HELSINKI UNIVERSITY OF TECHNOLOGY
Faculty of Electronics, Communications and Automation

Lauri Ahonen

A COMPUTATIONAL APPROACH TO ESTIMATION OF CROWDING IN NATURAL IMAGES

Thesis submitted for examination for the degree of Master of Science in Technology

In Helsinki, 3rd of December, 2008

Thesis supervisor: Professor Mikko Sams
Thesis instructor: Risto Näsänen, Ph.D., Docent of Experimental Psychology

Author:	Lauri Ahonen
Degree Programme:	Communications Engineering
Major Subject:	Cognitive Technology
Minor Subject:	Bioadaptive Technology

Title:	A Computational Approach to Estimation of Crowding in Natural Images
Number of Pages:	12+71
Chair:	S-114 Cognitive Technology
Supervisor:	Professor Mikko Sams
Instructor:	Risto Näsänen, Ph.D., Docent of Experimental Psychology

Crowding is a phenomenon where the identification of objects in peripheral vision is deteriorated by the presence of nearby targets. Crowding therefore reduces the extent of visual span, i.e. information intake during a single eye fixation. It is, thus, a limiting factor of many everyday tasks, such as reading. The phenomenon is due to wide area feature integration in the higher levels of visual processing. Despite the critical role of the phenomenon, complex natural images have so far not been used in the research of crowding. The purpose of the present study was to determine how the crowding effect affects object recognition in complex natural images, and whether the magnitude of the crowding could be modelled using methods introduced below.

The actual magnitude of the crowding effect was determined experimentally by measuring contrast thresholds for letter targets of different sizes on various natural image backgrounds. The results of the experiments were analyzed to evaluate the developed methods. The methods are based on image statistics and clutter modelling. Clutter models assess the complexity in the image. The image statistics and the clutter models were combined with basic knowledge of the crowding effect. In addition, an early visual system model was incorporated to assess the role of the visual acuity across the visual field.

The developed models predicted the induced crowding effect in an arbitrary natural image. The model of the visual system contributed to the results, as well. The differences between the methods for assessing the image properties were, however, negligible. Contrast energy, the simplest measure, can be regarded as the most efficient.

Natural images can cause very strong crowding effects. The conclusion is that predicting quantitative dimensions of the crowding effect in an arbitrary image is viable. However further research of the subject is necessary for developing the models. Computational assessment of the crowding effect potentially can be applied to e.g. user interface design, assessing information visualization techniques, and the development of augmented reality applications.

Keywords: Crowding, Visual system, Image statistics, Clutter, Natural images, Modelling.

TEKNILLINEN KORKEAKOULU
Elektroniikan, tietoliikenteen ja automaation tiedekunta

DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Lauri Ahonen
Koulutusohjelma:	Tietoliikennetekniikan koulutusohjelma
Pääaine:	Kognitiivinen teknologia
Sivuaine:	Bioadaptiivinen tekniikka

Työn nimi:	Laskennallinen malli ärsyketungoksen arvioimiseen luonnollisissa kuvissa
Sivumäärä:	12+71
Professuuri:	S-114 Kognitiivinen Teknologia
Valvoja:	Professori Mikko Sams
Ohjaaja:	FT, Kokeellisen psykologian dosentti, Risto Näsänen

Ärsyketungos on ilmiö, jonka vuoksi ääreisnäössä olevien kohteiden tunnistus heikkenee, mikäli kohteen läheisyydessä on häiriöärsykkeitä. Ärsyketungos on siis havaintokentän, eli yhdellä fiksaatiolla havaittavan kohdemäärän, kokoa rajoittava tekijä. Tämä johtuu näköjärjestelmän korkeammilla tasoilla tapahtuvasta laajasta piirreintegraatiosta kohteen ympäristössä. Kriittisestä roolistaan huolimatta ärsyketungosilmiötä ei ole tutkittu monimutkaisilla luonnollisilla kuvilla (mikä tahansa valokuva). Tässä tutkimuksessa tutkittiin, kuinka ärsyketungos vaikuttaa kohteen havaitsemiseen monimutkaisissa luonnollisissa kuvissa. Lisäksi haluttiin selvittää voidaanko ilmiön voimakkuutta ennustaa näköjärjestelmä malleja ja kuvien tilastollisia ominaisuuksia käyttäen.

