairing). The dashed vertical line represents the normal mean value of the saccadic reaction time (in the order of 200 ms) to unpredictable targets [Becker, 1991]. **(b)** Box plot comparing distributions of correlation-peaks over the movie set for the AS measures H, S, K, and random AS – ES pairing (middle line: median, box: upper and lower quartile, whiskers: data extent, plus sign: outliers).

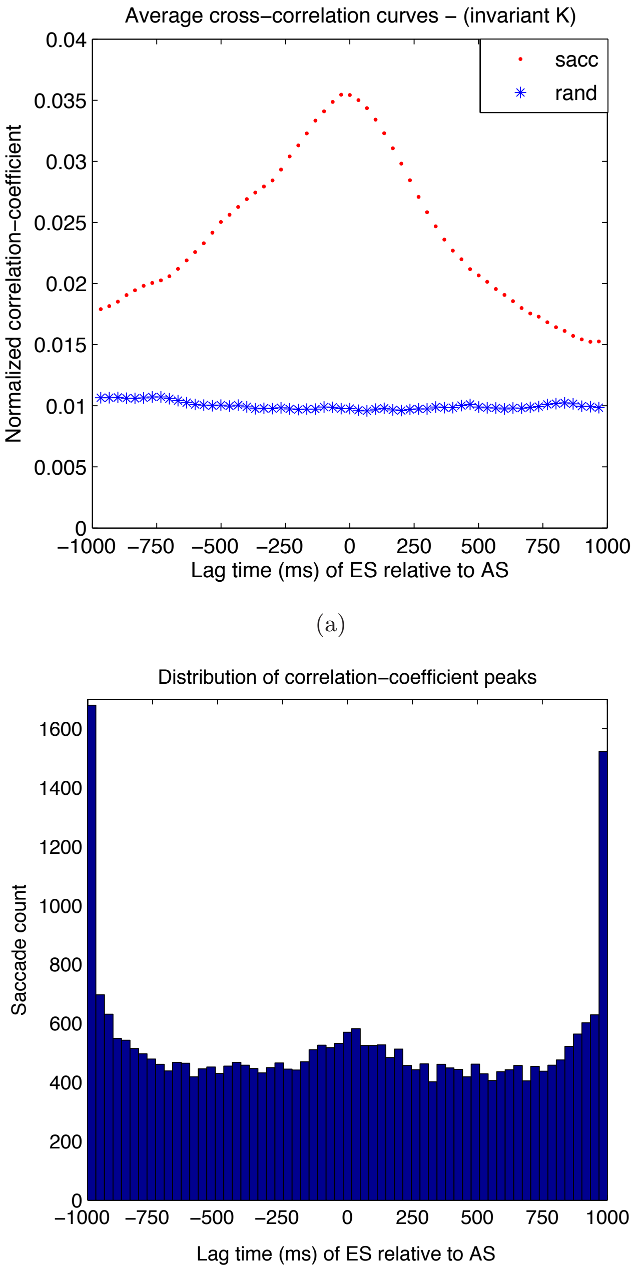
of the shape of the individual correlation curves indicates that these two curves are flatter than those of the other movies, with no pronounced peaks. Indeed, a closer look at the content of these movies reveals that they are of almost still-life character (e.g. unpopulated bridge) and so, as invariant K is only sensitive to dynamic content, it is not surprising that the correlation curves have no distinctive peaks. In the following, unrealistically large positive and negative shifts are considered outliers.

Overall, we found that the three distributions of the invariants are very similar with a median of one frame (33.367 ms) for invariant H, -16.68 ms for invariant S and 0 ms for K. Unlike the concentrated distributions of the invariants, the lags at which maximum correlation occurs in the random pairing case are scattered throughout the correlation window.

Results for correlating invariant K with individual saccades are shown in Figure 4.5. The maximum of the average correlation curves of all saccades is here, too, identified at -33.367 ms (see the peak of the red dotted curve in Figure 4.5(a)). The average correlation curve has again a pronounced peak when compared to the curve of the control condition. However, the peak around 0 ms in the distribution of the time shifts with maximum correlation (in Figure 4.5(b)) is not very distinctive. This is again due to low values of K resulting in a flat correlation curve with no pronounced peaks. To measure the curves' "flatness", we used the following simple measure: in Figure 4.6(a) we sorted the curves according to the difference of maximum and minimum correlation values over the frame shifts. The more curved the correlation line, the larger this difference. Indeed, when plotting, in Figure 4.6(b), only the distribution of saccades for which this difference exceeded the mean of all differences (i.e. 0.26), the peak becomes more prominent. Nevertheless, this simple measure cannot eliminate outliers, such as peaks at implausibly large time offsets.

4.3.2 Moving camera and edited videos

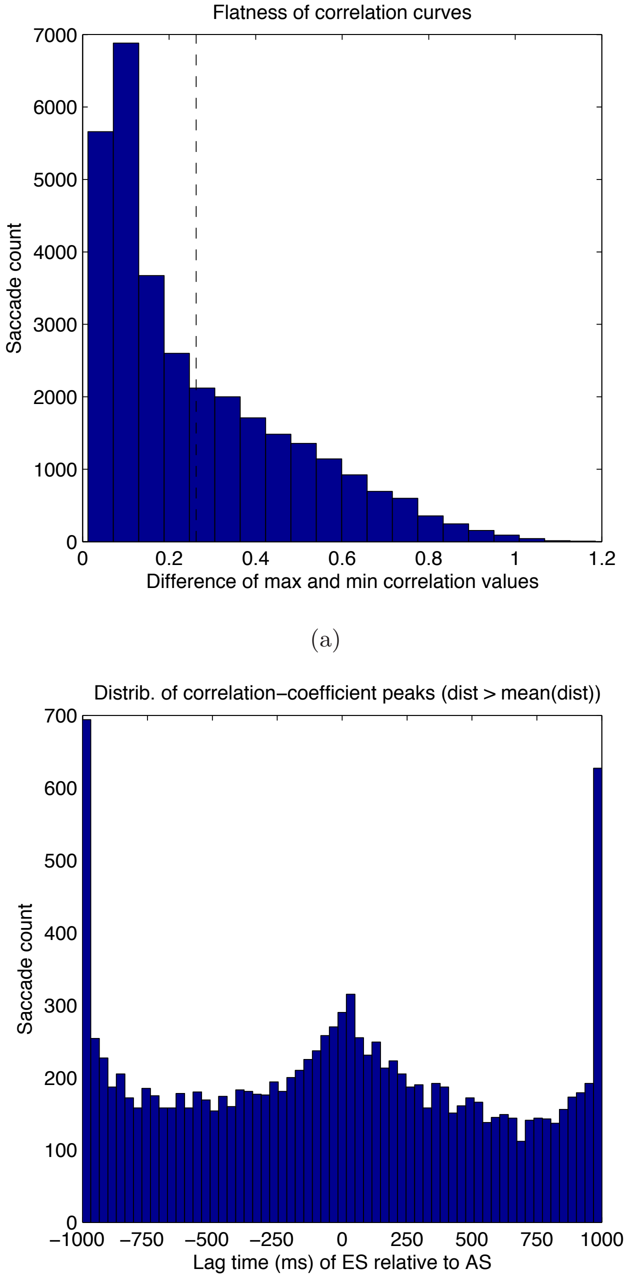
Next, we compare these findings with those obtained on the CRCNS eye-1 data set. When cross-correlating invariant K with individual saccades, a noticeable shift is observed in the location of the peak of the mean correlation coefficients (red dotted curve in Figure 4.7). The correlation maximum is here identified at about 133.33 ms (four frames). This larger average time shift could, however, be explained by the fact that a significant number of



(a)

(b)

Figure 4.5: Individual saccade landing points cross-correlated with the analytical saliency map K. **(a)** Average correlation coefficients over all saccade endpoints (red dot: landing points, blue asterisk: shuffled locations). Peak identified at $-33.36\,\text{ms}$ (-1 frame). **(b)** Distribution of time shifts (over all saccades) with maximum correlation.



(a)

(b)

Figure 4.6: Individual saccade landing points cross-correlated with the analytical saliency map K. **(a)** Histogram of the distribution of saccades sorted according to the difference between the correlation curves' extreme points. A threshold is set at the mean of the differences removing around 60 per cent of saccades with a flatness measure smaller than the mean measured "flatness". **(b)** Distribution of correlation peaks of curves after thresholding.

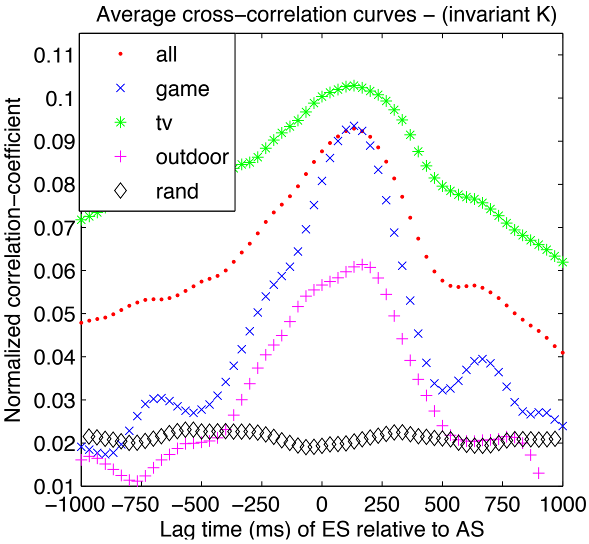


Figure 4.7: Mean correlation curves when cross-correlating individual saccades with the invariant K (CRCNS eye-1 data set). For the "original" experiment: averaging over correlation curves of all movies (red dot), computer game videos (blue cross), TV-clips (green asterisk), outdoor scenes (magenta plus sign), and randomly shuffled locations (black diamond).

the clips (television broadcasts and quasi-realistic computer game scenes) are physically quite different from real-world natural scenes. Jump cuts, camera movements, and movie-editing techniques introduce unnatural temporal discontinuities which could entail delayed oculomotor responses. For instance, movie cuts elicit reorienting saccades towards the centre of the screen [Dorr et al., 2010a]. To further investigate whether the presence of camera motion and movie-editing techniques affects average response delays, we categorized the fifty movie sequences into three groups based on stimulus type: TV-broadcasts (23 clips), computer games (9 videos), and outdoor scenes (17 sequences; parks, crowds, rooftop bar). Note that the outdoor scenes, too, were captured using basic camera movement techniques (i.e. tilt, pan, and zoom). We excluded from our analysis a synthetic clip of a disc drifting on a textured background. Average cross-correlation curves of the three stimulus groups are plotted in Figure 4.7. Although the three curves reach their maximum at very similar time shifts (at about 133.33 ms), notice the difference in how peaked the curves are. The correlation curve of the quasi-realistic computer game stimuli is the most sharply peaked, whereas

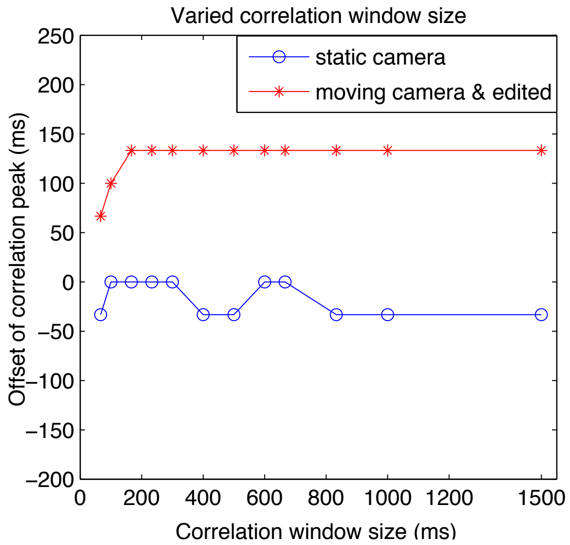


Figure 4.8: Offsets of correlation curve peaks when the correlation window size is varied. Individual saccades were cross-correlated with the invariant K. Blue circle: first movie set of outdoor scenes with static camera, red asterisk: moving camera and edited videos.

the curve of the more natural outdoor scenes reaches a plateau at around -66.66 ms after which only limited increase occurs. Considering that we are looking at averages of several individual correlation curves, a pronounced peak and high correlation values (e.g. as in the case of computer games) suggest that the majority of the underlying individual curves reach their maximum at roughly the same time lag. In case of the outdoor scenes, however, the distribution of time shifts at which a peak occurs is more scattered; therefore averaged coefficient values are lower and the maximum is not very pronounced.

Finally, we show that the size of the sliding window has no impact on the outcome of the correlation analysis. In Figure 4.8 we plotted, for various correlation window sizes, the offsets of the peaks (of mean correlation curves) for the two movie sets (blue circle – dynamic scenes with static camera, red asterisk – moving camera and edited videos). When the window is smaller than the actual optimal offset, the peak is detected at the border of the correlation window, otherwise the curves are almost flat, i.e. offsets are consistent, with only small fluctuations of one frame in case of the dynamic scenes with static camera. Here, peaks were detected at an offset of either -1 or 0 frames.

4.4 Discussion

An often neglected question in the design of computational models of saliency is what the typical response lag is to changes in the visual scene. The choice of a specific value is typically motivated by laboratory investigations of saccadic response latencies to synthetic stimuli. In [Carmi and Itti, 2006] for instance, authors manually choose a particular latency that agrees with the timing of human saccades in the context of a synthetic test clip. However, depending on the stimulus type, the average lag can vary quite substantially: in [Land and Furneaux, 1997], authors distinguished between “reactive saccades of the laboratory” (having positive lags) and “proactive saccades of normal life” (with near-zero or even negative lags). Here, we aimed to infer the mean response delay, in laboratory settings under head-fixed viewing conditions, when free-viewing dynamic natural scenes. Using cross-correlation analysis of analytical saliency maps — encoding saccade-triggering changes in the video — and spatiotemporal fixation maps — encoding eye movement responses to the salient events —, we identified the time shift at which the two maps have the maximum correlation. We then averaged results over several movies or individual saccades to determine the mean lag in the stimulus class of natural videos. In addition, we examined whether this average response delay differs from that obtained on similar natural and quasi-natural (video game) stimuli, which were captured using basic camera movement techniques and, depending on the movie type, post-processed with video-editing software.

In the first data set of dynamic natural scenes, we found a near-zero mean lag, meaning that, on average, reactions to salient events coincided with or even slightly preceded the events themselves. This result was consistent for all analytical saliency maps (invariants H, S, and K) and both when scanpaths of all subjects and individual saccades were cross-correlated with analytical maps. This somewhat surprising finding may be attributable to an adaptation of the human visual system to the environmental dynamics of the surrounding world. Most dynamic events in natural scenes are, at least to some extent, predictable. Such anticipatory mechanisms (e.g. looking ahead of the movement) imply some sort of scene knowledge of the dynamic characteristics of the environment that is due, for instance, to experience with the physical laws of motion.

In line with the studies on task-specific gaze control [Hayhoe and Ballard, 2005, Flanagan and Johansson, 2003], one could also speculate that, during the viewing of a particular scene, observers might identify certain higher-level (hidden) tasks and actions, such as playing beach ball, walking on a bridge, driving in a roundabout. If we think of the free viewing of natural videos as action observation of what is happening in the video, possessed knowledge about these actions could possibly generate anticipatory gaze behaviour.

Note that we are here not aiming at explaining gaze behaviour with a simple bottom-up model but merely at measuring the time lag between events in the video and the responding eye movements. We use a plausible model of bottom-up saliency simply to improve the measurement of this time lag. In other words, our bottom-up saliency model based on the invariants merely serves as an “event detector”. The fact that this time lag is small can indeed be attributed to top-down mechanisms but our result does not depend on such interpretations. The anticipation that we find can be due to many different predictive mechanisms starting from very simple (low-level) models, such as a Kalman filter, to more complex (high-level) ones, such as action planning. Given this possible continuum of mechanisms of increasing complexity, it seems unnecessary to draw a “bottom-up top-down borderline”.

The analysis of the second set of complex stimuli (CRCNS eye-1 data set) reveals a longer average delay of about 133 ms between a dynamic event in the scene and saccades responding to it. We argue that, due to the presence of jump cuts, camera motion, and other movie-editing techniques, the amount of bottom-up influence in these stimuli is, on average, higher than in truly natural scenes. The introduced temporal discontinuities and the sudden appearance of text overlays in television broadcasts trigger a high number of reactive saccades. Similarly, to passive observers, the moves of the video game character are less predictable than to the game player himself. Looking at the average correlation curves of the three video subsets (TV-clips, games, and outdoor scenes), the curve of the outdoor scenes pops out. Its global maximum is identified shortly after the overall average of

133 ms but mean correlation values are comparably high already beginning with −66.66 ms. This could suggest that, in comparison with the first set of natural outdoor movies in which the great majority of saccades were rather predictive (therefore, the peak shortly before zero), here the ratio of visually guided and anticipatory saccades is more balanced.

Computational models of attention either assume no time shift between their analytical saliency maps and the responding eye movements, or they do not try to optimize this value but use subjective observations [Carmi and Itti, 2006]. We argue that by introducing an artificial time lag adjusted to the stimulus type (i.e. eliciting maximum response in the analytical saliency at the time of the expected gaze response, not at the time of the event), saliency models significantly increase their performance in predicting eye movements. As an alternative, temporal uncertainty could be introduced in the model in order to account for the different stimulus-specific time lags [Vig et al., 2009].

The findings of this chapter are also highly relevant for our work on integrating gaze into future visual and communication systems by measuring and guiding eye movements. In such a scenario, the right timing of the so-called gaze-capturing events is critical for achieving the desired effect. In other words, for attention to be drawn to a specific movie region at a specific time, the temporal placement of the gaze-capturing event must take into consideration the stimulus-specific average response lag.

4.5 Chapter conclusion

In summary, in this chapter, we have characterized a special class of visual stimuli, namely, that of real-world natural scenes, in terms of the typical time lags between salient changes in the scene and the responding eye movements. To measure this typical time lag, we temporally aligned analytical spatiotemporal saliency maps with response maps encoding saccadic reaction to the salient events. We argue that the near-zero average lag could be attributable to an adaptation of the human visual system to the — often predictable — dynamics of the environment. We have shown that the degree of anticipation is reduced when moving away from natural scenes by introducing cuts, camera motion, and film editing. Finally, we suggest that the stimulus dependent mean response lag should be an important consideration in the design of computational models of visual saliency and gaze guiding systems, and provide a method for computing the average time shift between movie events and eye movements.

5
Prediction of eye movements on natural dynamic scenes

Since visual attention-based computer vision applications have gained popularity, evermore complex, biologically-inspired models — such as those reviewed in Chapter 2 and Appendix B — have been developed to predict salient locations or interest points in naturalistic scenes. However, it is well-known from machine learning theory that too much complexity can lead to overfitting and poor generalization performance. In this chapter, we therefore explore how far one can go in predicting eye movements by using only basic signal processing, such as image representations derived from the efficient coding principles presented in Chapter 3, and machine learning. To this end, we begin with simple single-scale saliency maps computed on grayscale videos and then gradually increase the complexity of our model to spatiotemporal multiscale and multispectral representations. Using a large collection of eye movements on high-resolution videos, supervised learning techniques fine-tune the (relatively few) free parameters whose addition is inevitable with increasing complexity. The proposed model, although very simple, demonstrates significant improvement in predicting salient locations in naturalistic videos over four selected baseline models and two distinct data labelling scenarios. Furthermore, we also evaluate the impact of the different labelling scenarios, which is a novel contribution as well. Finally, we show that our model can be extended to successfully predict eye movements even on transparently overlaid movies.

Parts of the work described in this chapter have previously been published in [Vig et al., 2009, 2010b, 2012, Barth et al., 2010, Dorr et al., 2010b, Vig et al., 2010a].

5.1 Motivation

As we have discussed in Chapter 2, computational saliency models range in complexity from few-parameter, empirical models to more complex, multi-parameter ones. We have seen that while evermore complex models seem to be needed to better predict gaze behaviour on realistic scenes, there are also a few counterexamples to the trend [Kienzle et al., 2007b, Guo and Zhang, 2010].

This thesis contributes to this latter line of research by exploring the potential of saliency models that make as few premises as possible. Once we have established such baseline, we can then investigate (and quantify) the potential gain from gradually increasing complexity. We propose to go back to the basics of signal processing to obtain efficient image representations (such as those presented in Chapter 3), and, if required, utilize powerful learning algorithms on these representations to predict visual saliency in videos. Specifically, we begin with the simple observation that many video regions, such as homogeneous areas, are highly redundant, and that it is local changes, i.e. intensity variations (along edges, corners, etc.), that are informative. As shown in Chapter 3, the degree of this signal redundancy can be mathematically described by the *intrinsic dimension* of a region, and we here use this concept as a simple measure of saliency. In order to further tune the model parameters so as to predict bottom-up attention on complex scenes, we adopt data-driven machine learning techniques. However, given the high dimensionality of a pixel-based video representation, current learning algorithms would require very large amounts of data and thus have only limited practical applicability. Even with only a moderate number of training data, i.e. human fixations on videos, we here overcome the curse of dimensionality through dimensionality reduction (specifically by spatial pooling of features). This allows us to incorporate more information, e.g. from multiple spatiotemporal scales. Furthermore, the concept of intrinsic dimensionality naturally leads to a unified representation of spatial and temporal saliency, such that no fusion of separate static and dynamic maps is required (as in the case of models derived from the Feature Integration Theory, see Section 2.2.4). Similarly, the definition can be extended to multispectral sequences, so that it becomes no longer necessary to combine separate saliency maps from each colour channel. In order to test the performance of our model, we use the large data set of human fixations on the diverse collection of high-resolution outdoor videos (captured with a static camera) presented in Chapter 4. Since top-down processes strongly modulate gaze behaviour, obviously, we cannot expect any bottom-up model to fully account for the complex nature of attentional orienting. Nevertheless, we shall show that our simple assumptions already account reasonably well for eye movements during free-viewing of dynamic real-world scenes. Indeed, the proposed simple approach shows significant improvement over several state-of-the-art models of bottom-up saliency, which base their prediction on numerous assumptions on perceptual processes and incorporate several basic features. Through a systematic analysis, we shall also set out to quantitatively evaluate the gain from more complex features by gradually extending a simple single-scale saliency map computed on the intensity videos to a multiscale and multispectral model. Our results support the (intuitive) assumption that a higher degree of variation in the visual signal leads to higher saliency.