Ärsyketungos-ilmiön voimakkuus määritettiin kokeellisesti, mittaamalla kontrasti kynnyksiä erikokoisille kirjainkohteille, jotka sijaitsivat luonnollisen kuvan päällä. Kokeellisen osan tuloksilla validoitiin kehitettyjä metodeja. Metodit perustuivat kuvan tilastollisiin ominaisuuksiin ja 'clutter-malleihin'. Tilastolliset ominaisuudet ja 'clutter-mallit' yhdistettiin sekä tutkimustietoon ärsyketungoksen ominaisuuksista, että näköjärjestelmän tunnettuihin ominaisuuksiin. Näköjärjestelmä huomioimalla pyrittiin arvioimaan spatiaalisesta näöntarkkuuden vaihtelusta aiheutuvia muutoksia kuvan tilastollisiin ominaisuuksiin.

Kehitetyt metodit ennustivat mielivaltaisen kuvan aiheuttaman ärsyketungos-ilmiön voimakkuuden. Myös näköjärjestelmän malli vaikutti tuloksiin. Erot eri laskentamallien välillä olivat kuitenkin merkityksettömiä. Täten yksinkertaisinta metodia, jossa laskettiin kontrastienergiaa, voidaan pitää tehokkaimpana.

Luonnolliset kuvat voivat aiheuttaa voimakkaan ärsyketungos-ilmiön. Pääteltiin, että ilmiö voidaan ennustaa kohtuullisella tarkkuudella jo nykyisellä tietämyksellä, mutta lisätuntemus ilmiön syistä ja mekanismeista mahdollistaisi tarkempien mallien kehittämisen. Tällaisilla malleilla on sovellutuskohteita esimerkiksi käyttöliittymien suunnitellussa, informaation visualisoinnin arvioinnissa ja lisätyn todellisuuden sovellusten kehityksessä.

Avainsanat: Ärsyketungos, Näköjärjestelmä, Kuvan tilastolliset ominaisuudet, Luonnolliset kuvat, Mallintaminen.

ACKNOWLEDGEMENTS

I first wish to thank my mentor and instructor Risto Näsänen for his enormous contribution to this work, and for many inspirational discussions about research.

This thesis was written at the Finnish Institute of Occupational Health (FIOH). I would like to express my sincere gratitude to the director of the centre, Kiti Müller, for the opportunity to work in the FIOH, and for paving my way to the world of research. To my esteemed supervisor Mikko Sams, I express my wholehearted gratitude for supporting my work.

I am indebted to all my colleagues at the FIOH for the fruitful atmosphere. To Andreas Henelius and Kristian Lukander I owe my sincere thanks for constructive comments and help in practical problems with the work. A special mentioning goes to all the participants of the experimental part.

My warmest thanks go to my friends and my family for their patience and support, particularly to my dearest Liisa Lalu, who served as my inspiration during my writing process.

Helsinki, December 3, 2008 Lauri V. Ahonen

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BWT	Berkeley wavelet transform, fast transformation for describing the RFs
FFT	Fast Fourier transformation, an algorithm for Fourier transformation
fMRI	Functional magnetic resonance imaging, a neuroimaging technique.
cRF	Classical receptive field, spatial convergence of feedforward connections
DC-term	Zero frequency value of an image
DoG	Difference of Gaussian, a filter type
GUI	Graphical user interface
IT	Inferior temporal gyrus, an area for object recognition (Brodmann area 20)
L-cone	Cone cell with maximal absorption efficiency at wavelength ($l_{\max}$) of 560 nm
LCD	Liquid crystal display, display in which a liquid polarizes the light
LGN	Lateral geniculate nucleus, processing centre for cells signalling from the retina
M-cone	Cone cell with maximal absorption efficiency at wavelength ($l_{\max}$) of 530 nm
M-pathway	Magnocellular pathway, detailed visual processing pathway
MT	Middle temporal gyrus, also known as V5 (Brodmann area 21)
MVC	Model-view-controller, a programming paradigm
P-pathway	Parvocellular pathway, fast visual processing pathway
PSC	Post-synaptic cell
RF	Receptive field, spatial convergence area of neurons
RGB	Red-green-blue, colour components in an additive colour model, e.g. in displays

S-cone	Cone cell with maximal absorption efficiency at wavelength ($l_{\max}$) of 420 nm
sd	Standard deviation
SciPy	Scientific Python, an algorithm library for the Python language
tRF	Total receptive field, all connections to a ganglion cell
V1	Primary visual cortex (Brodmann area 17)
V2	Secondary visual cortex (Brodmann area 18)
V3	Associative visual cortex (Brodmann asre 19)
V4	Extrastriate visual cortex, found to be active e.g. in the crowding effect

CHAPTER

ONE

INTRODUCTION

The terms ‘vista’ and ‘hearsay’ are an example that emphasizes the fundamental role of vision in the acquiring of information. Of all the senses, vision is considered as “the window to reality”. This, however, is not entirely true. Vision is rather a process in which a model of the environment is created using information in ambient light.

In reality, a tiny fraction of the information provided by the light in the environment is processed by the visual system, and the observer becomes aware of even tinier fraction. The incompleteness of the visual system creates an illusion, that we are, at all times, fully visually aware of our surroundings. However, the structure and complexity of what is perceived largely determines which part of the visual information one becomes aware of.

Visual perception consists of fairly well understood processes of feature detection and almost completely obscure processes of feature integration. In these processes the conception of the surrounding reality is created. The detection processes decompose the information and integration processes lead to object recognition and awareness of spatial positioning of objects.

Feature integration combines and compresses the information provided by feature detection. The compression leads to partial unawareness about ambient objects. The feature integration processes are largely unknown, but few exceptions exist. An example of early feature integration is the formation of illusory contours. Illusory contours are perceived in locations of apparent contours and can be detected with fMRI in the visual areas of the brain [1].

The purpose of the present work is to study one of the most important by-products of the feature integration processes: the crowding effect. The crowding effect is a phenomenon that deteriorates peripheral vision. Because of crowding effect, the *identification* of e.g. a letter in peripheral view is much harder in the presence of nearby objects, although the ability to *detect* the letter is unmodulated [2]. The preservation of the ability to detect the target, suggests that the effect is purely due to feature integration processes and not due to feature detection inability. There are several studies about the nature of the crowding effect and how it is related to feature integration processes of the brain.

The crowding effect has been suggested to be the restricting phenomenon for object recognition. It is a real-world phenomenon with which designers of advertisements, people working with information visualizations, and user-interfaces struggle. The phenomenon is for example a determining factor for the saccade, i.e. the eye movement, frequency in reading [3].

The characteristics that differentiate crowding from other masking effects in human vision are

- Crowding preserves the ability to detect a target
- The distracters can be non-overlapping
- Crowding depends on eccentricity of a target

The crowding effect was first quantified by Bouma [4]. Bouma suggested that the effect is related to the eccentricity of the target, i.e. how far in the peripheral vision the target is located. He stated that flanking objects closer than $d < 0.5\phi$ modulate the ability to identify the target, where ϕ is the distance of the target from the fovea, i.e. the most accurate part of vision, expressed in visual angle.

Because crowding is a higher-level phenomenon in visual processing it also includes many unsolved aspects. One of the main questions is the definition of an object. How is the distinction between figure and ground made e.g. in natural images, i.e. images one encounters in everyday life. Studies show that the crowding effect is largely determined by the complexity of the visual pattern (target and flankers) [5]. Thus, if the target and the flankers together produce an entity, the feature integration process is different and crowding may be suppressed.

Another important aspect in the study of the crowding effect is its relation to perceptual span. Perceptual span is the area of information intake with one glance. Perceptual span

is closely related to crowding examined in the peripheral vision [6]. Also, the relation to attention is under intensive debate [7].

Previous research has studied the crowding effect using a plain background with distinct flanking objects. The present study was established to examine the crowding effect caused by natural images. Examining the phenomenon in real complex imagery is crucial due to the fundamental role of the effect in object recognition. The protocol of the study consists of the evaluation of the crowding effect and outlining the influence of the early visual system on the effect.

Computational models are incorporated in the research of visual processing to simulate and evaluate the findings of psychophysical and physiological study. Computational modelling have revealed unforeseen relations and facts of the neural basis of the processing. In the present study, elementary models for assessing the influence of the crowding effect on natural images were developed. The models are not neuronal but based on image statistics. However, there is a certain degree of "bio-inspiration" in some of the elementary models. In these models the features, i.e. the low-level image properties, are extracted from the image in a similar manner to the early visual system. In addition to the feature extraction the models also integrate the psychophysical results of the characteristics of the crowding effect.

The extracted features were local contrast, contour orientation and entropy in the area around the target. With the feature extraction, an assessment of the information intensity of the image patch was generated. The information intensity that deteriorates the visual perception is referred to as clutter. Clutter is an important concept in visual perception. Rosenholtz et al. [8] define clutter as an excess of objects leading to degradation of performance in visual perception.