The remainder of this chapter is organized as follows. In the first and main part, we describe and empirically validate the major contribution of this thesis: the generic yet powerful saliency predicting framework derived from the simple assumption that the degree of local signal variation is related to informativeness (and thus, salience) of an image region. We start by describing the computational steps of the above outlined simple and efficient algorithm for bottom-up saliency. Then, in Section 5.3, we demonstrate its performance in predicting human fixations on the 18 natural videos of the previous chapter. There, we shall prove the validity of the approach for two different data labelling scenarios (Section 5.3.2), discuss implementation issues (Section 5.3.3), and present a systematic analysis of how the choice of free parameter values affects prediction performance (Section 5.3.4). In Section 5.3.5, we compare our results to those of four baseline models for bottom-up saliency. Finally, in Section 5.4, we interpret the results and summarize the major findings.

In the second part of this chapter (Section 5.5), we shall extend and apply our basic saliency to two further application scenarios. First, we examine the contribution to bottom-up saliency of spatiotemporal intensity variation along different subspaces of lower dimensionality (i.e. different combinations of the x, y, t axes). The findings of this work are particularly relevant for machine vision systems with limited computing resources. Then, to conclude this chapter, we shall extend our saliency prediction framework to predict eye movements on transparent overlaid movies based on the symmetric invariants of the *generalized* structure tensor.

5.2 Model description

An outline of our eye movement prediction approach is schematically illustrated in Figure 5.1. Before delving into details, we first provide a brief overview of the model structure and the main computational steps. Here, we learn the structural differences between salient and non-salient video locations on simple video representations (reviewed in Chapter 3) that characterize different types of spatiotemporal intensity changes. Given a collection of image sequences and a large set of recorded eye movements on them, we label areas in the videos as either salient or non-salient. For each video, we compute low-level feature maps that encode the intrinsic dimensionality of video regions. Such maps are computed on several spatiotemporal levels of multiresolution image pyramids. In a neighbourhood around each location (be it salient or not), we extract the *feature energy* from these maps: the root-mean-square of the pixels in the spatiotemporal neighbourhood. Feature energy (a single scalar) is computed on each pyramid level; thus, each location is described by a low-dimensional vector whose components are the energy values on different scales. Such feature energy vectors are finally fed into a classifier (a Support Vector Machine), which learns a mapping between feature energy vectors and the saliency level of a certain location. In the following, we shall describe the above steps in greater detail.

5.2.1 Multiscale geometric invariants

To characterize different types of local spatiotemporal variations, we use the structure tensor-based image representations reviewed in Chapter 3. The scale on which the intrinsic dimension is estimated depends on the bandwidth of the Gaussian smoothing function Ω and of the derivative operators (see Eq. 3.11). Therefore, we extract the geometrical invariants H, S, and K (of both grayscale and multispectral videos) on an *anisotropic Gaussian pyramid* with nS spatial and nT temporal levels. A detailed description of

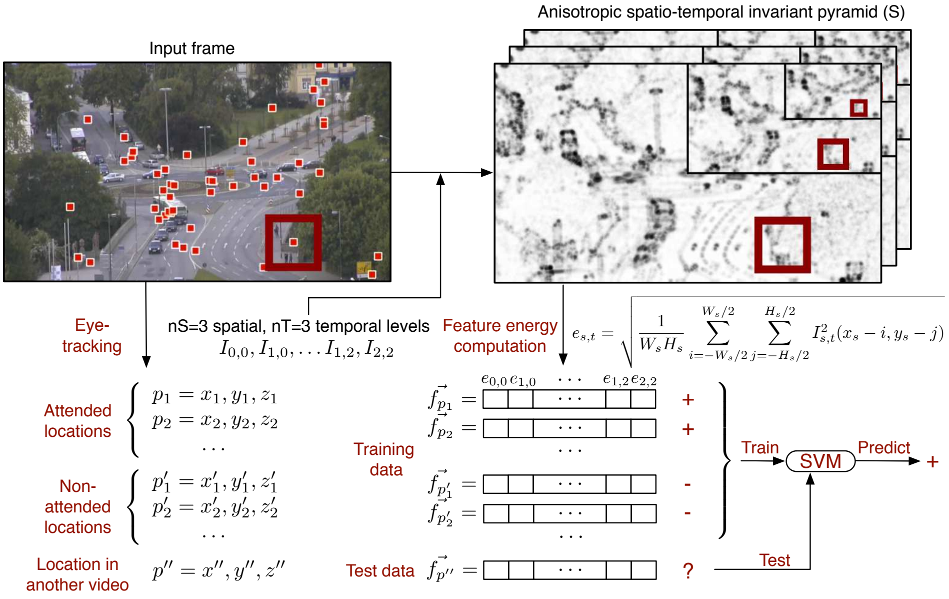

Anisotropic spatio-temporal invariant pyramid (S)
Input frame
Eye-tracking
nS=3 spatial, nT=3 temporal levels
$I_{0,0}, I_{1,0}, \dots I_{1,2}, I_{2,2}$
Feature energy computation
$e_{s,t} = \sqrt{\frac{1}{W_s H_s} \sum_{i=-W_s/2}^{W_s/2} \sum_{j=-H_s/2}^{H_s/2} I^2_{s,t}(x_s - i, y_s - j)}$
Attended locations
$p_1 = x_1, y_1, z_1$
$p_2 = x_2, y_2, z_2$
Non-attended locations
$p'_1 = x'_1, y'_1, z'_1$
$p'_2 = x'_2, y'_2, z'_2$
Location in another video
$p'' = x'', y'', z''$
Training data
Test data
$e_{0,0}\, e_{1,0} \dots e_{1,2}\, e_{2,2}$
$\vec{f_{p_1}} =$
$\vec{f_{p_2}} =$
$\vec{f_{p'_1}} =$
$\vec{f_{p'_2}} =$
$\vec{f_{p''}} =$
+
+
-
-
?
Train
SVM
Predict
+
Test

Figure 5.1: Flow diagram of our approach. Using eye tracking data (fixations denoted by small red squares in the movie frame), we label movie regions as attended or non-attended. Image features (the geometrical invariants) are extracted on multiple scales of an anisotropic spatiotemporal pyramid. For a neighbourhood (large unfilled square shown schematically) around each location, the average feature energy is computed on each scale of the spatiotemporal pyramid. An SVM is trained on the obtained energy vectors and is then used to predict whether a new location will be attended or not.

such multiresolution Gaussian image pyramids is given in Section 3.3.

5.2.2 Dimensionality reduction

The saliency of a video location is strongly influenced by its spatiotemporal context. Centre-surround models exploit this property when they define saliency as the ability of some features to best discriminate between image structure in a centre and a surround window. Besides, in a data-driven approach, where fixational data is utilized to tune the model parameters, one also has to compensate for possible inaccuracies in both the eye tracking and the biological system. The size of the spatiotemporal neighbourhood that needs to be considered is still a matter of debate in the human vision community. While some studies use windows of the size of the high-resolution centre of the retina, the fovea (2–3 degrees), one can also optimize it with respect to the available eye movement data. Learning in the pixel space determined by the number of pixels of the neighbourhood is often problematic as the feature space dimensionality of a reasonable sized image patch, e.g. 64 by 64 pixels ($2.5 \times 2.5\,\mathrm{deg}$) grows rapidly (more than 4000 dimensions). In such a scenario, given a limited number of training data, the effects of the “curse of dimensionality” seriously degrade classification performance. Due to such constraints, the learning algorithm in [Kienzle et al., 2007b], for instance, was restricted to a single spatial scale.

In order to tackle the above problem and allow incorporating information from multiple scales, we perform a *spatial pooling*: we reduce pixel information in a window around the location to a single scalar, by taking the root-mean-square of the feature values (i.e. geometrical invariants) in the window. As a result of such pooling, an invariant representation of the local neighbourhood originates. This allows us now to compute such *feature energy* on every scale of the above multiresolution pyramids, as the dimensionality is still kept low. Here, we use a spatial neighbourhood only, as the uncertainty induced by measurement errors and saccade imprecision is higher in the spatial domain than in the temporal one.

More formally, for a movie location $p = (x, y, z)$ (with spatial coordinates x and y, and frame number z), we compute a vector

$$\mathbf{f_p} = (e_{0,0}, e_{0,1}, \cdots, e_{nS-1,nT-1}) \tag{5.1}$$

consisting of the feature energies extracted on each scale of an anisotropic pyramid with nS spatial and nT temporal levels. The feature energy of a window (centred around the location p) computed on the s-th spatial and t-th temporal pyramid level is defined as

$$e_{s,t} = \sqrt{\frac{1}{W_s H_s} \sum_{i=-W_s/2}^{W_s/2} \sum_{j=-H_s/2}^{H_s/2} I_{s,t}^2(x_s - i, y_s - j)} \;, \tag{5.2}$$

where $I_{s,t}$ represents the s-th spatial and t-th temporal level of one of the invariant pyramids, H, S, and K, computed beforehand for every pixel. W_s and H_s stand for the (subsampled) spatial width and height of the neighbourhood on the s-th spatial scale (independent of the temporal scale). W_s and H_s are decreased by a factor of two per level, so that the effective window size is the same on all scales. The spatial coordinates of the location are also subsampled on the spatial scale s: $(x_s, y_s) = (x/2^s, y/2^s)$. In time, one frame of a lower pyramid level corresponds to several frames on the original level, so that we implicitly integrate over time, as well. Given a learning scenario, the optimal window size can be inferred from the eye movement data by systematically evaluating, in terms of performance in predicting fixations, a range of different neighbourhood sizes.

5.2.3 Learning

Given a collection of videos together with a set of salient and non-salient locations on these videos, the task of predicting interesting locations can be naturally viewed as a binary decision problem, to which efficient methods from machine learning can be applied.

Thus, the task of learning to distinguish salient locations consists in finding a confidence value quantifying the patch's level of interestingness. Formally, we look for a function $g : \mathbb{R}^{nS \times nT} \rightarrow \mathbb{R}$ that returns such a confidence value for a new movie location p, based on its energy vector $\mathbf{f_p}$. The training data comprises the feature energy vectors of previously seen locations and associated class labels (salient or not), $(\mathbf{f_{p_i}}, l_i) \in \mathbb{R}^{nS \times nT} \times \{-1, 1\}$.

The available fixational data is partitioned "movie-wise" into a training and a test set: gaze data of all viewers on one movie are retained for testing, while the fixations on the remaining movies are used for the training. For the classification we use a standard soft margin Support Vector Machine (SVM) with Gaussian kernels. A brief description of the theory of Support Vector Machines is included in the Appendix. Prior to training, we linearly scale each attribute (i.e. the feature energy on a particular spatiotemporal scale) to $[-1, 1]$. Optimal model parameters are found with cross-validation on the training sequence. To measure the quality of prediction, we perform an *ROC analysis* using the collected human gaze data as ground truth.

Receiver operating characteristic (ROC) analysis is a common evaluation metric from signal detection theory, which in recent years has been used increasingly in machine learning for model comparison. ROC curves illustrate possible tradeoffs between *true positive rate* (i.e. the fraction of correctly classified fixations among all fixations) and *false positive rate* (the fraction of non-fixated locations forecasted as fixations) for classifiers that have continuous output. A systematic variation of the threshold used to discriminate between salient and non-salient movie locations leads to a change in both the false positive and true positive rate, which can be plotted as a curve. The area under the ROC curve (AUC) is commonly used to summarize performance across all possible thresholds in a single value. The smaller the area under the ROC curve the more the predictor resembles a random classifier, which has an AUC of 0.5. An AUC of 1.0 means perfect discrimination.

To quantify the benefits of incorporating information from multiple scales, we compare the model with simpler variants of the above classifier that operate on *single scales* only. For this, we evaluate the performance of one-dimensional maximum-likelihood classifiers when the feature energies from individual pyramid levels are treated as inputs to the decision algorithm. Results for the “most predictive” scale are then compared to the performance of the (learned) multi-scale model.

5.3 Experimental evaluation

Here, we test the quality of the structure tensor-based predictors on a large set of eye movement data and compare their predictive power with that of four state-of-the-art models of bottom-up saliency.

5.3.1 Videos and eye movement data

Our experiments examined the performance of the proposed approach on the data set of eye movements on 18 high-resolution naturalistic videos (captured with a static camera) presented in Chapter 4. From the recorded gaze data, about 40,000 saccades were extracted using a dual-threshold velocity-based procedure [Böhme et al., 2006].

5.3.2 Data set labelling

The learning algorithm takes as input a set of positive, salient examples and a set of negative, non-salient ones. Whereas the set of fixations, more precisely saccade landing points, appears as a straightforward choice for the positive class, obtaining negative examples is non-trivial. An intuitive and commonly used approach is to arbitrarily pick locations from a uniform distribution either from the entire scene or (better) from areas that were not fixated, i.e. where spatiotemporal distance to the nearest fixation is large enough. However, as we have argued in Chapter 4, several recent studies have pointed out that such approaches do not account for a common problem inherent in most eye movement data sets: the tendency of viewers to fixate preferably in the centre of the display [Tatler et al., 2005, Tseng et al., 2009]. To remove possible artifacts due to the *centrally biased* distribution of gaze positions, it has been suggested that the non-salient locations of a video should be taken from real scanpaths on *different* movies. That way, an identical spatiotemporal distribution of the positive and negative examples over the set of all movies is obtained, but such artefact minimization also comes at a price. The above procedure of picking the negative examples may lead to overlap between the two classes and, hence, to an underestimation of the real model performance.

Existing approaches typically report results for only one of the aforementioned methods, so that it is not clear how sensitive the models are to labelling conditions, and whether or not the different conditions lead to significant deviation in performance. To investigate this and provide a fair comparison of the different models that might otherwise benefit from (labelling) biases, here, we consider both of the above labelling procedures: the “bias-free”, where we account for the central fixation bias and allow for overlap, and the “default” one, which minimizes the overlap. Loosely speaking,

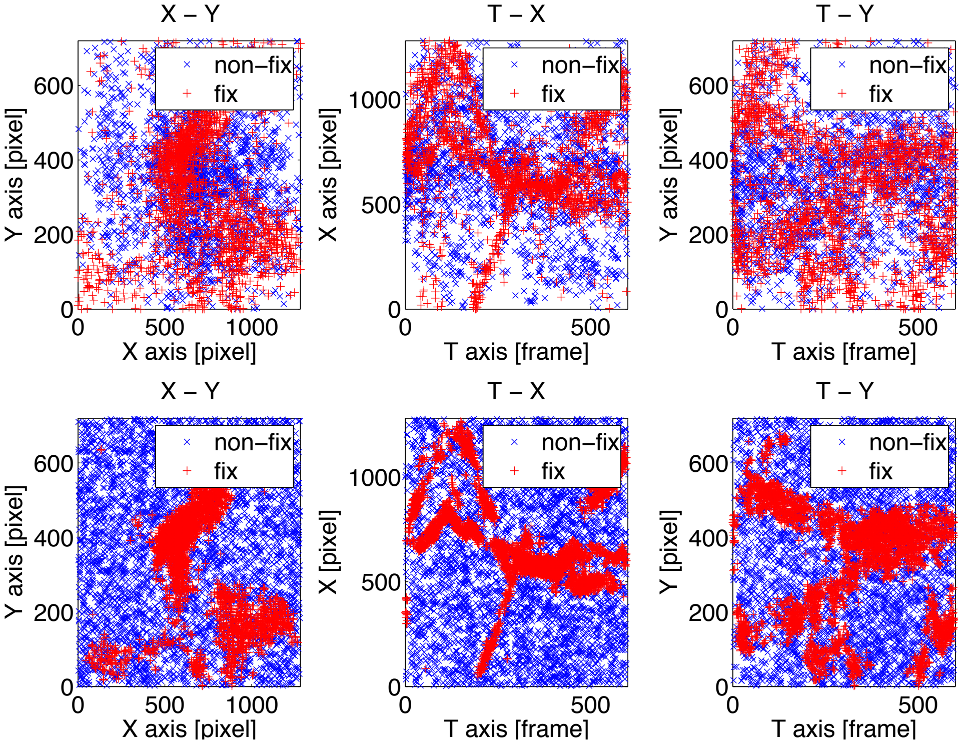


Figure 5.2: Salient (red plus) and non-salient (blue cross) locations on a movie. These locations are shown on the 2D projections (xy, tx, and ty) of the 3D spatiotemporal volume of the video. Upper row: “bias-free” labelling with saccade landing points in the salient class, and fixations on other movies in the non-salient class. Lower row: “default” labelling — salient and non-salient locations are chosen from the maxima and minima of the empirical saliency measure. Note the difference in overlap between the two classes under the two labelling schemes.

the “bias-free” scheme samples negative training data from different movies, whereas the “default” scheme samples from different spatial locations.

In the first case, the full set of saccade landing points is used to label the salient locations (about 40,000 over all movies and subjects). For the negative class, the non-salient locations of a movie are chosen using randomly selected scanpaths from different movies (see upper row of Figure 5.2). Because of latencies of the oculomotor system, the time of the gaze response to a specific salient event does not necessarily coincide with the time of the event. Hence, existing approaches usually introduce a temporal offset (between 150-250 ms) based on well-established results on reaction time to synthetic stimuli. However, as we have shown in Chapter 4 and in [Vig et al., 2011b], the typical reaction time is stimulus dependent and in natural scenes this average lag is near zero (i.e. no offset needs to be considered) due to the highly predictive nature of salient real-world events.

As argued before, such a "bias-free" labelling procedure introduces overlap in the salient and non-salient classes, i.e. the data set is contaminated with wrongly labelled samples (outliers) that deteriorate the model performance. In an attempt to avoid such overlap, in the "default" labelling scheme, we rank video regions according to the *"empirical" saliency measure* (introduced in Chapter 4), which is derived from the recorded eye movement data. As shown in Chapter 4, such maps are defined as the density of the gaze points averaged over all viewers and therefore constitute an upper limit of prediction, i.e. an inter-subject agreement. We compute such a probability map for each video, by superposing spatiotemporal Gaussians placed at each gaze location of all subjects. Samples of the salient and non-salient classes are picked from regions with the highest (for the positive class) and lowest (for the negative class) density of fixations. In our analysis, the Gaussian filter had a spatial support of 2.4 degrees of visual angle, a temporal one of 0.17 s, with the standard deviations 0.6 degrees (spatial) and 600 ms (temporal). An equal number (40,000) of salient (non-salient) locations was then chosen randomly from locations where the empirical saliency exceeds (is below) a given global threshold (see lower row in Figure 5.2). Threshold values were set at the upper ten percent (for salient) and lower one percent (for non-salient locations) of the maximum empirical saliency estimated over all movies. These values were chosen so as to obtain an equal number of data points in the two (salient and non-salient) classes.

5.3.3 Implementation

Here, we provide a more detailed discussion of how implementation considerations were integrated in our analysis.

To extract the proposed salient features (i.e. the geometrical invariants) on different spatiotemporal scales, we constructed an anisotropic Gaussian pyramid with $nS = 5$ spatial and $nT = 5$ temporal levels, as described in Section 3.3. This rather high number of pyramid levels (a free parameter) was chosen so as to ensure that frequency components that are potentially relevant for visual saliency are represented. For the structure tensor $\mathbf{J}$, partial derivatives in Equation 3.11 were calculated by first smoothing the input with spatiotemporal 5-tap binomial kernels $(1, 4, 6, 4, 1)/16$ and then applying $[-1, 0, 1]$ kernels to compute the differences of neighbouring pixel values. For the smoothing of the products of derivatives (with Ω), we chose the same spatiotemporal 5-tap Gaussian.

Besides being symmetric, the above filter kernels are non-causal, so that the temporal filtering requires video frames with future time stamps. As a consequence, depending on the number of temporal scales, a certain number of the initial and final output frames of the invariants are distorted. To avoid such temporal border effects, we only considered fixations from (and restricted the analysis to) valid frames. For a temporal pyramid with $nT = 5$ levels, this meant discarding quite a notable number of frames: the first and last 3.2 s (96 frames) were not considered for further analysis. Since the invariants H, S, and K comprise of products of one, two, and three eigenvalues, respectively, their dynamic range is not identical. For a fair comparison of the three, we therefore mapped them to the same dynamic range: they were raised to the power of six, three, and two, respectively.

To increase computational efficiency in the subsequent steps, the invariants were stored to disk using lossless compression. We normalized output invariant videos to pixel intensity values between $[0, 255]$ by taking the eighth root and linearly scaling the maximum over all levels to 255.

Once these features were extracted on multiple scales, we computed the feature energy in windows of varying size at each salient and non-salient location (about 25,000 per class over all movies, after discarding invalid invariant frames). We cropped the window at the boundaries if it was too large.

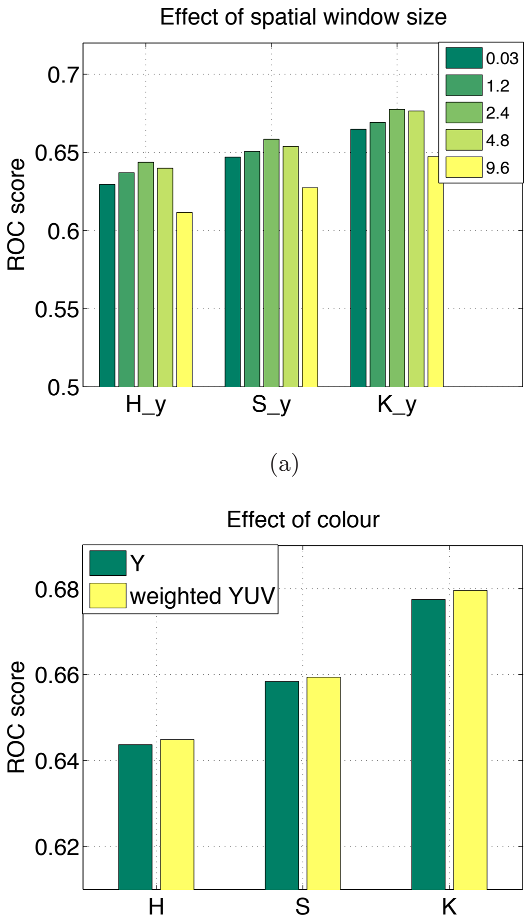
Finally, a classifier was trained with feature energy vectors on all but one video from the movie set and testing was performed on the withheld movie. The optimal parameters of the kernel Support Vector Machine (i.e. the width γ of the Gaussian and the penalty term C) were found by 8-fold cross-validation on the training sequence. Given a low number of videos (18 in total), and since eye movement predictability varies quite considerably between different video clips, the whole procedure (including the training and search for optimal parameters) was repeated 18 times so that each movie served as test data once.