The concept of clutter is adapted here to measure the crowding effect. The connection between the concepts of crowding and clutter is not very simple. However the results show coarse but congruent correlation between basic image properties and the intensity of the crowding effect.

The structure of the thesis is as follows: The second chapter presents the literature of the related subjects, i.e. the crowding effect, clutter and early visual system in general. The third chapter introduces the methods used and the experimental part of the study. The methods are the mathematical procedures used to resolve the image statistics and the experiment is the psychophysical paradigm used. The results are presented in chapter four. The main findings are discussed and the conclusions are presented in chapter five. The program used in this study is presented in Appendix A.

CHAPTER

TWO

REVIEW OF THE LITERATURE

This chapter reviews literature on the applicable parts of the wide field of vision science essential to the purpose of the study. The physiology of the human visual system is reviewed and related aspects of modelling are presented, in particular the phenomenon of crowding.

2.1 Physiology of the Visual System

The human visual system consists of the eyes and a considerably large part of the central nervous system that is linked to them. The system can be divided into following functional parts: the optics of the eye, the retina in the fundus, the intermediate segments before the visual cortex (LGN and superior colliculus) and, finally, a diversity of parts of the neo-cortex, for example the areas called V1, V2, MT, and IT.

2.1.1 Optics of the Eye

The term ‘visible light’ refers to a certain energy range of electromagnetic radiation. Light travels to the neural parts of the visual system via the optical system of the eye. The optical system causes aberrations in the travelling light [9]. The optical system is, however, quite efficient and retains most of the information unlike the latter parts of the visual system, in

particular at locations outside the central visual field [10]. Thus, the optical system and its transfer function can be omitted when considering the peripheral vision. However, optics play an important role e.g. in preventing aliasing in the peripheral vision that is due to the low sampling density of receptor cells [11].

The optical system of the eye includes the cornea, the lens, and the vitreum. The vitreum fills the interiors of the eye around and behind the lens. The purpose of the parts of the optical system is to refract the incoming light wave to form a sharp, focused, image on the retina. Most of the refraction is caused by the difference in optical density and associated refractory indices between air and the anterior surface of the cornea. In a normally functioning eye the ability to accommodate to nearby objects is provided by the lens. The refracting power of the lens is adjusted by the small ciliary muscle located around the lens. The refraction properties are about 40 D for the cornea and up to 17 D for the lens. The parts of the eye are presented in figure 2.1.

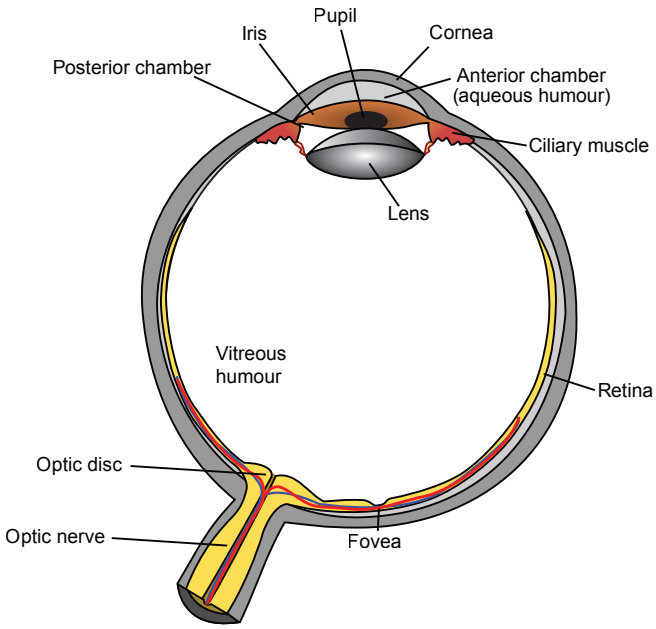


Figure 2.1: The primary features of the eye. Figure modified from [12]

2.1.2 Retina

The first neural component of the visual system is the retina. The retina consists of 60 to 80 cell types: $3-4$ receptor cell types, circa $40-50$ types of interneurons, and circa $15-20$ ganglion cell types plus supporting and connecting tissue [13]. The anatomical arrangement of the retina is reversed as compared to the neuronal signalling chain of the visual system. On the anterior surface of the retina are the ganglion cells, and below the ganglions the interneurons and receptors, respectively.