To estimate the performance gain from incorporating information from multiple spatiotemporal scales, the predictability of the single scales was also tested. For this, an ROC analysis was performed (without further SVM prediction) on the energies from single pyramid levels. Here, multiscale results are compared with the outcome of the single "best" scale over all movies (in terms of ROC analysis), i.e. the frequency component that is most relevant for attentional selection. In case of multiscale analysis, the delivered decision values on the test movie are determined with respect to the training data, that is, the energy vectors from the remaining 17 videos. For single scales, however, a separate ROC analysis on each single movie would not take into account the overall distribution of feature energies in the two classes, and thus overestimate performance. Therefore, for single scales, instead of 18 ROC tests for the individual movies, we perform a single ROC analysis on the *entire* set of salient and non-salient locations from *all* 18 videos. This assures that during decision making the approximated true distribution of the fixated and non-fixated energies is used.

5.3.4 Quantitative analysis

In this section, we systematically investigate how different feature types contribute to model performance. We vary three main variables: the window size considered in extracting the feature energy, the colour channels (luminance alone or multispectral representations) on which the geometrical invariants are extracted, and, finally, the number of pyramid scales considered (single-scale vs. multiscale approach). The following analysis was performed for all three geometrical invariants. Since the qualitative results for the two types of data set labelling were identical, in this section, we only consider one: the "bias-free" labelling.

We started with the simplest scenario, considering salient features that are extracted on single spatiotemporal scales of the grayscale videos (i.e. no multiscale and multispectral analysis yet). Here, we report results for the pyramid level that gave best predictability, in terms of a single ROC analysis over the entire set of fixated and non-fixated locations from all 18 movies. To quantify the gain of the final spatial pooling (i.e. feature energy computation) on predictability, we varied the spatial window in size between a single pixel (i.e. no spatial pooling) to about 10 degrees of visual angle, with the exact window sizes used as follows: 0.03, 1.2, 2.4, 4.8, and 9.6 degrees. As seen in Fig. 5.3(a), the trend is consistent for all three



(a)

(b)

Figure 5.3: (a) Eye movement predictability as a function of window size for the "bias-free" labelling. Range tested: $\{0.03, 1.2, 2.4, 4.8, 9.6\}$ degrees of visual angle. For all three invariants, highest ROC scores were found at 2.4 deg. (b) Predictability using the geometrical invariants of the structure tensor on the luminance channel (Y) and of the multispectral structure tensor (YUV) given an optimal window size of 2.4 deg. Performance does not increase much with the addition of the UV colour channels. In both (a) and (b), invariants that extract features with higher intrinsic dimensions (K) are more predictive than lower intrinsic dimensions (S and H).

invariants: predictability increases with the window size, peaking at around 2.5 degrees, after which it slowly decreases. A window of 4.8 degrees still yields prediction rates close to the maximum. This is in agreement with psychophysical studies that claim the size of the influencing spatiotemporal context has roughly the size of the fovea. Since the relative gain in predictive power from no window to one of 2.4 degrees is 11% for invariant H, and 8% for S and K, a rather large pooling is justified. Therefore, for further analysis we fix the window size to the optimal 2.4 degrees.

The qualitatively most relevant result, however, is that the prediction performance increases with the intrinsic dimension: invariants that extract features with higher intrinsic dimension are more predictive. Thus, invariant K with an ROC score of 0.68 is best, followed by S (AUC of 0.66), whereas the worst performing is H with an AUC of 0.64.

Results for geometrical invariants computed on the luminance channel alone versus on multispectral representations (the weighted $Y'C_bC_r$ colour space) are shown in Fig. 5.3(b). Colour information has surprisingly little effect on saliency: it improves prediction performance, but only slightly.

Finally, we evaluate how much improvement can be achieved when including information from multiple scales. Thus, the single-dimensional ROC analysis is replaced by a kernel SVM that operates on 25-dimensional feature energy vectors computed on anisotropic invariant pyramids with $nS = 5$ spatial and $nT = 5$ temporal levels. As expected, results in Fig. 5.4 show some benefits of multiscale processing: prediction performance improved by 11% for invariant H, for S by 7%, while a slightly smaller increase of 4.5% is found for K.

5.3.5 Comparison to existing bottom-up models

We compared the proposed generic method with four state-of-the-art models of bottom-up saliency for dynamic scenes: the Bayesian "surprise" [Itti and Baldi, 2009], SUNDAy [Zhang et al., 2009], and the models of [Itti et al., 1998] and [Itti and Koch, 2001] (denoted by "Maxnorm" and "Fancy"), which are in fact implementations of the classical saliency map of Koch and Ullman [1985] (detailed in Chapter 2) but which employ different fusing schemes of the individual saliency maps into a master map. "Maxnorm" (normalized summation) yields smooth, more continuous saliency maps, while the iterative "Fancy" scheme yields increasingly sparser maps, with

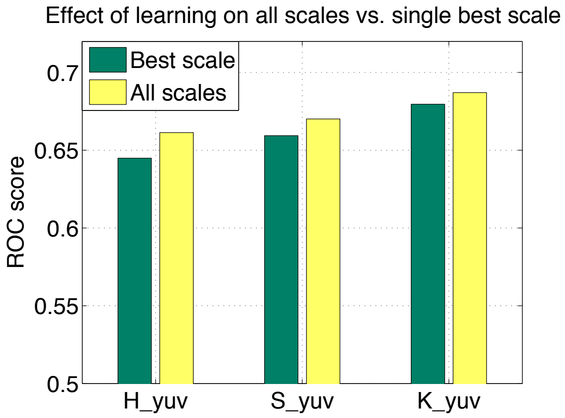


Figure 5.4: Predictive power of single-scale (i.e. "best" scale in terms of ROC analysis) and multiscale approaches (window size = 2.4 degrees, multispectral structure tensor, "bias-free" labelling). Using information from multiple scales improves performance, but only slightly.

only a few sharp peaks of activity. In Appendix B, we give an overview of the architecture of the Itti ("Maxnorm") and SUNDAy models. Default model parameters, detailed on the web pages of the toolboxes, were used to obtain saliency maps for the same video set. To discriminate between salient and non-salient movie locations, these maps were treated as maximum likelihood binary classifiers. By thresholding these maps, movie regions above the threshold were classified as salient. A systematic variation of the threshold – "movie-wise" – resulted in 18 ROC scores listed in Table 5.1. As before, the labelling scheme used to obtain the results in Table 5.1 was the "bias-free". For comparison, the geometrical invariants were extracted on multiscale and multispectral representations (with feature energies computed in the optimal window of 2.4 degrees). The prediction performance of the various models was compared with a paired Wilcoxon signed rank test. Statistical significance was obtained for $K > H$ ($p = 0.034$) and $K > S$ ($p = 0.013$), however, not for $S > H$ ($p = 0.395$). Also, results on the invariants proved to be significantly different from those of the four baseline models (except for $H >$ SUNDAy with $p = 0.07$). However, no statistical differences were found among the four state-of-the-art models.

Possible ROC values range from 0.5, which indicates chance performance, to 1.0, which means perfect discrimination. Note, however, that different class labelling strategies narrow the effective range of ROC scores. On the

Table 5.1: ROC scores of various bottom-up saliency models on the collection of 18 outdoor videos ("bias-free" labelling; numbers in bold indicate highest prediction rate). Regions with higher intrinsic dimension (encoded by invariant K) are significantly more predictive for saliency (paired Wilcoxon's test).

Movie	H	S	K	Maxn.	Fancy	Surp.	SUN
beach	0.67	0.68	**0.71**	0.64	0.61	0.61	0.65
breite_strasse	0.71	**0.76**	**0.76**	0.73	0.70	0.70	0.70
bridge_1	**0.63**	0.61	0.59	0.53	0.52	0.52	0.50
bridge_2	0.57	0.53	0.53	0.59	0.61	**0.64**	0.60
bumblebee	0.57	0.54	**0.63**	0.53	0.55	0.54	0.56
doves	0.80	0.82	**0.83**	0.67	0.70	0.71	0.72
ducks_boat	0.58	0.64	**0.70**	**0.70**	0.63	0.65	0.63
ducks_children	0.73	**0.78**	**0.78**	0.52	0.59	0.56	0.70
golf	0.75	0.76	**0.77**	0.70	0.60	0.67	**0.77**
holsten_gate	0.62	0.62	**0.66**	0.61	0.53	0.51	0.61
koenigstrasse	**0.64**	0.62	0.60	0.57	0.53	0.60	0.62
puppies	0.68	0.73	0.75	0.68	**0.76**	0.71	0.65
roundabout	0.68	0.69	**0.70**	0.63	0.63	0.62	0.63
sea	0.84	**0.86**	**0.86**	0.82	0.77	0.83	0.84
st_petri_gate	0.56	0.58	**0.60**	0.52	0.56	0.56	0.51
st_petri_market	0.62	0.60	**0.63**	0.57	0.56	0.52	0.58
st_petri_mcdon.	0.51	0.52	0.50	0.51	**0.59**	0.51	0.57
street	0.74	0.76	**0.77**	0.71	0.68	0.58	0.68
Average	0.66	0.67	**0.69**	0.62	0.62	0.61	0.64

one hand, the "bias-free" method that accounts for the central fixation bias may lead to erroneous labelling, which results in lower prediction rates. On the other hand, with no bias correction ("default" labelling), the model benefits from the differences in the spatiotemporal location distributions, which amounts to a substantial jump in performance. To estimate the effective performance range related to the two different labelling strategies, we additionally considered two simple control measures: (1) the spatial distance of the salient/non-salient location to the video-centre as a (possible) lower bound to this range, and (2) the "empirical saliency" measure – a fixation density map – as a "perfect" predictor of eye movements and, as such, as an upper bound. Note that when existing scanpaths from other movies serve as non-fixated points, the salient and non-salient location distributions are identical, hence, the distance to centre performs roughly at chance level. However, the empirical saliency is obviously an optimal predictor (with an AUC of 1.0) when the locations of the two classes are picked by thresholding this map.

The performance of the various methods for the two labelling strategies is summarized as averages over all 18 test sets/movies in Figure 5.5. With no bias correction ("default" labelling), the distance to the centre alone achieves a mean ROC score of 0.75, which is in agreement with previously reported results [Zhang et al., 2008, Judd et al., 2009]. At the same time, in the case of "bias-free" labelling, an empirical saliency measure built on the fixation positions discriminates these same locations from non-salient ones with a mean AUC of 0.79. The non-optimal performance is here due to noisy labelling and overlap in the two classes.

Despite its simplicity, our generic model based on the invariants of the structure tensor outperforms all four baseline models when accounting for the central fixation bias. Invariant K (average 0.69 AUC) comes closest to the upper bound marked by empirical saliency (0.79 AUC), but even the "weaker" invariants S and H still perform better than the baseline models; of those, SUNDAy achieves the highest average AUC (0.64).

Invariant K gives best prediction results (0.84 AUC) also for the second labelling procedure. Here, the two Itti models ("Maxnorm" and "Fancy", 0.81 and 0.80) perform better than SUNDAy and Surprise; the latter two surprisingly seem to be only as good as the "distance to centre" classifier.

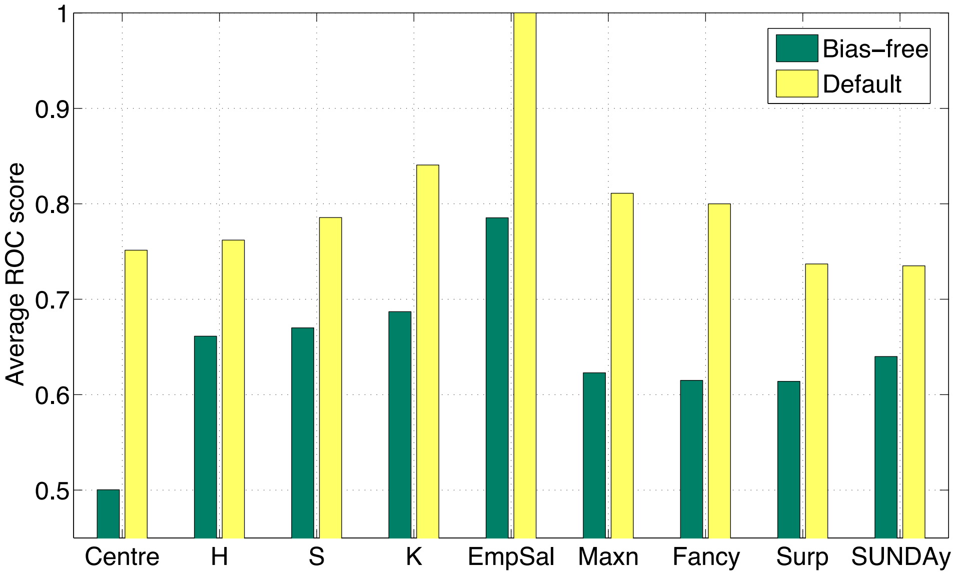


Figure 5.5: Average ROC scores of the various models for the prediction of eye movements on naturalistic videos. The two data labelling scenarios (green – "bias-free" and yellow – "default") differ on whether or not viewing biases are accounted for, and whether all fixations or only the most salient areas are modelled. To estimate the effective performance range, two control measures were introduced: (1) Centre – distance to the video-centre as a lower bound and (2) EmpSal - the empirical saliency as the upper bound. The invariants (H, S, and K) were computed with the optimal parameters: a multispectral anisotropic pyramid with five spatial and five temporal levels, and feature energy was averaged in a window of 2.4 degrees. Performance is compared to that of four baseline models: Itti's Maxnorm (Maxn) and Fancy algorithm, Itti and Baldi's Surprise model (Surp) and SUNDAy.

5.4 Discussion

In this chapter, we have derived a generic yet powerful model for bottom-up saliency from the simple assumption that the degree of local intensity variation is related to the informativeness of an image region. The concept of intrinsic dimensionality measures this degree and yields a basic description (or "alphabet") of how a multidimensional signal may change. We characterize typical video structures based on the geometrical invariants H, S, and K of the structure tensor, which correspond to the minimum intrinsic dimension of a movie region. Our model of bottom-up saliency combines such simple low-level visual features — the geometrical invariants extracted on multiple spatiotemporal scales — with machine learning to predict salient locations in natural dynamic scenes. We found that this simple approach proves successful in explaining human fixation data on a diverse collection of real-world videos. All three geometrical invariants were found to have good prediction capability. More importantly, however, our results provide strong evidence that the human visual system preferentially allocates its processing resources to more informative image regions; invariants that extract features with higher intrinsic dimension yield a sparser representation and they are more predictive for eye movements. Conversely, movie regions with lower intrinsic dimensions, i.e. redundant locations in case of $i0D$ and $i1D$, are less often fixated. Taken together, this provides indirect evidence for the efficient coding strategy of the brain [Olshausen and Field, 1996], and indeed $i2D$ operators emerge as non-linear filters when sparse overcomplete bases are learned [Labusch et al., 2009]. Our structure tensor-based approach is closely related to the space-time interest points of Laptev [2005]. In their approach, the spatiotemporal structure tensor is employed to detect local 3D corners in videos, which are highly useful in providing a compact representation of a movie. Such space-time interest points are popular in computer vision, e.g. for learning and recognizing human activities in videos.

Despite being based on simple, low-dimensional representations (1 to max. 25 scalars), the proposed model shows significant improvement over the four selected baseline models of bottom-up saliency. This finding becomes even more striking given the fact that such cognitive models rest on several assumptions, employ a high number of hand-tuned parameters, and involve complex computations. However, the straightforward hypothesis that during visual processing signals with lower intrinsic dimension are suppressed renders our model biologically plausible as well. Indeed, previous work has shown that this simple hypothesis can already explain the occurrence of lateral inhibition ($i0D$ signals are suppressed), end-stopping ($i1D$ signals are suppressed) [Zetzsche and Barth, 1990], and motion selectivity [Barth and Watson, 2000].

Existing approaches are typically tuned towards optimal performance for specific tasks: while the SUNDAy model yields smooth, continuous saliency maps that are more adequate for the prediction of real fixations, the Itti models (especially the normalization scheme Fancy) produce sparser maps with few peaks that rather account for the most salient scene locations only. To test how well our simple approach can generalize to both tasks, we defined two data-labelling scenarios: one that aims to model all human fixations, but picks non-salient locations so as to account for viewing biases, too; and a second, where salient and not salient locations are chosen from the most and least salient video regions without viewing bias correction. To our surprise, we find that while existing models typically excel in only one scenario, our approach, more specifically invariant K, is generic enough to provide optimal prediction for both problems.

We also have shown that although different labelling schemes allow the comparison of the relative performance of the different models, they also narrow down the effective performance range. Knowledge of the upper and lower bounds of the model performance is essential as it allows the assessment of the true performance gain and the estimation of the closeness to the optimal model behaviour achievable for a given problem formulation.

In order to understand the potential gains from more complex (but biologically motivated) features, that is from additional information (be it for instance multiscale or multispectral), we performed a comprehensive analysis by gradually extending our simplest saliency map, the geometrical invariants computed on a single scale of the intensity videos. With the integration of more features, the introduction of additional free parameters becomes inevitable, but their values are here fine-tuned in a supervised learning scenario.

Our first extension, the spatial pooling through feature energy computation, allowed us to consider movie sub-volumes (i.e. a salient context) of arbitrary size around the fixation. Thus, we could overcome the limitations of learning algorithms operating in high-dimensional (pixel) spaces. This is, however, only one simple way of decreasing dimensionality, and we are aware that by such a notable reduction also an information loss is introduced. Still, this step enabled the computation of visual features on multiple spatiotemporal scales, thereby modestly increasing the dimensionality again.

A key issue in the design of bottom-up saliency maps is how to combine separate feature maps coming from different modalities to create a unique master map. A main advantage of the concept of intrinsic dimensionality is that it leads to a unified representation of spatial and temporal saliency and, moreover, it can be readily extended to multispectral sequences. However, we found no strong difference between the invariants on luminance and those on a multispectral representation. This could be partly due to the fact that colour channels are highly correlated with each other, so that only redundant information is added with colour. Also, other colour spaces, such as the perceptually uniform CIELAB space, as well as the approximately equidistant HSV space, may better capture the true role of colour in attentional guidance.

Overall, we found that including more information and fine-tuning the model parameters through learning algorithms increased the predictability, but the gain was less than intuitively expected. Learning appears to partially compensate for the lower quality of an image or video representation, when quality is measured in terms of how compact a representation is. Note, however, that our eye movement prediction results are better than those of the reference models even without multiscale learning.

Obviously, as with any purely bottom-up model of visual saliency, the present approach cannot fully account for the complex nature of human fixation patterns. Nevertheless, such models may predict top-down behaviour reasonably well when the high-level task is implicit or unknown Elazary and Itti [2008]. Indeed, our proposed model further improves upon previous approaches and successfully predicts human eye movements during free-viewing of dynamic real-world scenes. Note that incorporating other known properties of active vision, such as scanpath statistics, temporal correlations of scanpaths, and preference for the centre, could lead to even better performance.

5.5 Extensions

In the remainder of this chapter, we report two extensions of the basic saliency framework presented above. The first extension consists of the substitution of the invariants of the classical (three-dimensional) structure tensor with the invariants of the nD structure tensor ($1 \leq n \leq 3$) with the aim of investigating what kind of local spatiotemporal variation is particularly predictive for saliency in natural scenes. A systematic analysis of the different types of variations shall reveal which dimensions may be sacrificed for a faster computation of saliency (or regions of interest) in systems with limited computing resources, e.g. mobile robot applications.

The second extension makes again use of the geometric interpretation of multidimensional signals. Our basic framework for saliency prediction is now extended to the case of multiple overlaid movies. Although such stimuli do not constitute the natural input the human visual system is exposed to, still, locally, multiple motions (in forms of occlusions) are common in natural scenes. We shall show that the invariants of the generalized structure tensor (see Section 3.2.4), which better characterize (and discriminate between) various motion types are more adequate for the prediction of eye movements on such complex stimuli than the invariants of the classical structure tensor.

5.5.1 Contribution of spatiotemporal intensity variation to bottom-up saliency

In this section, we use the above saliency prediction framework to quantify the contribution of local spatiotemporal variation of image intensity to visual saliency. To measure different kinds of variation, we compute, for the set of natural outdoor videos used above, invariants of the n-dimensional structure tensor ($1 \leq n \leq 3$). Considering a video to be represented in spatial axes (x, y) and temporal axis t, the nD structure tensor is evaluated for different combinations of axes (2D and 3D) and also for the (degenerate) case of only one axis. To obtain a simple measure of bottom-up saliency now the symmetric invariants of the nD structure tensors are used, which we compute on several spatiotemporal scales. The resulting features are evaluated in terms of how well they can predict eye movements on our complex videos. We shall show that a 3D structure tensor is optimal: the most predictive regions of a movie are indeed those where intensity changes along all spatial and temporal directions. Among two-dimensional variations, the axis pair yt, which is sensitive to horizontal translation, outperforms xy and xt by a large margin, and is even superior in prediction to two baseline models of bottom-up saliency.

As discussed in Chapter 3, for three-dimensional signals, i.e. the spatiotemporal volume of the video, usually a three-dimensional structure tensor is defined. However, on subspaces of the video volume (e.g. combinations of two axes, or even considering the degenerate case of a single axis only) 1D or 2D structure tensors can be constructed. For instance, considering only the vertical spatial dimension y and the temporal dimension t, the two-dimensional structure tensor $\mathbf{J_2}$ is defined as

$$\mathbf{J}_2 = \omega(y,t) * \begin{pmatrix} f_y^2 & f_y f_t \\ f_y f_t & f_t^2 \end{pmatrix} , \tag{5.3}$$

where $\omega(y,t)$ is a 2D-Gaussian smoothing function and f_y and f_t stand for the first order partial derivatives $\delta f/\delta y$ and $\delta f/\delta t$.

The intrinsic dimension is then obtained from the symmetric invariants of $\mathbf{J_2}$:

$$\begin{array}{lll} H & = 1/2 \text{ trace}(\mathbf{J_2}) & = \lambda_1 + \lambda_2 \\ K & = |\mathbf{J_2}| & = \lambda_1 \lambda_2 \end{array} , \tag{5.4}$$

where λ_i denote the eigenvalues of $\mathbf{J_2}$. Regions where $H > 0$ are at least intrinsically one-dimensional ($iD \geq 1$), e.g. non-vertical stationary edges, vertically translating edges, and uniform regions that change in time, whereas $K > 0$ indicates an $i2D$ feature such as yt corners (changing motion) and structures that appear or disappear in yt, which correspond to non-vertical translation. The generalization of the formulas for the n-dimensional case ($1 \leq n \leq 3$) is summarized in Table 5.2.

In the following, we shall quantitatively compare the power in predicting eye movements on complex natural videos of the above simple tensor-based representations that characterize different types of spatiotemporal changes. For our evaluation, we used the same data set of 18 high-resolution movie clips of natural scenes that proved useful in testing the predictability of our saliency prediction scheme. As a preprocessing step, all movies were cropped to the same size along the spatial axes (preserving the central 600 by 600 pixels), to make the resulting space-time cubes rotation-invariant

Table 5.2: n-dimensional structure tensors and their invariants, which correspond to the minimum intrinsic dimension (iD) of a region. Invariants that encode features of higher iD were previously shown to be better predictors of eye movements; therefore, they are used for further analysis (these are marked with a box).

n	nD Structure Tensor	Invariants (eigendecomp. of $\mathbf{J_n}$)	
1	$\mathbf{J_1} = \omega(u) * f_u^2$ $u \in \{x, y, t\}$	$\boxed{H = \lambda_1}$	$iD = 1$
2	$\mathbf{J_2} = \omega(u, v) * \begin{pmatrix} f_u^2 & f_u f_v \\ f_u f_v & f_v^2 \end{pmatrix}$ $u, v \in \{x, y, t\}, u \neq v$	$H = \lambda_1 + \lambda_2$ $\boxed{K = \lambda_1 \lambda_2}$	$iD \geq 1$ $iD = 2$
3	$\mathbf{J_3} = \omega(x, y, t) * \begin{pmatrix} f_x^2 & f_x f_y & f_x f_t \\ f_x f_y & f_y^2 & f_y f_t \\ f_x f_t & f_y f_t & f_t^2 \end{pmatrix}$	$H = \lambda_1 + \lambda_2 + \lambda_3$ $S = \lambda_1 \lambda_2 + \lambda_1 \lambda_3 + \lambda_2 \lambda_3$ $\boxed{K = \lambda_1 \lambda_2 \lambda_3}$	$iD \geq 1$ $iD \geq 2$ $iD = 3$

with regard to size (because movies had 600 frames). The total number of saccades that remained after the cropping was 24,370.

In the main part of this chapter, we showed that invariants that encode features of higher intrinsic dimensionality are better predictors of eye movements; therefore, here only these were considered (see Table 5.2). For each video, we computed the invariants of the tensors $\mathbf{J_1}$, $\mathbf{J_2}$, and $\mathbf{J_3}$ along *all* possible dimensions/combinations of dimensions. See Figure 5.6 for still shots from a movie and the corresponding invariants. The above invariants were computed on each scale of an anisotropic spatiotemporal multiresolution pyramid with $nS = 2$ spatial and $nT = 2$ temporal scales, in which each spatial pyramid was decomposed further into its temporal bands.

Following the prediction scheme detailed in the first part of this chapter, we labelled areas in the videos as salient and non-salient (according to the "bias-free" labelling scheme). To account for imprecisions in both the oculomotor and the eye-tracking system, we considered a spatial window (of 32 pixels, i.e. about 1.2 deg, on the highest pyramid level), and computed the window's *energy*, as defined in Section 5.2.2.

The predictive power of the different representations was assessed by evaluating (through ROC analysis) the performance of one-dimensional maximum-likelihood classifiers when the feature energies from the single pyramid levels are used as inputs to the decision algorithm. In Table 5.3, we report

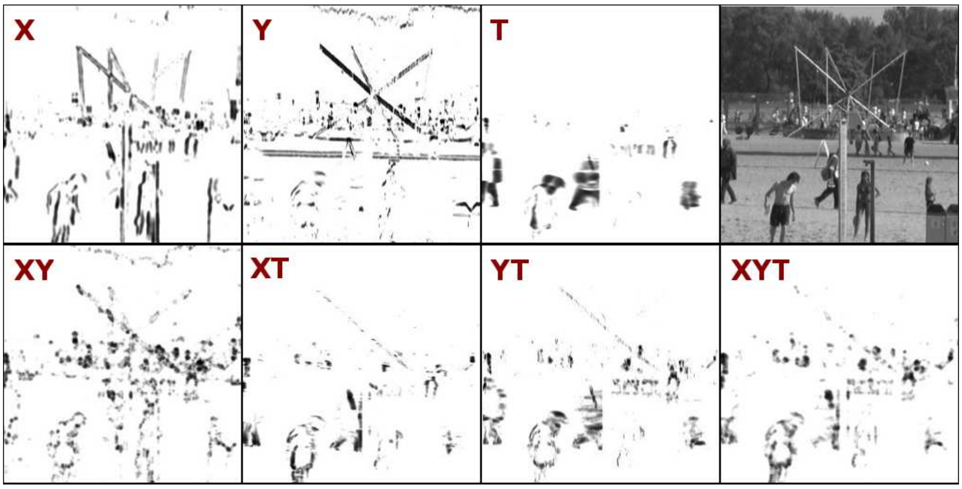


Figure 5.6: Top row (from left): H of $\mathbf{J_1}$ computed along the individual axes x, y, and t; original frame also shown. Bottom row (from left): K of $\mathbf{J_2}$ computed along the axes xy, xt, and yt; below the original image: K of $\mathbf{J_3}$ along all three axes.

average ROC scores (over the 18 movies) obtained for the "most predictive" scale (i.e. the pyramid level with the highest average ROC score).

For comparison, the saliency maps computed by two of the four state-of-the-art algorithms considered above [Itti et al., 1998, Zhang et al., 2009] were treated as maximum-likelihood classifiers for discriminating between fixated and not fixated video regions. By thresholding these maps, movie regions above the threshold are classified as salient. A systematic variation of the threshold parameter gives us a single ROC curve per movie and model. The averaged ROC scores over all videos are reported in Table 5.3.

We find that with an average ROC score of 0.673 the three-dimensional structure tensor $\mathbf{J_3}$ is optimal, suggesting that the most predictive regions of a movie are indeed those where intensity varies along *all* spatial and temporal dimensions. Surprisingly, the second best predictor operates on the axis pair yt; this predictor is sensitive to horizontal translations, which are most common in typical natural scenes. $\mathbf{J_2}$ evaluated on the axes yt outperforms xy and xt by a large margin (with an ROC score of 0.656 compared to 0.639 and 0.637, respectively), and is even superior to the two baseline models with ROC scores 0.644 (Itti & Koch) and 0.635 (SUNDAy), which incorporate a number of different features such as colour, contrast,

Table 5.3: Average ROC scores (AUC) of the different models and representations.

Model		AUC	Model		AUC	Model	AUC
$\mathbf{J_1}$	x	0.621	$\mathbf{J_2}$	xy	0.639	$\mathbf{J_3}(xyt) - K$	0.673
	y	0.617		xt	0.637	Itti & Koch	0.644
	t	0.623		yt	0.656	SUNDAy	0.635

and orientation. Although one-dimensional variations perform worst (with $\mathbf{J_1}$ along the vertical axis giving the lowest score – 0.617), their average prediction rate is still significantly higher than chance (ROC score of 0.5).

Our results can be used to choose efficient active vision strategies. Under the assumption that the human visual system is near-perfectly optimized for natural environments, the spatiotemporal structure tensor $\mathbf{J_3}$ thus picks the most informative regions. However, with our data, it is now also possible to choose which dimension should be sacrificed for faster computation in resource-limited systems, e.g. in an embedded real-time module of a robot with active vision sensors: for natural environments, the axis pair yt is more informative than xy or xt.

5.5.2 Prediction of eye movements on overlaid movies

In the following, we shall extend our framework to predict eye movements on transparently overlaid videos. In Chapter 3, we introduced a mathematical formalism, based on the invariants of the generalized structure tensor, to characterize multidimensional signal variation and certain motion patterns. Here, we use such generic tensor-based representations combined with the prediction framework presented in this chapter to investigate how eye movements are influenced by multiple overlaid motions and how well gaze can be predicted on such stimuli. Since the generalized structure tensor $\mathbf{J_G}$ — as opposed to the classical $\mathbf{J}$ — is able to distinguish between complex motion patterns [Barth et al., 2010], our hypothesis is that the invariants of $\mathbf{J_G}$ might better predict viewing behaviour on such complex stimuli than the invariants of $\mathbf{J}$.

In psychophysics, multiple transparent motions have often been used to probe the human visual system (see e.g. Braddick and Qian [2001] for

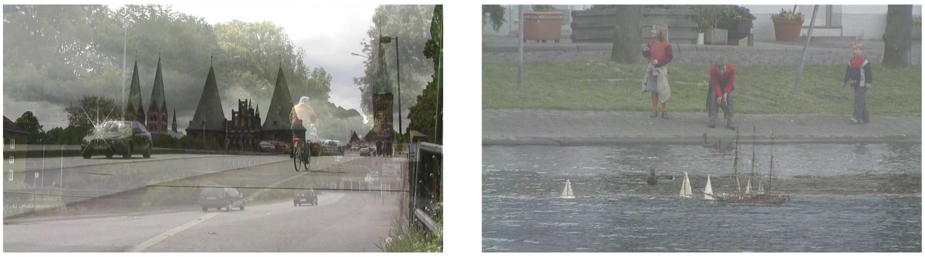


Figure 5.7: Stillshots from two blended movies.

a review). Obviously, such visual stimuli do not faithfully represent the kind of sensory input that biological visual systems are faced with every day. However, multiple motions are locally quite common in natural scenes because of reflections, occlusions, and transparencies. Here, we use such stimuli to test our hypothesis (already formulated in the first part of this chapter) that by suppressing all the redundancies that arise if the visual signal does not change in a particular direction(s) efficient representations can be obtained. The brain might employ such efficient representations for visual information coding (see Chapter 3).

For our evaluation, we used 19 transparently overlaid movie clips that were obtained by blending two videos randomly selected from our set of 18 outdoor sequences. Superimposing videos with very different spatiotemporal spectral energy distribution can lead to the perceptual dominance of one of the videos in the overlaid result. To avoid this, spatiotemporal frequency bands were equalized prior to blending. To perform pyramid-based blending, the videos were first decomposed into an anisotropic spatiotemporal Laplacian pyramid with five spatial and five temporal levels. Then, on each pyramid level, the blending weights were derived as the reciprocal of their standard deviation. Example stimuli are shown in Figure 5.7. The 19 resulting overlaid clips were shown on an Iiyama Master Pro 514 display screen (covering about 43 by 23 deg) to ten human subjects. From the raw gaze data, collected with an SMI Hi-Speed eye tracker running at 1250 Hz, about 10,000 saccades were extracted with the velocity-based procedure of Böhme et al. [2006].

Feature extraction was performed in a similar manner as for the single movies in the first part of this chapter. The invariants of $\mathbf{J}$ and $\mathbf{J_G}$ were computed on an anisotropic Gaussian pyramid with five spatial and five temporal levels. For a set of salient and non-salient locations (again, “bias-free” labelling), 25-dimensional feature vectors were extracted with which a Support Vector Machine was trained. The way the set of attended and non-attended locations was divided into a training and a test set differed slightly from the division on the single movies. The training set contained the fixations of two-thirds of all subjects (on all 19 movies), whereas the test set consisted of the fixations of the remaining one-third of the subjects (also on all movies). Hence, gaze data from all movies were used both for training and testing, but for the sake of generality, eye movements of any particular viewer were only present in one of the two data sets. Thus, our model is here put to test in terms of how well it can predict gaze behaviour of new viewers on videos that have already been “seen” (i.e. learned on) by the classifier. Apart from being able to predict eye movements on new, “unseen” videos, for various computer vision problems it is often important to model the expected gaze behaviour on stimuli that have already been seen by several viewers (e.g. websites). Here, we also test this ability of the presented saliency prediction framework. Due to the differences in the training/test set division, results cannot be compared directly between overlaid and single movies.

Quantitative differences in the distribution of prediction rates (ROC scores over 20 realizations into a training and a test set) are plotted for the invariants of $\mathbf{J}$ and $\mathbf{J_G}$ in Figure 5.8. As expected, prediction scores are now higher than for single movies, but qualitatively our results confirm the previous findings: eye movement predictability increases with the intrinsic dimension (i.e. the rank of the structure tensor) both for $\mathbf{J}$ (left part of figure) and $\mathbf{J_G}$ (right part). More importantly, the higher-order representation of $\mathbf{J_G}$ that allows a more precise characterization of motion types significantly outperforms $\mathbf{J}$ (paired Wilcoxon's signed rank test, $p < 1.1 \cdot 10^{-4}$). Thus, the generalized structure tensor is indeed able to better capture and characterize the complex nature of multiple signals (here motions) than the classical structure tensor. Results confirm our hypothesis that redundancies are suppressed even in the more complex case of transparent overlaid movies (as eye movements tend to avoid redundant regions).

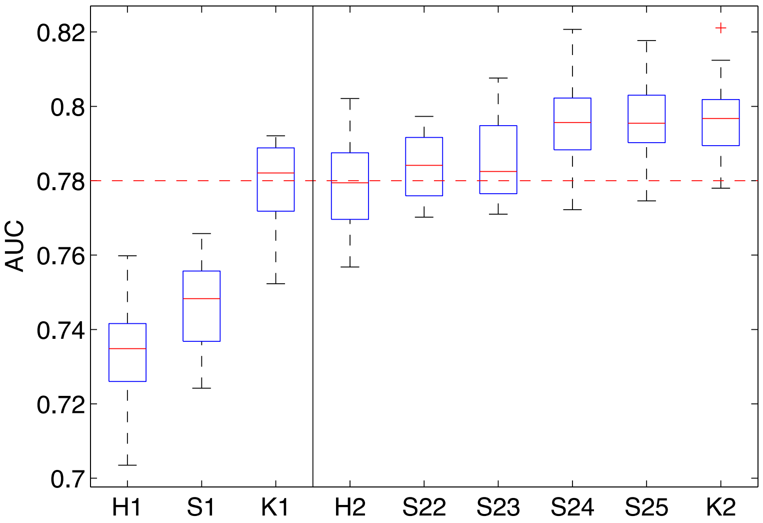


Figure 5.8: Box plot comparing AUC values obtained for the prediction of eye movements on overlaid movies with different invariants of $\mathbf{J}$ ($H1$, $S1$, and $K1$ on the left) and $\mathbf{J_G}$ ($H2$, $S22$, $S23$, $S24$, $S25$, and $K2$ on the right). Comparison of the prediction performance was done by Wilcoxon's signed rank test. Predictability is found to increase significantly with the rank. Overall, the invariants of $\mathbf{J_G}$ and the highest-order invariant ($K1$) of $\mathbf{J}$ give comparably high performance (median 78%, indicated by the red dashed line).

5.6 Chapter conclusion

In summary, in this chapter we have demonstrated how standard supervised learning techniques can fine-tune the free parameters of a simple image processing-based model of bottom-up saliency to account for eye movements in natural dynamic scenes. Grounded in the intuitive assumption that the visual signal must change in order to attract attention, in the first part of this chapter we proposed a generic model and tested its predictive power on a large set of eye movements in two distinct data labelling scenarios. Despite its conceptual simplicity, our model outperforms state-of-the-art baseline models. In the second part, we have presented two extensions of this model: i) to examine the contribution of intensity variation along different combinations of spatiotemporal dimensions to bottom-up saliency and ii) to predict eye movements on transparent overlaid movies based on the symmetric invariants of the generalized structure tensor. The results of the first extension are particularly relevant for machine vision systems with limited computing resources as they shed light on which dimensions may be sacrificed for a faster computation of interest points. In the second extension, our saliency prediction framework has proven successful in predicting eye movements also on more complex natural stimuli such as transparently overlaid movies.

6

Learned saliency transformations for gaze guidance

Through unconscious steering of visual attention to goal-relevant scene regions, gaze-guiding systems — as presented in Chapter 1 — promise to aid and complement human vision in many areas of human-computer communication and interaction. In previous chapters, we have already addressed two critical components of this process. In Chapter 5, we have put forth a simple yet powerful saliency model for the prediction of a limited set of salient candidate locations, from which the next desired saccade target is selected. For an unconscious guiding process, the optimal timing of gaze-capturing events is decisive. Therefore, in Chapter 4 we have characterized various video types with respect to the typical saccadic response lags to salient events. A major question, however, is yet to be answered: what image transformations are suitable and effective for altering the saliency level of a specific image or video region? In the present chapter, a generic saliency modification scheme is proposed that is built upon the saliency learning framework detailed in the previous chapter. Once the structural differences between attended and non-salient video regions have been distilled, transformation rules can be derived that manipulate some saliency-relevant properties of video regions. The proposed generic scheme is implemented in practice by considering spatiotemporal contrast manipulations (on an anisotropic Laplacian pyramid), and is evaluated both conceptually and empirically, in a psychophysical study.

Parts of the work described here have previously been published in [Vig et al., 2011a] and [Dorr, 2010].

6.1 Motivation

Apart from the prediction of scanpaths, only very few studies have addressed the intriguing question of how one can change an image or video locally to *influence* the emerging scanpath, i.e. how human gaze can be *guided* by image-based manipulations to the input. In situations where specific information must be found on a large visual display (e.g. while driving, analysing medical and geological images) [Rasche and Gegenfurtner, 2010], it is often crucial in which order the salient and relevant objects and events are attended to, i.e. how we look at a certain visual stimulus. Eye movement studies have shown that in several domains the gaze patterns of experts differ considerably from that of novices. For example, search strategies of expert and novice radiologists are substantially different [Nodine and Mello-Thoms, 2000], and experienced drivers' and pilots' gaze patterns exhibit shorter dwell times and are better defined [Kasarskis et al., 2001]; in other words, experts have learned to direct their eyes more efficiently. Moreover, in safety-critical situations, such as driving, assistance in where to look next, for example in order not to overlook a pedestrian, can prove more than beneficial.

Barth et al. [Barth, 2001, Barth et al., 2006] proposed *gaze-guidance systems* that steer the observer's gaze in a visual scene in order to enforce a predetermined, optimal scanpath and, through this, to aid the information uptake of the human viewer. The goal is to augment human vision with computer vision technology in a least-obtrusive way. Gaze guidance is realized by *gaze-contingent interactive displays* that use an eye tracker to monitor the viewer's gaze. In order to achieve an alteration of the gaze patterns, the saliency distribution of the visual scene is modified in real time by local changes to the visual input. Based on the original visual input and the eye position of the viewer, first, a limited set of salient, candidate locations is predicted that would attract the user's gaze. Then, using real-time video processing, the probability of being attended (i.e. its saliency) is increased for one selected candidate location, and simultaneously decreased for all other candidates. That such modifications are not perceived consciously is assured by the fact that they are embedded gaze-contingently in the periphery.

A few other attempts had been made to influence gaze patterns, either by filtering potentially salient targets [Su et al., 2004, Nyström and Holmqvist, 2010] or by adding synthetic gaze attractors such as high-frequency noise [Einhäuser et al., 2006] or flashing Gabors [McNamara et al., 2009]. However, these attempts were limited to static natural images and computer-generated content, where eye movements are more idiosyncratic and less driven by bottom-up saliency than on natural movies [Dorr et al., 2010a], and they were also not rendered gaze-contingently, so that subjects presumably quickly became aware of the changes and could consciously decide to ignore their effect.

In the above formulation, a critical issue is to identify *optimal image transformations* that can make a video region more (or less) eye-catching (i.e. salient) to the viewer. Here too, we use a data-driven approach to the problem by *learning*, from eye movements collected on real-world dynamic scenes, how to alter the saliency level of the video locally. As in the case of saliency prediction described in the previous chapter, we here consider a two-class classification scenario in which the video regions fixated by humans form the salient class and non-fixated locations represent the non-salient class. To the best of our knowledge, the general problem of "moving" a sample of a class into the other class, in an optimal way and under certain constraints, is novel in the machine learning and computer vision literature.

Before we describe our saliency modification framework in detail, we provide a short overview of past work on gaze-contingent displays, also briefly presenting the gaze-contingent display based on an anisotropic spatiotemporal Laplacian pyramid that we use for our experiments on gaze guidance.

6.2 Gaze-contingent displays — state of the art

Gaze-contingent displays (for extensive reviews see Duchowski et al. [2004], Parkhurst and Niebur [2002], Reingold et al. [2003]) manipulate an image or a video in real time based on the observer's gaze direction. The two main components of a gaze-contingent display are i) an eye tracker and a ii) displaying system that modifies its output as a function of the measured gaze position. The gaze-contingent paradigm has first been adapted in reading research (see e.g. Rayner [1998] for a review). Lately, it has been used extensively both in clinical applications (e.g. to understand and simulate visual field defects such as scotoma) as well as in psychophysical studies. Gazecontingency can be incorporated in various ways in experimental paradigms. One form of gaze-contingent displays during saccades modifies some properties of the visual input, such as contrast, motion content, colour, etc. With such displays one can investigate e.g. change blindness and transsaccadic integration. Other displays mask parts of the visual field permitting e.g. either central or peripheral vision only, and thus allow vision scientists to probe different perceptual mechanisms.

Gaze-contingent displays continue to receive great research interest also in more technical areas. The probably most well-known application of such displays is *foveation*. Foveated displays present high-resolution visual information at the point of gaze and gradually decreasing spatiotemporal resolution as the distance from the fixation increases. Thus, they simulate the non-uniform sampling of visual information implemented in the retina. If the width of the decay function matches the resolution distribution of the retina, an undisturbed (natural) visual experience can be evoked. Because of less high-frequency content, foveated images and videos can be compressed more efficiently and, hence, foveation is useful in reducing the bandwidth and storage requirements for video transmission and encoding [Geisler and Perry, 1998, Sheikh et al., 2003, Böhme et al., 2008].

Until recently, such foveated displays only allowed the lowpass filtering of the input (using a Gaussian pyramid) and rarely met real-time constraints. For the purpose of gaze guidance, however, a more sophisticated weighting scheme — such as the individual weighting of frequency bands — is highly desirable, not to mention the real-time requirements of gaze-guiding systems. For this purpose, Dorr et al. [Dorr, 2010] developed a gaze-contingent display that is based on an anisotropic spatiotemporal Laplacian pyramid, and thus allows the space-variant spatiotemporal filtering (by individual frequency-band weighting) of high-resolution videos. Extensions i) from an underlying isotropic Gaussian to an anisotropic Laplacian pyramid, as well as ii) from only the spatial to the spatiotemporal domain greatly increased the computational complexity, but algorithmic improvements and the implementation of the software framework on dedicated graphics hardware allowed for the realization of gaze guidance systems with very low latencies.

6.3 Saliency transformations

In the current gaze-guidance scenario, saliency transformations are limited to subtle changes in the video patch that go "unnoticed" — as they are embedded gaze-contingently in the periphery — yet still have a gaze-guiding effect.

In principle, such image modifications could be derived directly in the pixel (or intensity) space of image or video regions (also referred to as "patches"). However, as natural image patches are known to be samples of an unknown low-dimensional manifold in the space of all possible image patches (i.e. randomly drawing intensity values for the pixels of an image patch does not result in a natural image), transforming them in the original, high-dimensional pixel space will most probably result in unnatural, white noise images. In other words, there are only a limited number of modification types that can be applied to a given image while still keeping its natural look. Moreover, these modification rules would be specific to the image patch at hand, and would not apply to all patches.

Alternatively, one could map the high-dimensional pixel patch onto some lower-dimensional (parameter) space by performing local *feature extraction* on the patch. Such an approach clearly limits the range of possible image modifications to changes in the chosen feature space. This could mean, for example, an increase/decrease in either luminance contrast, colour contrast or intensity, or motion velocity. Nevertheless, it has the advantage that any meaningful feature modification still yields a natural looking image. However, a strong constraint is imposed on the chosen feature space by the need to be able to apply (or map back) the changes in the feature space to the pixel image. Additionally, as we intend to derive transformation rules from information on the characteristics of salient and non-salient image regions that was obtained with machine learning algorithms, the proposed feature space must be characterized by a good separability of the salient and non-salient classes.

Here, we propose to use the *local spectral energy* as a feature space that satisfies the above constraints. It is a low-dimensional representation of a movie patch computed on each level of a spatiotemporal Laplacian pyramid by averaging the squared pixel intensities within the patch. Learned transformations within this space can be implemented as local spatiotemporal contrast manipulations on a spatiotemporal Laplacian pyramid. We will show that such transformations lead to a modification of the saliency distribution, which in turn results in a change in the eye movement statistics.

In the following, we present the machine learning framework used for deriving transformations in the spectral energy space. Then, in Sec. 6.5, we evaluate the effect of the spatiotemporal contrast modifications on saliency distribution in a preliminary experiment, where such energy modifications are embedded offline in our naturalistic videos. The desired effect — an increase or decrease in absolute saliency — is observed in different saliency maps of the modified movies — maps computed by our saliency predictor (outlined in the previous chapter) and two other state-of-the-art models of eye movements. We end this chapter with an empirical evaluation of the effect of gaze-contingent saliency transformation on eye movements in a psychophysical experiment.

6.4 Transformations in the spectral energy space

To derive saliency alteration rules, we again explore a data-driven approach that takes advantage of learning the discriminative characteristics of salient video regions directly from human-labelled data (i.e. fixated video areas). Note that this approach does not make any strong assumptions per se on what constitutes saliency in natural movies. Our strategy is to first learn the structural differences between fixated and non-fixated movie regions by building a classifier that operates on the spectral energy representation of the patches, and then use information on the classification boundary to move elements of one class into the other.

6.4.1 Spectral energy as a simple saliency measure

The flow diagram of our joint saliency classification/modification scheme is depicted in Figure 6.1. The classification part of the model is built around the learning framework outlined in the previous chapter, the only difference being that instead of the geometrical invariants now a band-pass Laplacian serves as image feature. As in Chapter 5, we use eye movements collected on our videos to label movie areas as either attended or non-attended. The videos are first decomposed into their Laplacian pyramid representation (see

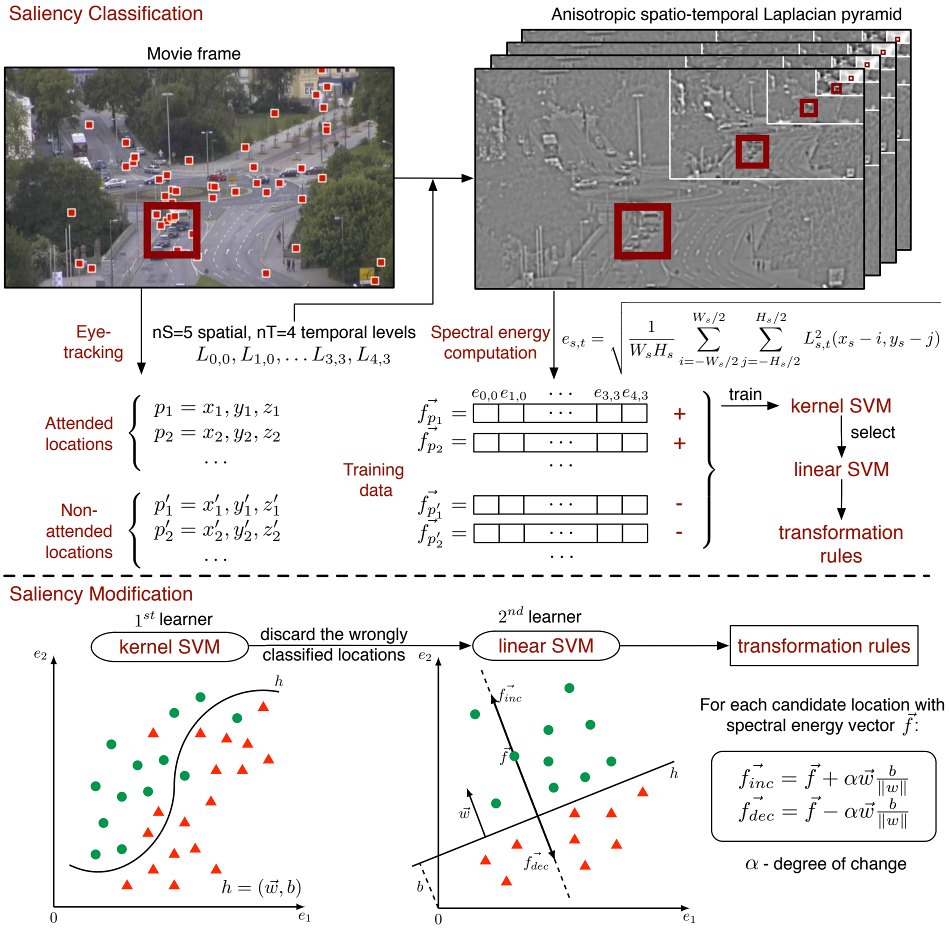

Saliency Classification
Movie frame
Anisotropic spatio-temporal Laplacian pyramid
Eye-tracking
nS=5 spatial, nT=4 temporal levels
$L_{0,0}, L_{1,0}, \dots L_{3,3}, L_{4,3}$
Spectral energy computation
$e_{s,t} = \sqrt{\frac{1}{W_s H_s} \sum_{i=-W_s/2}^{W_s/2} \sum_{j=-H_s/2}^{H_s/2} L_{s,t}^2(x_s - i, y_s - j)}$
Attended locations
$p_1 = x_1, y_1, z_1$
$p_2 = x_2, y_2, z_2$
Non-attended locations
$p'_1 = x'_1, y'_1, z'_1$
$p'_2 = x'_2, y'_2, z'_2$
Training data
$e_{0,0}\, e_{1,0} \cdots e_{3,3}\, e_{4,3}$
$\vec{f}_{p_1} =$ +
$\vec{f}_{p_2} =$ +
$\vec{f}_{p'_1} =$ -
$\vec{f}_{p'_2} =$ -
train
kernel SVM
select
linear SVM
transformation rules
Saliency Modification
1^{st} learner
kernel SVM
discard the wrongly classified locations
2^{nd} learner
linear SVM
transformation rules
e_2
h
$h = (\vec{w}, b)$
0
e_1
$\vec{f}_{inc}$
$\vec{f}$
$\vec{w}$
$\vec{f}_{dec}$
b
For each candidate location with spectral energy vector $\vec{f}$:
$\vec{f}_{inc} = \vec{f} + \alpha \vec{w} \frac{b}{\|w\|}$
$\vec{f}_{dec} = \vec{f} - \alpha \vec{w} \frac{b}{\|w\|}$
α - degree of change

Figure 6.1: Flow diagram summarizing the proposed approach. In the saliency classification phase (top), a classifier is trained with the spectral energy profiles of attended and non-attended video patches (fixations are denoted by small red squares in the movie frame). This feature is extracted as the mean-square-root of pixel intensities in a neighbourhood around the locations (large unfilled square) on each level of a spatiotemporal Laplacian pyramid. Bottom: Schematic view of transformation rules (for illustration purposes, only a two-dimensional feature space is shown: $\mathbf{f} = (e_1, e_2)$). An iterative SVM approach (kernel + linear SVM) is utilized to learn an optimal separation (a hyperplane $h = (\mathbf{w}, b)$) of salient (green dots) and non-salient (red triangles) video regions. To avoid saccades to a particular salient region whose energy profile is $\mathbf{f}$, the energy profile of the patch is moved perpendicular to the hyperplane h in the direction of the class of non-salient regions (along $\mathbf{f}_{\text{dec}}$). To increase the saliency of the patch, its energy profile is moved away from h (along $\mathbf{f}_{\text{inc}}$).

Chapter 3), i.e. a dissection of the original movie into a hierarchy (or pyramid) of videos such that each pyramid level corresponds to a different spatiotemporal frequency band. For each movie location $p = (x, y, z)$ in the two classes, the local spectral energy is extracted on each level of the spatiotemporal Laplacian pyramid. The spectral energy $e_{s,t}$ on the s-th spatial and t-th temporal pyramid level ($L_{s,t}$) is computed in a spatial neighbourhood centred around p as

$$e_{s,t} = \sqrt{\frac{1}{W_s H_s} \sum_{i=-W_s/2}^{W_s/2} \sum_{j=-H_s/2}^{H_s/2} L^2_{s,t}(x_s - i, y_s - j)}\ , \tag{6.1}$$

where W_s and H_s stand for the width and height of the neighbourhood on the s-th spatial scale (fewer pixels on lower-resolution spatial scales, but independent of the temporal scale). The spatial coordinates of the location p are also subsampled on the spatial scale s: $(x_s, y_s) = (x/2^s, y/2^s)$. The size of the neighbourhood considered is, here too, a free parameter whose value needs to be determined either from data fitting or chosen in accordance with the results of perceptual experiments.

With this low-dimensional representation — the spectral energy profile — of a video patch, a non-linear kernel Support Vector Machine (SVM) is trained that can discriminate between salient and non-salient movie regions. We here note only briefly that despite its simplicity this algorithm yields similar results to state-of-the-art saliency models. On our collection of natural videos, the leave-one-out ROC score for predicting eye movements — averaged over all movies and after removing the already discussed biases inherent in eye tracking data — is 0.62 for the above simple algorithm, 0.62 for the classical Itti and Koch model (with the Maxnorm normalization scheme), and 0.64 for SUNDAy (see Table 5.1). When the invariants of the structure tensor are used as underlying image features, better prediction results can be obtained (see previous chapter). However, these generic representations are not invertible (as non-linearities are involved in their computation — see Equation 3.12) and thus cannot be used for saliency modifications.

6.4.2 Spectral energy modification

Support vector machines search for an optimal "hyperplane", a decision boundary that separates the two classes with maximum margin. The hyperplane h is described by a vector $\mathbf{w}$ perpendicular to the plane and the bias b, which specifies its shift from the origin. The closer an instance to the plane, the more difficult it is to classify it into either group, because the more it resembles instances of the other class. The classification confidence of those points located far from the plane is high since, in our case, they are "truly" salient/non-salient video areas. Therefore, in order to change the saliency level of a movie region (in terms of its spectral energy) it suffices to move its energy profile relative to the plane, either towards the plane or away from it. Thus, a separating plane imposes a meaningful direction for transformations of spectral energy profiles in the feature space.

Still, an important question remains: how can we map back a modified feature vector (i.e. an energy profile) to an image patch? How to apply the learned transformations to the original video patch? Obviously, this mapping can only be approximate, but there are various ways of increasing or decreasing the spectral energy of a video patch. A straightforward approach, applied here, is to multiply every pixel in the patch with the ratio of the desired and actual energy, thus increasing or decreasing contrast in the specific pyramid scale.

One complication in our scenario relates to the fact that the classifier that best discriminates salient video regions from non-salient ones is kernel-based, i.e. it non-linearly maps its input data into a higher-dimensional space, where the problem becomes linearly separable. The non-linear mapping between the input space and the high-dimensional feature space is performed implicitly using the kernel trick, hence the ϕ non-linear embedding function is unknown. As a result, the reverse mapping (with an unknown ϕ^{-1}) from the feature space back to the input (energy) space of the modified data points is difficult. This is known as the *pre-image problem* in the kernel methods literature [Kwok and Tsang, 2004, Bakır et al., 2004]. It has been shown that exact pre-images typically do not exist but need to be approximated, in the process of which they can easily get distorted. To remediate the issue of a further non-linear mapping, we reformulate the task of learning a saliency classifier by considering only a subset of the attended and non-attended locations, thereby making the problem “easier”. Assuming that the video patches correctly classified by the kernel Support Vector Machine approximate well the manifolds of their respective classes, we train a second, *linear Support Vector Machine* with only these patches, in case of which the separating plane is defined in the input (i.e. energy) space — see Figure 6.1 for a visual illustration.

Recall that with our problem formulation (gaze guidance through saliency manipulations), only the alteration (in terms of the relative probability of being attended) of potentially gaze-capturing locations (i.e. candidate points) is intended. To modify the saliency of a candidate, i.e. salient, video patch, we move its energy profile perpendicular to the separating hyperplane of the linear SVM, either towards the non-salient class (i.e. towards the hyperplane, to make the patch less salient), or away from the hyperplane (to increase its saliency) — as shown schematically in Figure 6.1. Thus, for a candidate location with spectral energy vector $\mathbf{f}$, the transformation rules are defined as

$$\begin{aligned} \mathbf{f}_{\text{inc}} &= \mathbf{f} + \alpha_1 \mathbf{w} \tfrac{b}{||\mathbf{w}||} \\ \mathbf{f}_{\text{dec}} &= \mathbf{f} - \alpha_2 \mathbf{w} \tfrac{b}{||\mathbf{w}||} \end{aligned} , \tag{6.2}$$

where α_i denotes the degree of change.

One might argue whether the learning of such contrast modification rules (or weights) from eye movement data really is necessary. An analysis of the average spectral energy at attended and non-attended locations reveals that, on every scale, the attended movie regions have higher spectral energy than non-attended ones. Thus, it may suffice to increase/decrease energy by a constant factor — relative to the average spectral energy of the specific class — in each frequency band. However, we chose to learn these weights, since this way the local structure of the manifold of natural video patches is also considered, and the relative weighting of individual frequency bands becomes possible. Different spatiotemporal frequency bands may play different roles in guiding bottom-up attention, and individually weighting them can account for these differences.

To avoid artefacts, such as pixel saturation, due to strong contrast enhancements (occurring in the “saliency-increase” case), elaborate normalization schemes that map back the output videos to pixel intensity values in $[0, 255]$ are required. Because natural videos usually already use up the limited dynamic range of the display, we reduce the to-be-modified videos to $x\%$ overall contrast and adjust the energy weights (through the strength factors α_i in Equation 6.2) such that the intensity range at the modified location is stretched maximally without overflows. Also, in order to avoid strong and unnatural luminance changes in the candidate video patch, the DC component (i.e. the lowest pyramid level of the Laplacian) is left unaltered.

6.5 Conceptual evaluation

To evaluate the effect of the spatiotemporal contrast modifications on saliency and eye movements, in a first preliminary experiment, we embed such local energy transformations *offline* (i.e. not in a gaze-contingent manner) in our high-resolution videos of natural outdoor scenes. Our saliency prediction model outlined in the previous chapter and two baseline saliency models briefly presented in Appendix B (the Itti and Koch [Itti et al., 1998] and SUNDAy [Zhang et al., 2009] models), are used to compute saliency maps both for the unmodified and transformed movies. Using statistical tests, we verify whether the embedded spectral energy modifications really bring the desired change, i.e. an increase or decrease in absolute saliency.

6.5.1 Learning the contrast modification rules

For the experiment, we use our collection of 18 natural videos for which eye movements of 54 human subjects freely viewing these movies are available. The extracted saccades are used to find an optimal hyperplane for separating salient and less salient video regions.

The energy profiles of the attended and non-attended locations are computed on an anisotropic Laplacian pyramid decomposition of the videos (the pyramid having $nS = 5$ spatial and $nT = 4$ temporal levels), in a 5 by 5 degree spatial neighbourhood on all scales (which corresponds to 128×128 pixels on the highest spatial levels). In the periphery, the highest spatial frequency information is known to contribute little to attentional selection because it is discernible only near the fovea, and high spatiotemporal frequencies in general might contain a significant amount of noise from the recording system (e.g. camera sensor noise). Therefore, we leave the energies in these scales (8 out of the 20 pyramid levels) unaltered, i.e. we fix their weights to 1.0. Thus, the soft-margin kernel SVM [Chang and Lin, 2001] operates in a low-dimensional space: on the only 12-dimensional vectors containing energies from all but the highest spatial and temporal scales. The optimal SVM parameters, the width of the Gaussian γ and the penalty term C, are found with 5-fold cross-validation. Different from classical machine learning tasks, here, we do not wish to improve the performance of the above simple classifier on independent test data, but rather optimize it to better fit the given training data. Even though not relevant here, performance on test data is also good (see Sec. 6.4.1). The quality of prediction on the training data is measured through ROC analysis, which reports an ROC score of 0.82. After discarding the wrongly classified video patches, about 28,000 locations are left per class, with the energy profiles of which a linear SVM is trained. Its C parameter is again determined with 5-fold cross-validation. Now, with this linear SVM, on the selection of "truly" (i.e. easily discriminable) salient and non-salient video patches, an ROC score of 0.819 is achieved. The optimal separating hyperplane $h = (\mathbf{w}, b)$ found by this linear SVM shall be used to derive the rules in Equation 6.2.

6.5.2 Embedding the modifications in natural movies

For our evaluations, in the above 18 movies, about every second, 10 candidate locations are determined. In principle, we could have used the above simple saliency predictor based on the spectral energies (or the saliency model of the previous chapter) to generate these locations. However, for our testing purposes, the most precise determination of gaze-capturing areas is important, and human observers' eye movements are still best predicted by other observers' eye movements. Hence, we created a spatiotemporal fixation density map (the already discussed "empirical" saliency map) for each movie by placing a two-dimensional Gaussian with standard deviation 0.75 deg at each gaze sample of the 54 subjects. After normalizing the superposition of these Gaussians, the candidate locations are iteratively extracted from these maps by picking the location with the highest "empirical" saliency, and subsequently laterally inhibiting this location with an inverted Gaussian of standard deviation 2.35 deg. In this way, it is also assured that within a neighbourhood of about 5 × 5 deg no overlaps of candidates occur. With Equation 6.2, for each of these candidates a pair of new spectral energy vectors is computed based on the candidates' actual profiles, which were extracted with the parameters used for the SVM learning. The scalar α_i, which controls the degree of change, is first set to a fixed initial value independent of the candidate's energy vector. The rationale is that, at this point, we only define the directions of change in the feature space of spectral energies, and adjust the strength of the modification later, separately for each test condition, in which the effectiveness related to different modification strengths is examined. Thus, for contrast modifications, initial weights $\mathbf{w}_{\text{inc}}$ and $\mathbf{w}_{\text{dec}}$ are derived as the ratio between the desired and actual energies $\frac{\mathbf{f}_{\text{inc}}}{\mathbf{f}}$ and $\frac{\mathbf{f}_{\text{dec}}}{\mathbf{f}}$, respectively.

As mentioned above, if contrast is increased beyond what the dynamic range of the display allows, artefacts occur. Therefore, to leave room for contrast enhancements, we reduce the overall contrast of our movies by different amounts, and embed modifications in each of these contrast-decreased videos.

The final, video patch specific saliency-increase weights $\mathbf{w}_{\text{inc}}{}'$ are defined for each contrast level so as to stretch the dynamic range in the neighbourhood of the candidate between the extrema (i.e. 0 and 255, black and white – as we are operating on the brightness channel only). Thus, with different overall contrasts, it becomes possible to quantify the strength of the modification and evaluate its effect on saliency. We introduce a simplified notation for the *synthesis* of the Laplacian pyramid: $\sum_{s=0}^{nS-1}\sum_{t=0}^{nT-1} L_{s,t}$, which in fact involves the iterative upsampling and addition of the Laplacian levels. To avoid overflows, for each pixel $p = (x, y, z)$ in the modified spatiotemporal video patch the following must hold:

$$0 \leq \sum_{s=0}^{nS-1} \sum_{t=0}^{nT-1} w'_{s,t} L_{s,t}(x, y, z) \leq 255, \qquad \boxed{6.3}$$

where $w'_{s,t}$ is the patch-specific weighting coefficient for the spatiotemporal frequency band (s, t). These weights are obtained from the initially derived ones $(w_{s,t})$:

$$w'_{s,t} = (w_{s,t} - 1)\beta + 1, \qquad \boxed{6.4}$$

where β takes now the role of α_i from Equation 6.2 in controlling the strength of the manipulation. To stretch the intensity range to the extrema but not beyond, β is derived from Equation 6.3 for each spatiotemporal video patch individually as

$$
\begin{aligned}
\beta \quad & = \min_{(x,y,z)\in\text{patch}} \begin{cases} \dfrac{255 - \sum_s \sum_t L_{s,t}(x,y,x)}{d}, & d > 0 \\ -\dfrac{\sum_s \sum_t L_{s,t}(x,y,z)}{d}, & d < 0 \end{cases} \\
\text{where } d \quad & = \textstyle\sum_s \sum_t w_{s,t} L_{s,t}(x,y,z) - \sum_s \sum_t L_{s,t}(x,y,z)\,.
\end{aligned}
\tag{6.5}
$$

For each pixel in the video patch, exactly one of the following conditions holds: (1) the denominator d is larger than zero, i.e. the manipulation brings an increase in pixel intensity, and so (in the limit) β should stretch the new intensity to 255; (2) the denominator is negative, i.e. the modified pixel intensity is smaller than the original, and should, therefore, be decreased further to 0; (3) the denominator is zero, i.e. the pixel intensity remains the same, hence, β is not affected. By picking the smallest ratio over all pixels in the patch, we assure that the modified intensities remain in the allowed range.

The quantification of the strength of the *saliency-decrease* rules cannot be tied to the overall contrast level of the video. Instead, the strength is varied by scaling the initial $\mathbf{w}_{\text{dec}}$ weights so that n weights closest to zero are actually brought to zero ($0 \leq n < nS \cdot nT$). Setting the energies in certain scales to zero means removing those frequencies. For example, by carefully selecting the bands in which the energies are set to zero an object of a certain size moving with a certain speed can be filtered out.

For the experiment, three saliency-increase and one saliency-decrease strengths were tested; for the increase rules, the original videos were decreased to 70, 80, and 90 percent overall contrast. For simplicity, we only report results for one condition: the 80% overall contrast case. The same qualitative results were obtained in the other two conditions, with the obvious difference that saliency-increase modifications at 70% contrast were stronger than at 80% or 90%. For decrease rules, n was set to 4, i.e. frequencies in four spatiotemporal levels – with weights closest to zero – were removed. Every second, the saliency of 5 randomly chosen candidate points was increased further, and the remaining 5 candidates were decreased in their saliency. For the results reported below, spatiotemporal contrast manipulations were embedded in a 5 by 5 deg spatial and 700 ms temporal neighbourhood centred around the candidates. An example still shot from a

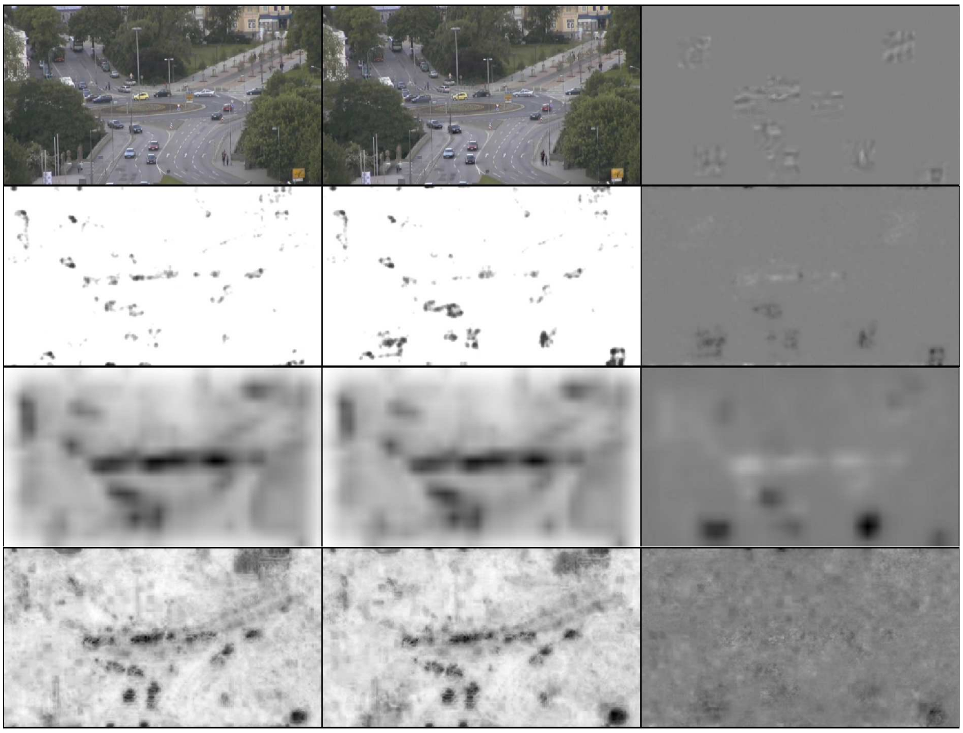


Figure 6.2: Saliency maps for one frame of an original (first column) and altered (second column) video. Ten non-overlapping candidate (i.e. salient) regions undergo saliency manipulations: the five candidates in the upper part of the scene are reduced in saliency, while the remaining five in the lower part are rendered more salient. Three baseline models are used to obtain the saliency maps: the geometrical invariant K (second row), the model of Itti and Koch (third row), and SUNDAy (last row). In the differences of the saliency maps before and after the modification (third column), the desired alteration in saliency can be clearly detected for the saliency-increase case (dark areas in the difference maps), while the decrease rules (bright areas in the differences) have a weaker effect on the saliency (in particular for SUNDAy).

movie and its altered version is shown in the first row of Figure 6.2. Lack of temporal change in the printed figure renders the modifications less visible than in the actual movie; however, in the difference map of the two, the 10 modified patches are clearly discernible. In this specific frame of the "roundabout" scene, the 5 locations in the upper part of the scene are decreased in saliency, while those in the lower part are increased.

6.5.3 Results

The effect of spectral energy modifications on overall saliency is evaluated by pairwise comparison of the saliency maps of unmodified and transformed videos — maps which were generated by three independent models of bottom-up attention: the classical Itti and Koch model (with the Maxnorm normalization scheme), SUNDAy, and our saliency predictor that relies on the estimation of the intrinsic dimension by means of computing the geometrical invariants of the structure tensor. As we have seen in Chapter 5, the geometrical invariant K, which encodes spatial and temporal changes and is computed as the product of the eigenvalues of the structure tensor, even outperforms baseline models in predicting eye movements. All of these models compute saliency on spatially downsampled versions of the original movie in order to reduce computational cost and to increase resilience against noise. The lowpass-filtered videos (6.6 cycles/degree) were created by filtering the video with a 5-tap spatial binomial filter and downsampling it (in space) by a factor of two. Note, though, that the highest spatial levels remained unchanged in our transformations anyway.

Saliency maps for the "roundabout" scene from Figure 6.2 are shown in subsequent rows of the same figure (in the order: invariant K, Itti and Koch, and SUNDAy – second to fourth rows). Alterations in the saliency distribution are visually more striking in the image differences between the saliency maps of unchanged and modified videos (third column). Here, a deviation from the gray value indicates an alteration in the saliency level: at darker areas a saliency-increase occurs, while brighter regions experience a decrease in saliency. Visually, saliency-increase seems to have a more pronounced effect than decrease, especially in the case of the maps computed by SUNDAy.

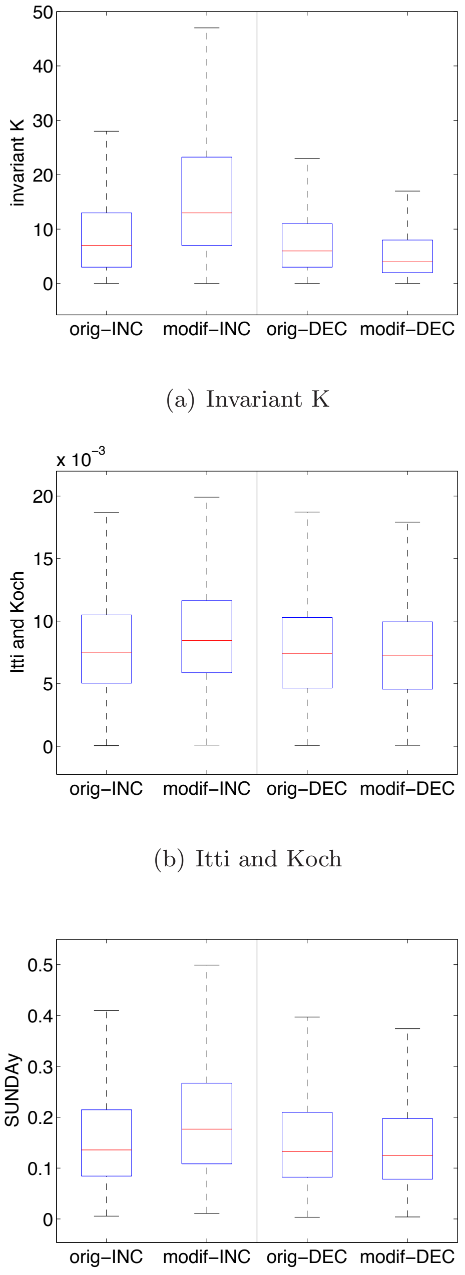
The saliency of candidate locations before and after the energy modification was compared with a paired Wilcoxon signed rank test, and proved to be significantly different for all three saliency models and both increase and decrease (see Figure 6.3). Results confirm our observation on the effectiveness of the modifications: the differences in saliency level are substantially greater where a saliency-increase manipulation was performed than at decrease locations. However, comparing the effectiveness of the two types of changes is not entirely fair, as the quantifications of the strength of manipulation for increase rules is independent of that of the decrease rules. Also, the modifications are the most effective (in changing the saliency distribution) for invariant K ($p = 1{,}9 \cdot 10^{-240}$ for increase rules, $p = 1{,}7 \cdot 10^{-165}$ for decrease rules). Nevertheless, the desired effect is reached also in the saliency maps of Itti and Koch (increase, $p = 4{,}9 \cdot 10^{-213}$; decrease, $p = 5{,}0 \cdot 10^{-67}$) and SUNDAy (increase, $p = 1{,}0 \cdot 10^{-145}$; decrease, $p = 8{,}0 \cdot 10^{-25}$). Unlike invariant K, which detects spatiotemporal intensity variations (space-time corners, non-constant translations), the two state-of-the-art models base their prediction of saliency on additional low-level features, such as colour and orientation. This explains why modifications to contrast only have a more modest (yet significant) impact on overall saliency in the case of the Itti and Koch and SUNDAy models.

It should be noted that, since the saliency transformation rules are learned from eye movement data and validated on existing saliency models, our results are also indicative of the biological relevance of these models.

Since the learned energy modifications indeed resulted in the desired alteration of the saliency level, in this first evaluation stage we have successfully demonstrated that gaze guidance is feasible in principle. The next step is to prove its usefulness also empirically, by examining what effect such low-level video modifications have on eye movement statistics.

6.6 Empirical validation

Although the above conceptual evaluation delivered encouraging results, a further empirical validation needs to be performed to prove the effectiveness of gaze-contingent energy modifications in guiding eye movements within an eye-tracking paradigm. To this end we performed a psychophysical experiment, in which six subjects viewed our gaze-contingently manipulated 18 videos of natural outdoor scenes. The experimental setup was identical



(a) Invariant K

(b) Itti and Koch

(c) SUNDAy

Figure 6.3: Box plots comparing the saliency distributions of candidate locations extracted from the saliency maps of original and modified videos. The distributions at saliency-increase (INC) and decrease (DEC) locations are treated separately. In all cases, the differences between the original and modified saliency distributions are statistically significant (paired Wilcoxon signed rank test) (middle line: median, box: upper and lower quartile, whiskers: data extent, outliers not shown).

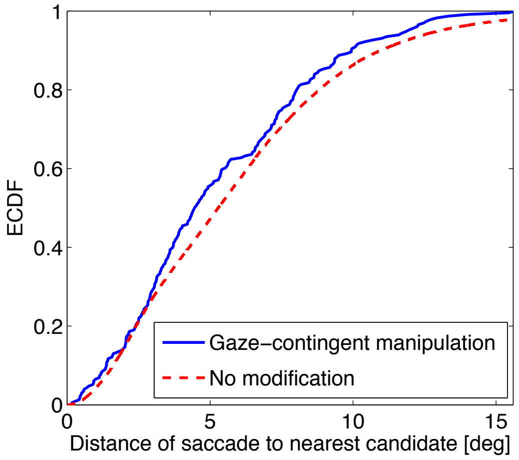


Figure 6.4: Empirical cumulative distribution function (ECDF) of the distribution of distances of saccades to the nearest manipulated location. Saccades tend to converge towards the modified regions in the gaze-contingently manipulated videos — therefore the shift of the solid curve to the left.

to the one considered in Chapter 4. The energy modification rules derived from the information on salient and non-salient video regions were now embedded online, in a gaze-contingent manner. Six different saliency increase and three decrease strengths were tested (in the same manner as above), and subjects were asked to press a button whenever they detected contrast manipulation during the viewing of the videos. To ensure correct coupling of responses and modifications, contrast manipulations were embedded only every three seconds. During each modification, up to 20 candidate locations were detected, one of which was increased in saliency whereas all others underwent a decrease in saliency.

As above, for simplicity, we only report results for a single increase–decrease strength combination. For gaze guidance to be successful, a significant difference is expected in the distribution of fixations recorded on the manipulated videos vs. those on the originals. Figure 6.4 summarizes results on the effect of gaze-contingent energy modifications on eye movement statistics. There, we plot the distribution of the distances measured between saccade endpoints and the nearest modified location. When compared with the control condition (i.e. saccades collected on the unmodified videos), the cumulative distribution of the distances is shifted to the left, indicating that saccades were made significantly more often to manipulated regions than when no manipulations were present at these locations. Interestingly, subjects became aware of the manipulations in only 4.5% of the trials, suggesting that an unconscious gaze-guiding process was indeed achieved in the rest of the trials [Dorr, 2010].

Thus, results confirm the preliminary conclusions of the previous section: gaze guidance proves feasible and useful even with simple spatiotemporal contrast manipulations implemented as local weighting of the spectral energies extracted on the levels of an anisotropic Laplacian pyramid. The evaluation of a considerably larger body of experimental data with a more systematic exploration of the parameter space (concerning e.g. the strength, duration, and spatiotemporal distribution of manipulations) is still ongoing, but interim results support the above findings.

6.7 Chapter conclusion

Directing visual attention to particularly goal-relevant areas in the visual field is a promising new strategy to integrate into future visual and communication systems. Our goal in this chapter was to explore techniques that allow to alter the saliency distribution of the scene, by embedding subtle low-level changes in the visual stimulus. With effective changes that do not introduce objectionable image artefacts, an unconscious gaze guiding process may be achieved.

In this chapter, we proposed a generic saliency modification scheme in which, first, the structural differences between attended and non-attended video locations are learnt. The information on the class boundary that separates the two classes was then used to derive the desired image transformations that lead to an alteration in saliency. Our scheme is generic because it does not assume any specific low- or high-level image feature space in which the manipulation rules are derived. However, two constraints have to be met by the selected feature(s). First, for effective saliency transformations, in this space, a high separability of the salient and non-salient video areas is highly desirable. Second, modifications in the chosen feature space need to be mapped to manipulation rules in the original input or pixel space of videos.

The spectral energy, computed on a spatiotemporal Laplacian pyramid, has proven to be a simple feature that fulfills the above constraints. Transformations performed in this low-dimensional space were implemented as local spatiotemporal contrast manipulation rules (on the spatiotemporal Laplacian). Normalization schemes to avoid visual artefacts and ways to quantify the modification strengths were also discussed. In a preliminary experiment, which aimed at evaluating the potential of such local video manipulations, we used three independent saliency models to compare the saliency maps of the unmodified and altered videos. The desired effect was reached in the saliency maps of modified movies, where a saliency-increase (or -decrease) rule applied to a video patch led to an increase (or decrease) in absolute saliency relative to the original movie patch. The second experiment confirmed the effectiveness of the modifications in a real-world eye-tracking experiment: even though gaze-contingent modifications usually remained invisible, they had a guiding effect on eye movements.

Note, however, that with such simple modifications the guiding effect is provable but still rather modest. We expect that the use of other, more powerful image features (such as motion and flicker) would greatly enhance the effectiveness of these video manipulations.

7
Conclusion

Gaze guidance, as envisaged in this thesis, holds considerable promise for improved human-machine communication, by complementing human perception with computer vision technology in a least-obtrusive way. The potential for such augmented vision aids is indeed huge. Gaze-guidance systems can provide "support" to an untrained eye in various scenarios and accelerate the mastering of task-specific skills. They may have a great impact on medical applications, too, e.g. by aiding patients with attentional deficits such as neglect. Also, this technology permits the guidance of gaze in safety-critical situations, for instance in traffic when a driver would otherwise overlook a pedestrian. Yet, before such systems become a reality, significant technological developments, on the one hand, and a deeper understanding of visual perception, on the other, must be achieved. The critical aspects of the technical side, such as the need for accurate and low-latency gaze-tracking and real-time image processing, may be more evident. However, a complete knowledge of the mechanisms involved in attentional orienting and the ability to simulate these biological processes are also required.

With the work presented in this thesis, we have contributed significantly to this latter challenge — the computational modelling and simulation of attentional processes — and, through this, opened the road towards the practical implementation of such attention-directing systems. Specifically, within the context of gaze guidance, our model of bottom-up attention, described in detail in Chapter 5, provides a solution to the problem of identifying a small set of salient locations in videos that the viewer may next attend to. Using the saliency modification framework put forth in Chapter 6, we can then apply appropriate transformation rules to the selection of salient locations in order to modify their interestingness, and thus, to influence where people look. Apart from the critical questions of *where* and *how* to apply the optimal video transformations, in a gaze-contingent scenario it is also essential to know *when* exactly to perform these manipulations to optimally bias saliency. To this end, Chapter 4 investigated the oculomotor response time to various types of video data.

Beyond their relevance for gaze guidance, the proposed models and findings have far-reaching implications in a much wider context. We have, for instance, demonstrated how simple principles rooted in the signal processing properties of human perception can serve as tools both to investigate perceptual phenomena and to predict likely saccade targets in videos. The simple geometrical framework reviewed in Chapter 3 proved very useful in quantifying the degree of anticipation in both truly natural and edited videos. The temporal component of gaze allocation is a rarely studied aspect of visual orienting in complex scenes, and we could show that the typical response times in naturalistic videos differ considerably from that in quasi-realistic scenes such as video games and TV clips. More importantly, we found that the average oculomotor lag in natural scenes is near zero, indicating an adaptation — in the course of evolution — of the human visual system to the often predictable dynamics of the real world.

The same geometrical scheme served as a basis for the novel, low-complexity computational model of attention put forth in Chapter 5. Hence, in this thesis, visual saliency has been quantified as a measure of the degree of local (video) signal variation, and special emphasis was placed on the generic nature of the approach: machine learning techniques and simple image representations of videos derived from efficient coding principles were combined to advance the state of the art in eye movement prediction. Our aspiration towards simplicity in design contrasts with existing, usually more complex approaches which, in order to grant biological plausibility, make several assumptions about perceptual processes, and whose results depend on the optimal choice of many free parameters. Despite its conceptual simplicity, our model outperforms baseline methods, and hence, holds considerable promise for practical applications in machine vision.

Finally, with the work on saliency transformations described in Chapter 6 we hope to have opened up a new research direction in attentional modelling. The proposed saliency modification scheme is built upon our saliency-learning framework, and is generic enough to operate on a diverse set of image features on which the manipulation rules are derived. We have demonstrated empirically the capability of the approach to alter the saliency distribution of scenes and guide the eyes in two validation experiments, thereby delivering a proof of concept of the feasibility of gaze-guidance systems.

The saliency prediction and modification framework put forth in this thesis suggests many promising directions for future research. Possible extensions of the saliency model include the incorporation of a top-down component that allows modulation by task and prior knowledge. Also, the consideration of a foveated input promises to reduce the computational load of the operations, and hence, to notably speed up the prediction. The recently developed principles of compressed sensing might lead to even more efficient image representations. With respect to computer vision, the possible practical problems where our saliency predictor could act as an efficient attentional front-end (or preprocessor), e.g. by filtering out the irrelevant information, are numerous. Work on using the attentional model as a component of an existing very powerful object recognition system has already been planned for the near future. The problem of deriving saliency transformations is an example of the much more generic problem of moving data points in high-dimensional manifolds under a set of constraints. We believe our work on saliency transformations is primarily a methodological contribution, and we acknowledge that the feature space — of the spectral energy — for which the use of the proposed modification scheme is demonstrated is rather simple. Therefore, a natural extension of this work would be the consideration of a wider set of more complex image features with which more sophisticated saliency transformations could be learned.

In summary, this thesis has contributed several essential building blocks towards the development of future information and communication systems that incorporate attention.

A

Support Vector Machines

In this appendix, we briefly recall the theoretical foundations of Support Vector Machines (SVM) for the discrimination of both linearly separable and non-separable data. For additional information on Support Vector Machines we refer to seminal papers [Vapnik, 1998, Osuna et al., 1997, Burges, 1998] and textbooks [Cristianini and Shawe-Taylor, 2000, Schölkopf and Smola, 2002, Bishop, 2006].

Support Vector Machines are supervised learning techniques that have been applied successfully to a variety of classification and regression problems and are considered state of the art. We here formalize the theory for the two-class classification scenario. The problem setting is as follows: given a set of training examples of the form $(\mathbf{x_i}, y_i)$, $i = 1, \ldots, N$, $\mathbf{x_i} \in \mathbb{R}^d$, with $y_i \in \{-1, +1\}$ class labels, we wish to predict whether a new (so-called *test*) example belongs to one of the two (positive or negative) classes. We first consider the simplest case, when the data is linearly separable, i.e. a function $f(\mathbf{x}) = \mathbf{w^T x} + b$, $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$, exists (in the 2D case a straight line) such that $\forall i$

$$\begin{aligned} \mathbf{w^T x_i} + b &\geq 0, \qquad \text{if } y_i = 1, \\ \mathbf{w^T x_i} + b &< 0, \qquad \text{if } y_i = -1. \end{aligned} \tag{A.1}$$

Given such a separating line (or a *hyperplane* in higher dimensional spaces), a new test sample $\mathbf{x}_t$ is classified according to the rule $y_t = \text{sign}(f(\mathbf{x}_t))$. However, there are multiple solutions, i.e. hyperplanes that can separate the two classes, but our goal is to find the one that gives the smallest generalization error. Support Vector Machines approach the problem through the concept of the *margin*: the smallest distance between the separating plane and the closest samples to the plane. Support Vector Machines search for

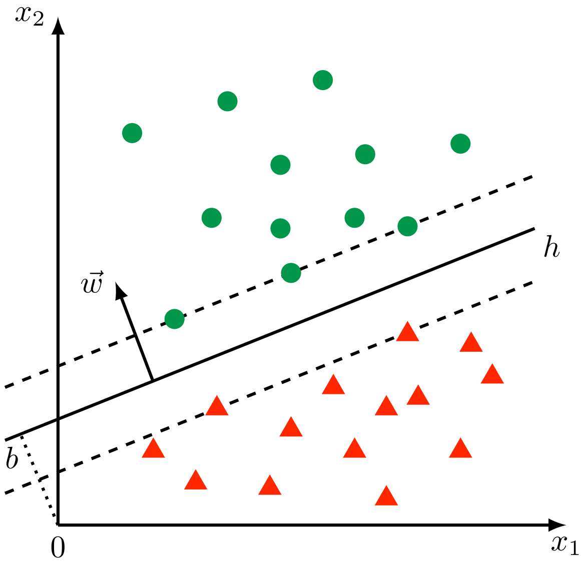


Figure A.1: Optimal separating hyperplane h (solid line) with maximal margins. Data samples that lie on the margins (dashed lines) are called support vectors. The two classes are in this case linearly separable.

the separating plane (the function $f(.)$) with the *largest margin*. The rationale is that the larger the margin, the higher the likelihood that novel test points are classified correctly. Figure A.1 illustrates how Support Vector Machines operate.

Let d_+ and d_- be the shortest distance from the hyperplane to the closest positive and negative example, respectively. The margin is thus defined as $d_+ + d_-$. It can be shown that $d_+ = d_- = \frac{1}{\|\mathbf{w}\|}$; hence, the margin that we want to maximize is $\frac{2}{\|\mathbf{w}\|}$, which is equivalent to minimizing $\frac{1}{2}\|\mathbf{w}\|$.

As defined above, any solution must satisfy the constraints

$$\begin{aligned} \mathbf{w^T x_i} + b &\geq +1, \quad \text{if } y_i = 1 \\ \mathbf{w^T x_i} + b &\leq -1, \quad \text{if } y_i = -1 \end{aligned} \tag{A.2}$$

which can be combined together as $y_i(\mathbf{w^T x_i} + b) - 1 \geq 0$. Thus, in order to maximize the margin, we wish to optimize the parameters $\mathbf{w}$ and b. We formulate the *primal* problem of the SVMs as

$$\begin{aligned} \text{minimize} \quad & \tfrac{1}{2}\|\mathbf{w}\|^2, \\ \text{subject to} \quad & y_i(\mathbf{w^T x_i} + b) - 1 \geq 0, \forall i\,. \end{aligned} \tag{A.3}$$

The data points for which $y_i(\mathbf{w^T x_i} + b) - 1 = 0$ are called *support vectors* since the location of the decision boundary is determined solely by this subset of the data points.

In order to solve the above constraint optimization problem, we rewrite Equation A.3 using the Lagrangian as

$$L(\mathbf{w}, b, \alpha) = \tfrac{1}{2}\|\mathbf{w}\|^2 - \textstyle\sum_{i=1}^{N} \alpha_i [y_i(\mathbf{w^T x_i} + b) - 1], \tag{A.4}$$

where $\alpha = (\alpha_1, \ldots, \alpha_N), \forall \alpha_i \geq 0$ is a set of Lagrangian multipliers. Thus, we must minimize L with respect to $\mathbf{w}$ and b, and maximize it with respect to α. Hence, the solution that minimizes the primal problem subject to the constraints is given by the *saddle point problem*:

$$\min_{\mathbf{w}} \max_{\alpha} L(\mathbf{w}, \alpha). \tag{A.5}$$

We set the derivatives of $L(\mathbf{w}, b, \alpha)$ with respect to $\mathbf{w}$ and b to zero and obtain

$$\begin{array}{rl} \mathbf{w}^* & = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x_i} \\ 0 & = \sum_{i=1}^{N} \alpha_i y_i \end{array} \tag{A.6}$$

The elimination of $\mathbf{w}$ and b in Equation A.4 using these conditions leads to the *dual* problem

$$\begin{array}{rl} \text{maximize} & L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j \mathbf{x_i}^T \mathbf{x_j}, \\ \text{subject to} & \sum_{i=1}^{N} \alpha_i y_i = 0 \\ & \alpha_i \geq 0, \forall i\,. \end{array} \tag{A.7}$$

The dual problem can be solved using classical quadratic programming based, for example, on constrained gradient descent.

The advantage of the dual formulation over the primal one is that the problem now only depends on $\mathbf{x_i}$ through the inner product $\mathbf{x_i}^T \mathbf{x_j}$ which we can substitute with *kernel matrices* $k(\mathbf{x_i}, \mathbf{x_j})$, thus projecting the data into a higher dimensional space where the separation of the two classes may be easier. Although the data set may not be linearly separable in the input space $\mathbf{x}$, it can often be separated linearly in the nonlinear feature space defined implicitly by the nonlinear kernel function.

A number of conditions (called the Karush-Kuhn-Tucker conditions) must hold at saddle points. They are derived partly from the primal problem by setting the derivatives with respect to $\mathbf{w}$ and b to zero. The constraints of the primal problem are also part of these conditions. Finally, the so-called "complementary slackness" constraint needs to be satisfied. The KKT conditions can thus be summarized as

$$\tfrac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \textstyle\sum_{i=1}^{N} \alpha_i y_i \mathbf{x_i} = 0 \tag{A.8}$$

$$\tfrac{\partial L}{\partial b} = \textstyle\sum_{i=1}^{N} \alpha_i y_i = 0 \tag{A.9}$$

$$y_i(\mathbf{w}^T \mathbf{x_i} - b) - 1 \geq 0 \tag{A.10}$$

$$\alpha_i \geq 0, \quad \forall i \tag{A.11}$$

$$\alpha_i[y_i(\mathbf{w}^T \mathbf{x_i} - b) - 1] = 0 \quad \text{(complementary slackness)} \tag{A.12}$$

These conditions are used to estimate the solution for b^* (after $\mathbf{w}^*$ has been found during training)

$$b^* = \textstyle\sum_j \alpha_j y_j \mathbf{x_j}^T \mathbf{x_i} - y_i, \quad i - \text{support vector} \tag{A.13}$$

The training instances where $\alpha_i > 0$ are the support vectors and they define the solution. Typically, α is very sparse (which means that the number of support vectors is low), i.e. not all kernel entries need to be evaluated to predict the class membership of a test sample.

So far, we have assumed that the training data is linearly separable. However, this rarely holds, and the proposed minimization problem does not have any solution if the two classes are not separable. To account for this case, the *soft margin* method will relax the constraints by introducing *slack variables* ξ_i that penalize (but nevertheless allow) mislabelled samples

$$\begin{aligned} \mathbf{w^T x_i} + b &\geq +1 + \xi_i, \qquad \text{if } y_i = 1, \\ \mathbf{w^T x_i} + b &\leq -1 - \xi_i, \qquad \text{if } y_i = -1, \end{aligned} \tag{A.14}$$

where $\xi_i \geq 0, \forall i$. To penalize the objective function for these constraint violations, we add a new term $C \sum_i \xi_i$ to the primal problem, where the constant C controls the tradeoff between a large margin and a small penalty error. Thus, the primal form can be rewritten as

$$\begin{array}{rl} \text{minimize} & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i, \\ \text{subject to} & y_i(\mathbf{w^T x_i} + b) - 1 + \xi_i \geq 0, \forall i \\ & \xi_i \geq 0, \forall i \end{array} \tag{A.15}$$

leading to the Lagrangian form

$$\begin{aligned} L(\mathbf{w}, b, \xi, \alpha, \mu) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i - \sum_{i=1}^{N} \alpha_i[y_i(\mathbf{w^T x_i} + b) - 1 + \xi_i] \\ - \sum_{i=1}^{N} \mu_i \xi_i. \end{aligned} \tag{A.16}$$

The dual form is derived using the KKT conditions to get rid of $\mathbf{w}$, b, and ξ

$$\begin{array}{rl} \text{maximize} & L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{ij} \alpha_i\alpha_j y_i y_j \mathbf{x_i}^T \mathbf{x_j}, \\ \text{subject to} & \sum_{i=1}^{N} \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C, \quad \forall i \end{array} \tag{A.17}$$

Surprisingly, the only difference to the quadratic programming problem from the linearly separable case, is the extra constraint (i.e. the upper limit) on the α_i multipliers. Also note that L_D can again be easily kernelized by replacing the inner products $\mathbf{x_i}^T\mathbf{x_j}$ with a kernel matrix $k(\mathbf{x_i}, \mathbf{x_j})$.

Throughout the thesis, for our analysis, we used the publicly available LIBSVM package, a standard implementation of various types of Support Vector Machines [Chang and Lin, 2001].

B
State-of-the-art saliency models

Throughout the present dissertation, we have extensively evaluated and compared our saliency prediction and modification approach to a number of baseline models of saliency. Two such state-of-the-art methods, the SUNDAy and the classic Itti and Koch model, are used recurringly in this work, and therefore, apart from their brief mention in Chapter 2, they merit further attention. Understanding the working principles behind these models can help placing our saliency model in the right context. In the following, we therefore briefly review the main steps of the algorithms employed by the two approaches.

B.1 The SUNDAy model

The model of Zhang et al. [2009], SUNDAy, defines saliency as the *self-information* of some low-level visual features. In the context of attentional modelling, self-information adequately quantifies the assumption that novel items draw attention. Self-information and the probability of a visual feature are inversely proportional, i.e. rarer features are more informative. In the formulation of Zhang et al. [2009], the probability distribution of the visual features are learned "through experience", from a large collection of natural videos. Figure B.1 shows the saliency map computed by SUNDAy on one of our videos.

The features used in their model are separable linear filters and they are computed on the intensity (I), red-green (RG), and blue-yellow (BY) channels of the videos, defined as $I = r + g + b$, $RG = r - g$, and $BY = b - \frac{r+g}{2} - \frac{\min(r,g)}{2}$. r, g, and b stand here for the red, green, and blue video components. The feature response functions take the form of $F = V * g * h$,

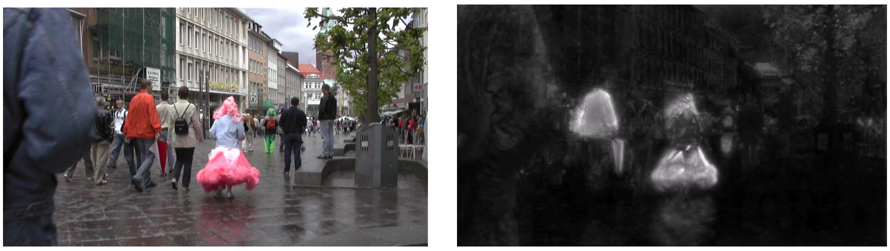


Figure B.1: The SUNDAy saliency map for one video frame.

where V is one of the above video channels, and g and h are filter components applied along the spatial and temporal dimensions, respectively. Difference of Gaussian filters (DoGs – with various σ's) are used as spatial filters:

$$g(x, y; \sigma) = \frac{1}{\sigma^2} e^{-\frac{x^2+y^2}{\sigma^2}} - \frac{1}{(1.6\sigma)^2} e^{-\frac{x^2+y^2}{(1.6\sigma)^2}}, \tag{B.1}$$

while the temporal components take the form of a Difference of Exponential (DoE):

$$h(t; \tau) = h'(t, 2\tau) - h'(t; \tau) \tag{B.2}$$

due to the resemblance of $h'(t, \tau)$ to an exponential distribution:

$$h'(t; \tau) = \frac{\tau}{1+\tau}(1+\tau)^t. \tag{B.3}$$

From each channel, all possible spatiotemporal combinations of five σ spatial and τ temporal scale parameters are extracted. Thus, overall, 75 feature responses are obtained. Since all filters are linear, the final filter response can be computed as

$$F(x, y, t; \sigma, \tau) = F'(x, y, t; \sigma, 2\tau) - F'(x, y, t; \sigma, \tau), \tag{B.4}$$

where $F'(x, y, t; \sigma, \tau) = V(x, y, t) * g(x, y; \sigma) * h'(t; \tau)$.

To learn the distribution for each feature, first the feature responses on about two hours of documentaries are computed, after which a generalized Gaussian distribution is fitted to the estimated distributions. The log-probability of a feature is written as

$$\log p(F_{i,j,k}) = -\left| \frac{f_{i,j,k}^{\theta_{i,j,k}}}{\varsigma_{i,j,k}} \right| + \text{const} \tag{B.5}$$

where $\theta_{i,j,k}$ is the shape parameter, and $\varsigma_{i,j,k}$ is the scale parameter of each of the 75 filters $f_{i,j,k}$ (spatial: $i = 1, \ldots, 5$, temporal: $j = 1, \ldots, 5$, colour: $k = 1, \ldots, 3$). The saliency of a location corresponds to the *self-information* ($-\log p(F = f)$) calculated as the sum of scaled and shaped filter responses:

$$\log s = -\log p(F = f) = \sum_{i=1}^{5} \sum_{j=1}^{5} \sum_{k=1}^{3} \left| \frac{f_{i,j,k}^{\theta_{i,j,k}}}{\varsigma_{i,j,k}} \right| + \text{const.} \qquad \boxed{\text{B.6}}$$

Source code for computing SUNDAy saliency maps is publicly available as part of the FastSaliency toolbox at `http://mplab.ucsd.edu/~nick/NMPT/`.

B.2 The Itti and Koch model

The biologically-inspired saliency framework of Itti et al. [1998, 2003] is perhaps the most popular model of bottom-up attention, against which all other approaches are compared. A schematic overview of the model architecture is given in Figure 2.3. The input video is here, too, decomposed into an intensity I ($I = \frac{r+g+b}{3}$) and two colour opponency channels RG and BY (as above). The decomposition is performed at nine spatial scales using Gaussian pyramids. Local orientation features (for four preferred orientations O_θ, $\theta \in \{0°, 45°, 90°, 135°\}$) are extracted, again on nine scales, by applying steerable filters to the intensity pyramid levels. Two types of temporal features, flicker (FL) and motion (M_θ) filters are used. The difference between the luminance of the current frame and that of the previous frame yields the flicker response. Motion features are extracted from spatially-shifted differences between steerable pyramids from adjacent frames. The steerable pyramids are those considered for orientation, and only shifts of one pixel orthogonal to the Gabor orientation are used.

To simulate centre-surround receptive fields, the authors perform across-scale subtraction of the individual feature maps. The centres are taken at pixels from the pyramid levels $c \in \{2, 3, 4\}$, while the surround at the corresponding pixels in the pyramid levels $s = c + \delta$, where $\delta \in \{3, 4\}$. Thus, six centre-surround maps are computed for each of the above features, yielding a total of 72 maps. Formally, the centre-surround feature maps are obtained as

$$\begin{aligned}&\mathcal{CS}_l(c,s) = |\mathcal{F}_l(c) \ominus \mathcal{F}_l(s)|, \quad \forall l \in L \\ &\text{with} \quad L = \{I, RG, RB, O_\theta, FL, M_\theta\}, \quad \theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}\end{aligned} \tag{B.7}$$

$\ominus$ denotes across-scale map subtraction, and $\mathcal{F}_l$ is either an intensity, colour, orientation, flicker, or a motion feature.

Next, the individual centre-surround feature maps are normalized and summed across scales using the across-scale addition operator $\oplus$:

$$\overline{\mathcal{F}_l} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{CS}_l(c,s)), \forall l \in L \tag{B.8}$$

The normalization step (with $\mathcal{N}(.)$) assumes an iterative convolution with a large Difference of Gaussian (DoG), which results in simultaneous self-excitation and inhibition of neighbouring locations, a behaviour associated with long-range connections in V1.

For colour, orientation, and motion the normalized feature maps are combined into a single *conspicuity map*. Thus for each feature type the following conspicuity maps are obtained:

$$\begin{aligned}&\textbf{Intensity} \quad \mathcal{C}_I = \overline{\mathcal{F}_I} \\ &\textbf{Colour} \quad \mathcal{C}_C = \overline{\mathcal{F}_{RG}} + \overline{\mathcal{F}_{BY}} \\ &\textbf{Orientation} \quad \mathcal{C}_O = \textstyle\sum_\theta \mathcal{N}(\overline{\mathcal{F}_{O_\theta}}) \\ &\textbf{Flicker} \quad \mathcal{C}_F = \overline{\mathcal{F}_{FL}} \\ &\textbf{Motion} \quad \mathcal{C}_M = \textstyle\sum_\theta \mathcal{N}(\overline{\mathcal{F}_{M_\theta}})\end{aligned} \tag{B.9}$$

These maps are again normalized and then summed into the final saliency map $\mathcal{S}$:

$$\mathcal{S} = \frac{1}{3}\left(\mathcal{N}(\mathcal{C}_I) + \mathcal{N}(\mathcal{C}_C) + \mathcal{N}(\mathcal{C}_O) + \mathcal{N}(\mathcal{C}_F) + \mathcal{N}(\mathcal{C}_M)\right) \tag{B.10}$$

From this map, winner-take-all (WTA) mechanisms select the most conspicuous location, which is attended to and then inhibited within a given radius. An iteration of WTA and inhibition-of-return steps assures the generation of attention shifts to locations of successively decreasing saliency.

In our work, we are concerned with the generation of saliency maps rather than scanpaths. Therefore, for all our analysis, WTA and inhibition-of-return mechanisms are turned off.

A complete real-time implementation of the above model is freely downloadable from `http://ilab.usc.edu/toolkit`.

Bibliography

Edward H. Adelson and Peter J. Burt. Image data compression with the Laplacian pyramid. In *Proceeding of the Conference on Pattern Recognition and Image Processing*, pages 218–223. Los Angeles, CA: IEEE Computer Society Press, 1981.

Yiannis Aloimonos, Isaac Weiss, and A. Bandyopadhyay. Active vision. *International Journal of Computer Vision*, 1(4):333–356, 1988.

C. H. Anderson, Peter J. Burt, and G. van der Wal. Change detection and tracking using pyramid transformation techniques. In *Proc. SPIE Conf. on Intelligent Robots and Computer Vision*, volume 579, pages 72–78, 1985.

Fred Attneave. Some Informational Aspects of Visual Perception. *Psychol Rev*, 61(3):183–193, May 1954.

Tamar Avraham and Michael Lindenbaum. Esaliency (extended saliency): Meaningful attention using stochastic image modeling. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(4):693–708, 2010.

Gökhan H. Bakır, Jason Weston, and Bernhard Schölkopf. Learning to find pre-images. *Advances in Neural Information Processing Systems*, 16:449–456, 2004.

Dana H. Ballard and Mary M. Hayhoe. Modelling the role of task in the control of gaze. *Visual Cognition*, 17(6-7):1185–204, 2009.

Horace B. Barlow. Possible principles underlying the transformations of sensory messages. In W. A. Rosenblith, editor, *Sensory Communication*, pages 183–192. MIT Press, Cambridge edition, 1961.

Erhardt Barth. Information technology for active perception: Itap. In *First GRP-Symposium, Sehen und Aufmerksamkeit im Alter, Benediktbeuren*, 2001.

Erhardt Barth and Andrew B. Watson. A geometric framework for nonlinear visual coding. *Optics Express*, 7(4):155–165, 2000.

Erhardt Barth, Terry Caelli, and Christoph Zetzsche. Image encoding, labeling, and reconstruction from differential geometry. *CVGIP: Graphical Models and Image Processing*, 55(6):428–446, November 1993.

Erhardt Barth, Michael Dorr, Martin Böhme, Karl R. Gegenfurtner, and Thomas Martinetz. Guiding the mind's eye: Improving communication and vision by external control of the scanpath. In Bernice E. Rogowitz, Thrasyvoulos N. Pappas, and Scott J. Daly, editors, *Human Vision and Electronic Imaging*, volume 6057 of *Proc. SPIE*, 2006. Invited contribution for a special session on Eye Movements, Visual Search, and Attention: a Tribute to Larry Stark.

Erhardt Barth, Michael Dorr, Eleonora Vig, Laura Pomarjanschi, and Cicero Mota. Efficient coding and multiple motions. *Vision Research*, 50(22): 2190–2199, 2010.

Wolfgang Becker. Saccades. In R. H. S. Carpenter, editor, *Vision & Visual Dysfunction Vol 8: Eye Movements*, pages 95–137. CRC Press, 1991.

James R. Bergen and Bela Julez. Rapid discrimination of visual patterns. *IEEE Trans Systems, Man, Cybernetics*, 13(5):857–863, 1983.

Chistopher Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

Rick S. Blum and Zheng Liu, editors. *Multi-Sensor Image Fusion and its Applications*. CRC Press, Boca Raton, FL, 2005.

Giuseppe Boccignone, Angelo Chianese, Vincenzo Moscato, and Antonio Picariello. Foveated shot detection for video segmentation. *IEEE Trans. Circuits Syst. Video Technol*, 15:365–377, 2005.

Martin Böhme, Michael Dorr, Christopher Krause, Thomas Martinetz, and Erhardt Barth. Eye movement predictions on natural videos. *Neurocomputing*, 69(16–18):1996–2004, 2006.

Martin Böhme, Michael Dorr, Thomas Martinetz, and Erhardt Barth. A temporal multiresolution pyramid for gaze-contingent manipulation of natural video. In Riad I. Hammoud, editor, *Passive Eye Monitoring*, chapter 10, pages 225–243. Springer, 2008.

Oliver Braddick and Ning Qian. The organization of global motion and transparency. In Johannes M Zanker and Jochen Zeil, editors, *Motion Vision - Computational, Neural, and Ecological Constraints*, pages 86–111. Springer Verlag, Berlin Heidelberg New York, 2001.

Neil Bruce and John Tsotsos. Saliency based on information maximization. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 155–162. MIT Press, Cambridge, MA, 2006.

Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

Peter J. Burt and Edward H. Adelson. A multiresolution spline with application to image mosaics. *ACM Transactions on Graphics*, 2(4):217–236, 1983.

Guy Thomas Buswell. *How People Look at Pictures: A Study of the Psychology of Perception in Art.* Chicago:University of Chicago Press, 1935.

Ran Carmi and Laurent Itti. Visual causes versus correlates of attentional selection in dynamic scenes. *Vision Research*, 46:4333–45, 2006.

Roger H. S. Carpenter. Oculomotor procrastination. In D. F. Fisher, R. A. Monty, and J. W. Senders, editor, *Eye Movements: Cognition and Visual Perception*, pages 237–246. Hillsdale, NJ: Lawrence Erlbaum, 1981.

Kelly Chajka, Mary Hayhoe, Brian Sullivan, Jeff Pelz, Neil Mennie, and Jason Droll. Predictive Eye Movements in Squash. *Journal of Vision*, 6 (6):481–481, 6 2006. ISSN 1534-7362.

Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

Cristianini Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods.* Cambridge University Press, 2000.

David Crundall, Peter Chapman, Nicola Phelps, and Geoffrey Underwood. Eye movements and hazard perception in police pursuit and emergency response driving. *Journal of Experimental Psychology*, 9(3):163–174, 2003.

Christine A. Curcio, Kenneth R. Sloan, Robert E. Kalina, and Anita E. Hendrickson. Human photoreceptor topography. *The Journal of Comparative Neurology*, 292:497–523, 1990.

Doug DeCarlo and Anthony Santella. Stylization and abstraction of photographs. *ACM Trans. Graph.*, 21:769–776, July 2002.

Robert Desimone and John Duncan. Neural Mechanisms of Selective Visual Attention. *Annual review of neuroscience*, 18(1):193–222, 1995.

Heiner Deubel. The time course of presaccadic attention shifts. *Psychological Research*, 72:630–640, 2008.

R. W. Ditchburn and B. L. Ginsborg. Vision with a stabilised retinal image. *Nature*, 170:36–37, 1952.

Michael Dorr. *Computational models and systems for gaze guidance*. Phd, University of Lübeck, Luebeck, Germany, Apr 2010.

Michael Dorr, Thomas Martinetz, Karl Gegenfurtner, and Erhardt Barth. Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision*, 10(10):1–17, 2010a.

Michael Dorr, Eleonora Vig, and Erhardt Barth. Colour saliency on video. In *Proceedings of Bionetics 2010 - 5th International ICST Conference on Bio-Inspired Models of Network, Information, and Computing Systems*, 2010b.

Andrew T. Duchowski, Nathan Cournia, and Hunter Murphy. Gaze-contingent displays: A review. *CyberPsychology & Behavior*, 7(6):621–634, 2004.

Wolfgang Einhäuser, Ueli Rutishauser, E. Paxon Frady, Swantje Nadler, Peter König, and Christof Koch. The relation of phase noise and luminance contrast to overt attention in complex visual stimuli. *Journal of Vision*, 6(11):1148–1158, 2006.

Lior Elazary and Laurent Itti. Interesting Objects are Visually Salient. *Journal of Vision*, 8(3):1–15, 3 2008.

Charles W. Eriksen and James D. St. James. Visual attention within and around the field of focal attention: A zoom lens model. *Perception and Psychophysics*, 40:225–40, 1986.

Georgios Evangelopoulos, Konstantinos Rapantzikos, Alexandros Potamianos, Petros Maragos, A. Zlatintsi, and Yair Avrithis. Movie summarization based on audiovisual saliency detection. In *Proc. IEEE Int'l Conf. on Image Processing (ICIP-08)*, pages 2528–2531, San Diego, CA, 2008.

John M. Findlay. Spatial and temporal factors in the predictive generation of saccadic eye movements. *Vision Research*, 21(3):347–354, 1981.

John M. Findlay and Iain D. Gilchrist, editors. *Active Vision: The Psychology of Looking and Seeing*, volume 37 of *Oxford Psychology Series*. Oxford University Press, 2003.

J. Randall Flanagan and Roland S. Johansson. Action plans used in action observation. *Nature*, 424(6950):769–771, Aug 2003.

Simone Frintrop, Patric Jensfelt, and Henrik I. Christensen. Attentional landmark selection for visual slam. In *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'06)*, Beijing, China, 2006.

Dashan Gao and Nuno Vasconcelos. Decision-theoretic saliency: Computational principles, biological plausibility, and implications for neurophysiology and psychophysics. *Neural Computation*, 21(1):239–271, 2009.

Dashan Gao, Sunhyoung Han, and Nuno Vasconcelos. Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 31(6): 989–1005, 2009.

Wilson S. Geisler and Jeffrey S. Perry. A real-time foveated multiresolution system for low-bandwidth video communication. In Bernice E. Rogowitz and Thrasyvoulos N. Pappas, editors, *Human Vision and Electronic Imaging: SPIE Proceedings*, pages 294–305. 1998.

Benno Gesierich, Angela Bruzzo, Giovanni Ottoboni, and Livio Finos. Human gaze behaviour during action execution and observation. *Acta Psychologica*, 128(2):324 – 330, 2008.

Robert B. Goldstein, Russell L. Woods, and Eli Peli. Where people look when watching movies: Do all viewers look at the same place? *Computers in Biology and Medicine*, 3(7):957–64, 2007.

Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Addison-Wesley Longman Publishing Co., Inc, Boston, MA, USA, second edition, 2001.

Gösta H. Granlund and Hans Knutsson. *Signal Processing for Computer Vision*. Kluwer, Dordrecht, 1995.

Chenlei Guo and Liming Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *Trans. on Image Processing*, 19(1):185–198, 2010.

Chris Harris and Mike J. Stephens. A combined corner and edge detector. In *Proc. The Fourth Alvey Vision Conference*, pages 147–152, 1988.

Mary Hayhoe and Dana Ballard. Eye movements in natural behavior. *Trends in Cognitive Sciences*, 9(4):188–94, 2005.

John M. Henderson and Andrew Hollingworth. Eye movements during scene viewing: An overview. In Geoffrey Underwood, editor, *Eye Guidance in Reading and Scene Perception*, pages 269–93. Elsevier Science Ltd, 1998.

Edmund Burke Huey. Preliminary experiments in the physiology and psychology of reading. *American Journal of Psychology*, 9(4):575–586, 1898.

D.E. Irwin. Visual memory within and across fixations. *Eye movements and visual cognition: Scene perception and reading. New York: Springer-Verlag*, pages 146–165, 1992.

Laurent Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing*, 13(10):1304–1318, Oct 2004a.

Laurent Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing*, 13(10):1304–1318, Oct 2004b.

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

Laurent Itti and Pierre Baldi. Bayesian Surprise Attracts Human Attention. *Vision Research*, 49(10):1295–1306, May 2009.

Laurent Itti and Christoph Koch. Feature combination strategies for saliency-based visual attention systems. *Journal of Electronic Imaging*, 10(1):161–169, Jan 2001.

Laurent Itti, Christoph Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, Nov 1998.

Laurent Itti, N. Dhavale, and F. Pighin. Realistic avatar eye and head animation using a neurobiological model of visual attention. In B. Bosacchi, D. B. Fogel, and J. C. Bezdek, editors, *Proc. SPIE 48th Annual International Symposium on Optical Science and Technology*, volume 5200, pages 64–78, Bellingham, WA, Aug 2003. SPIE Press.

Laurent Itti, Geraint Rees, and John K. Tsotsos. *Neurobiology of Attention*. Academic Press, December 2005.

Bernd Jähne and Horst Haußecker, editors. *Computer Vision and Applications*. Academic Press, 2000.

Bernd Jähne, Horst Haußecker, and Peter Geißler, editors. *Handbook of Computer Vision and Applications*. Academic Press, San Diego, USA, 1999.

William James. *The Principles of Psychology*. Henry Holt, 1890. On-line edition at http://psychclassics.yorku.ca/James/Principles.

William James, Frederick Burkhardt, and Fredson Bowers. *The principles of psychology*, volume 1. Harvard University Press, 1981.

Tilke Judd, Krista Ehinger, Fre'do Durand, and Antonio Torralba. Learning to Predict Where Humans Look. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2009.

Peter Kasarskis, Jennifer Stehwien, Joey Hickox, and Anthony Aretz. Comparison of expert and novice scan behaviors during vfr flight. In *The 11th International Symposium on Aviation Psychology*, Columbus, OH, 2001.

Wolf Kienzle, Bernhard Schölkopf, Felix A. Wichmann, and Matthias O. Franz. How to find interesting locations in video: a spatiotemporal interest point detector learned from human eye movements. In *Proceedings of* *the 29th Annual Symposium of the German Association for Pattern Recognition (DAGM 2007)*, pages 405–414, Berlin, Germany, 2007a. Springer Verlag.

Wolf Kienzle, Felix A. Wichmann, Bernhard Schölkopf, and Matthias O. Franz. A Nonparametric Approach to Bottom-Up Visual Saliency. In *Advances in Neural Information Processing Systems*, pages 689–696, Cambridge, Mass. USA, 2007b. MIT Press.

Christoph Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4(4):219–227, 1985.

Helga Kolb, Eduardo Fernandez, and Ralph Nelson. Webvision – the organization of the retina and visual system. `http://webvision.med.utah.edu/`, 2010.

James T. Y. Kwok and Ivor W. H. Tsang. The pre-image problem in kernel methods. *Neural Networks, IEEE Transactions on*, 15(6):1517–1525, 2004.

Kai Labusch, Erhardt Barth, and Thomas Martinetz. Sparse Coding Neural Gas: Learning of Overcomplete Data Representations. *Neurocomputing*, 72(7-9):1547–1555, 2009.

Michael F. Land and Sophie Furneaux. The knowledge base of the oculomotor system. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 352:1231–1239, 1997.

Michael F. Land and Mary Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision Research*, 41:3559–65, 2001.

Michael F. Land and D. N. Lee. Where we look when we steer. *Nature*, 369: 742–744, 1994.

Michael F. Land and Peter McLeod. From eye movements to actions: how batsmen hit the ball. *Nature Neuroscience*, 3:1340–1345, 2000.

Michael F. Land, Neil Mennie, and Jennifer Rusted. The roles of vision and eye movements in the control of activities of daily living. *Perception*, 28: 1311–1328, 1999.

Ivan Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2):107–123, 2005.

Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 99, 2010.

Vijay Mahadevan and Nuno Vasconcelos. Spatiotemporal saliency in dynamic scenes. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32:171–177, 2010.

Stephane G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans Pattern Anal Machine Intell*, 11(7): 674693, July 1989.

David Marr. *Vision*. W H Freeman, 1982.

Susana Martinez-Conde, Stephen L. Macknik, and David H. Hubel. The role of fixational eye movements in visual perception. *Nature Reviews Neuroscience*, 5(3):229–240, 2004.

Ethel Matin. Saccadic suppression: A review and an analysis. *Psychological Bulletin*, 81(12):899–917, 1974.

Ann McNamara, Reynold Bailey, and Cindy Grimm. Search task performance using subtle gaze direction with the presence of distractions. *ACM Transactions on Applied Perception*, 6(3):1–19, 2009.

Neil Mennie, Mary Hayhoe, and Brian Sullivan. Look-ahead fixations: Anticipatory eye movements in natural tasks. *Exp Brain Res*, 179(3):427–442, May 2007.

Olivier Le Meur, Patrick Le Callet, Dominique Barba, and Dominique Thoreau. A coherent computational approach to model bottom-up visual attention. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28 (5):802–817, 2006.

Cicero Mota and Erhardt Barth. On the uniqueness of curvature features. In G. Baratoff and H. Neumann, editors, *Dynamische Perzeption*, volume 9 of *Proceedings in Artificial Intelligence*, pages 175–178, Köln, 2000. Infix Verlag.

Cicero Mota, Ingo Stuke, and Erhardt Barth. Analytic solutions for multiple motions. In *Proc. IEEE Int. Conf. Image Processing*, volume 2, pages 917–920, Thessaloniki, Greece, October 7-10, 2001. IEEE Signal Processing Soc.

Cicero Mota, Michael Dorr, Ingo Stuke, and Erhardt Barth. Categorization of Transparent-Motion Patterns Using the Projective Plane. *International Journal of Computer and Information Science*, 5(2):129–140, 2004a.

Cicero Mota, Ingo Stuke, Til Aach, and Erhardt Barth. Estimation of multiple orientations at corners and junctions. In *26th Pattern Recognition Symposium (DAGM'04), Tübingen*, pages 163–170, 2004b.

Cicero Mota, Ingo Stuke, and Erhardt Barth. The intrinsic dimension of multispectral images. In *MICCAI Workshop on Biophotonics Imaging for Diagnostics and Treatment*, pages 93–100, 2006.

Vidhya Navalpakkam and Laurent Itti. Modeling the influence of task on attention. *Vision Research*, 45(2):205–231, Jan 2005.

Ulrik Neisser. *Cognitive Psychology.* New York: Appleton, 1967.

Alexandre Ninassi, Olivier Le Meur, Patrick Le Callet, and Dominique Barba. Does where you gaze on an image affect your perception of quality? applying visual attention to image quality metric. In *Proceedings of the International Conference on Image Processing (ICIP)*, pages 169–172, 2007.

Calvin F. Nodine and Claudia Mello-Thoms. The nature of expertise in radiology. In Richard L. Van Metter, Jacob Beutel, and Harold L. Kundel, editors, *Handbook of Medical Imaging, Volume 1. Physics and Psychophysics*. SPIE Press, Bellingham, WA, 2000.

David Noton and Lawrence Stark. Eye movements and visual perception. *Scientific American*, 224(6):34–43, 1971.

Marcus Nyström and Kenneth Holmqvist. Effect of compressed offline foveated video on viewing behavior and subjective quality. *ACM Trans. Multimedia Comput. Commun. Appl.*, 6(1):4:1–4:14, February 2010.

Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381: 607–609, 1996.

Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37:3311–3325, 1997.

Edgar E. Osuna, Robert Freund, and Federico Girosi. Support vector machines: Training and applications. Technical report, 1997.

Nabil Ouerhani, Javier Bracamonte, Heinz Hugli, Michael Ansorge, and Fausto Pellandini. Adaptive color image compression based on visual attention. In *Proc. of the International Conference of Image Analysis and Processing (ICIAP)*, pages 416–421, 2001.

Stephen E. Palmer. *Vision Science: Photons to Phenomenology.* MIT Press, 1999.

Derrick J. Parkhurst and Ernst Niebur. Variable-resolution displays: A theoretical, practical, and behavioral evaluation. *Human Factors*, 44(4): 611–29, 2002.

Jeff B. Pelz and Roxanne Canosa. Oculomotor behavior and perceptual strategies in complex tasks. *Vision Research*, 41:3587–3596, 2001.

Michael I. Posner. Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32:3–25, 1980.

Christoph Rasche and Karl Gegenfurtner. Orienting during gaze guidance in a letter-identification task. *Journal of Eye Movement Research*, 3(4): 1–10, 2010.

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

Pamela Reinagel and Anthony M. Zador. Natural scene statistics at the centre of gaze. *Network: Comput Neural Syst*, 10:341–350, 1999.

Eyal M. Reingold, Lester C. Loschky, George W. McConkie, and David M. Stampe. Gaze-contingent multiresolutional displays: An integrative review. *Human Factors*, 45(2):307–28, 2003.

Ronald A. Rensink. The dynamic representation of scenes. *Visual Cognition*, 7:17–42, 2000.

Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2:1019–1025, 1999.

Giacomo Rizzolatti, Lucia Riggio, Isabella Dascola, and Carlo Umilta. Reorienting attention across the vertical and horizontal meridians: Evidence in favor of a premotor theory of attention. *Neuropsychologia*, 25:31–40, 1987.

Francesco Di Russo, Sabrina Pitzalis, and Donatella Spinelli. Fixation stability and saccadic latency in élite shooters. *Vision Research*, 43(17):1837 – 1845, 2003.

Ueli Rutishauser, Dirk Walther, Christof Koch, and Pietro Perona. Is bottom-up attention useful for object recognition. In *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 37–44, 2004.

Anthony Santella, Maneesh Agrawala, Doug DeCarlo, David Salesin, and Michael Cohen. Gaze-based interaction for semi-automatic photo cropping. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, CHI '06, pages 771–780, New York, NY, USA, 2006. ACM.

Cordelia Schmid, Roger Mohr, and Christian Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37 (2):151–172, 2000.

Bernhard Schölkopf and Alexander J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. the MIT Press, 2002.

Alexander C. Schütz, Doris I. Braun, and Karl R. Gegenfurtner. Improved visual sensitivity during smooth pursuit eye movements: temporal and spatial characteristics. *Visual Neuroscience*, 26:329–340, 2009.

Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio. Robust object recognition with cortex-like mechanisms.

IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(3): 411–426, 2007.

Claude E. Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, 1949.

Hamid R. Sheikh, Brian L. Evans, and Alan C. Bovik. Real-time foveation techniques for low bit rate video coding. *Real-Time Imaging*, 9(1):27–40, 2003.

Christian Siagian and Laurent Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(2):300–312, 2007.

A. C. Smit and J. A. M. Van Gisbergen. A short-latency transition in saccade dynamics during square-wave tracking and its significance for the differentiation of visually-guided and predictive saccades. *Experimental Brain Research*, 76:64–74, 1989.

Tim J. Smith and John M. Henderson. Edit blindness: The relationship between attention and global change blindness in dynamic scenes. *Journal of Eye Movement Research*, 2(2):1–17, 2008.

Sara L. Su, Frédo Durand, and Maneesh Agrawala. An inverted saliency model for display enhancement. In *Proceedings of 2004 MIT Student Oxygen Workshop*, pages 119–124, 2004.

Bernard Marius 't Hart, Johannes Vockeroth, Frank Schumann, Klaus Bartl, Erich Schneider, Peter König, and Wolfgang Einhäuser. Gaze allocation in natural stimuli: Comparing free exploration to head-fixed viewing conditions. *Visual Cognition*, 17(6/7):1132–1158, 2009.

Benjamin W. Tatler, Roland J. Baddeley, and Iain D. Gilchrist. Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, 45:643–659, 2005.

Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.

Monica A. Trifas, John M. Tyler, and Oleg S. Pianykh. Applying multiresolution methods to medical image enhancement. In *ACM-SE 44: Pro-*

ceedings of the 44th annual Southeast regional conference, pages 254–259, New York, NY, USA, 2006. ACM.

Po-He Tseng, Ran Carmi, Ian G. M. Cameron, Douglas P. Munoz, and Laurent Itti. Quantifying Center Bias of Observers in Free Viewing of Dynamic Natural Scenes. *Journal of Vision*, 9(7):1–16, 7 2009.

Geoffrey Underwood, Nicola Phelps, Chloe Wright, Editha van Loon, and Adam Galpin. Eye fixation scanpaths of younger and older drivers in a hazard perception task. *Ophthal. Physiol. Opt.*, 25:346–356, 2005.

Vladimir Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.

Eleonora Vig, Michael Dorr, and Erhardt Barth. Efficient visual coding and the predictability of eye movements on natural movies. *Spatial Vision*, 22 (5):397–408, 2009.

Eleonora Vig, Michael Dorr, and Erhardt Barth. Contribution of spatio-temporal intensity variation to bottom-up saliency. In *Proceedings of Bionetics 2010 - 5th International ICST Conference on Bio-Inspired Models of Network, Information, and Computing Systems*, 2010a.

Eleonora Vig, Michael Dorr, Thomas Martinetz, and Erhardt Barth. A learned saliency predictor for dynamic natural scenes. In K. Diamantaras, W. Duch, and L. S. Iliadis, editors, *ICANN 2010, Part III*, volume 6354 of *Lecture Notes in Computer Science*, pages 52–61, Thessaloniki, Greece, 2010b. Springer.

Eleonora Vig, Michael Dorr, and Erhardt Barth. Learned saliency transformations for gaze guidance. In Bernice E Rogowitz and Thrasyvoulos N Pappas, editors, *Human Vision and Electronic Imaging XVI*, volume 7865, pages W1–11. SPIE-IS&T, 2011a.

Eleonora Vig, Michael Dorr, Thomas Martinetz, and Erhardt Barth. Eye movements show optimal average anticipation with natural dynamic scenes. *Cognitive Computation*, 3(1):79–88, 2011b.

Eleonora Vig, Michael Dorr, Thomas Martinetz, and Erhardt Barth. Intrinsic dimensionality predicts the saliency of natural dynamic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(6):1080–1091, 2012. ISSN 0162-8828.

Hermann von Helmholtz. *Treatise on Physiological Optics*, volume 3. New York: Dover, 1962, 3rd edition, 1866.

Jeremy M. Wolfe. Visual Search. In Harold Pashler, editor, *Attention*, pages 13–73. Psychology Press, 1998.

David S. Wooding. Eye movements of large populations: II. Deriving regions of interest, coverage, and similarity using fixation maps. *Behavior Research Methods, Instruments, & Computers*, 34(4):518–28, 2002.

Alfred L. Yarbus. *Eye Movements and Vision*. Plenum Press, New York, 1967.

Christoph Zetzsche and Erhardt Barth. Fundamental limits of linear filters in the visual processing of two-dimensional signals. *Vision Research*, 30: 1111–1117, 1990.

Christoph Zetzsche, Erhardt Barth, and Bernhard Wegmann. The importance of intrinsically two-dimensional image features in biological vision and picture coding. In Andrew B. Watson, editor, *Digital Images and Human Vision*, pages 109–38. MIT Press, October 1993.

Christoph Zetzsche, Kerstin Schill, Heiner Deubel, Gerhard Krieger, E. Umkehrer, and S. Beinlich. Investigation of a sensorimotor system for saccadic scene analysis: an integrated approach. In R. Pfeifer et al., editor, *From Animals to Animats 5: Proceedings of the Fifth International Conference on Simulation of Adaptive Behavior*, volume 5, pages 120–126. MIT Press, Cambridge, 1998.

Lingyun Zhang, Matthew H. Tong, Tim K. Marks, Honghao Shan, and Garrison W. Cottrell. SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):1–20, 12 2008. ISSN 1534-7362.

Lingyun Zhang, Matthew H. Tong, and Garrison W. Cottrell. SUNDAy: Saliency Using Natural Statistics for Dynamic Analysis of Scenes. In *Proceedings of the 31st Annual Cognitive Science Conference, Amsterdam, Netherlands*, 2009.