Photons with energy in the range of visible light cause a photochemical reaction cascade called phototransduction when absorbed into the pigment molecules in the receptor cells of the retina. Phototransduction takes place in the upper cell levels (interneurons) as well, but it supports other functions of the eye, adjustment of the circadian rhythm among other things. Phototransduction causes changes in the chemical consistence of a receptor cell due to the transformation of the pigment molecules and reactions that follow the transformation. Hence voltage-gated channels in the membrane are closed and the sodium-potassium balance is disturbed. This alters electrochemical currents between the inner and the outer segments of the cell finally causing a change in polarization. This complex reaction chain is highly non-linear as well as the other electrochemical phenomena in the latter parts of the visual system [14]. However, the problem can be (partly) linearized considering firing rates, i.e. the frequency of polarizations, instead of potentials [13].

The photoreceptor cells consist of two segments: the inner and the outer segment. In a non-stimulated state, i.e. in darkness, the photoreceptor cell maintains a steady electrical current from the outer segment to the inner segment. When stimulated, the gated ion channels of the cell close and the 'dark current' ceases to hyperpolarize the cell. Thus compared to other neurons, the activation of a photoreceptor cell is reversed.

There are two main classes of receptor cells in the fundus. The first class, called the cones, has three (in pathological cases one, two, or four) subclasses with different absorption properties. The classes are named S, M, and L cone cells, referring to the wavelength of light they absorb. Each class is activated by photons of different energies, initiating the colour coding of the visual system. The second class is called the rods. The rods act in dimmer conditions, providing more sensitivity to light intensity. The density of the receptor cells and the division into the mentioned classes greatly varies depending on the eccentricity, i.e. the distance from the fovea. It also varies considerably between individuals. On average, each human eye contains 4.6 million cones and 92 million rods [15].

In the fovea, the most accurate area of the retina corresponding to the midpoint of the

visual field, there are $199,000\,{}^{\text{cones}}/_{\text{mm}^2}$ on average. There are no rods in the fovea and also the postreceptor cells are bent sidewards to provide the incoming light a direct interaction with the cones. The diameter of the fovea is 0.35 mm (corresponding to 1.25° of visual angle) on average. The density of the cones decreases towards the periphery of the visual field. The rod density is highest around the optic disk, a small elliptical region at the origin of the optic nerve and declines slowly to the far periphery. In the periphery the rods are in majority. On the junction of the optic nerve and retina, there are no receptor cells and it is thus also known as the blind spot. The decline of the receptor cells is anisotropic, thinning out faster in the vertical direction, for both receptor cell types. In the far nasal periphery the cone density even increases slightly. [16]

The changes in the potentials of the receptor cells are detected by the next cell layers, also located in the retina. These layers contain cells called interneurons. The interneurons include bipolar cells, amacrine cells, and horizontal cells. A single cell in these layers gathers and modulates the signals of a number of receptor cells. These groups of receptors differ in spatial configuration and size. From the interneurons the receptive field signals are conducted to the last cell layer in the retina: the retinal ganglion cells (or ganglion cells). The term *receptive field* (RF) refers to the spatial and functional configuration of the receptor cells connected to each ganglion cell. Receptive fields are discussed in detail below.

The ganglion layer concludes the visual processing in the retina. The amount of ganglion cells per eye varies from 0.7 to 1.5 million and on average, the cone cell to ganglion cell ratio is from 2.9 to 7.5. The amount of receptor cells linked to one ganglion cell varies greatly depending on the location in the retina. In the foveal parts, there is a low convergence to prevent information loss, whereas in the peripheral parts, the receptor signals converge from wider regions. In other words, there is a similar thin-out effect as for receptor cells, nevertheless, the effects are not correlated. It is also noteworthy that, due to the interneurons, the organization is not purely convergent: one bipolar cell is connected to more than one ganglion cell. The anisotropy in the distribution of ganglion cells is inherited from the receptors as well. The anisotropy is even more profuse in ganglion cells: there are 3 to 1 ganglion cells in the nasal side compared to the temporal side and almost 2 to 1 in the superior compared to the inferior side. This anisotropy suggests performance differences between directions in visual field. These differences are recorded in a number of studies, e.g. in a work by Intriligator and Cavanagh [17]. [16]

2.1.3 Receptive Fields

The receptive cells are linked to the ganglion cells with a specific pattern. This pattern forms one basic concept of vision modelling: the receptive field (RF). This physiological relation between cell layers was first observed in the retina of the cat [18]. In the retina, the functioning of an RF is described by a centre-surround organization. The organization is depicted in figure 2.2a. RFs have two regions called ON and OFF regions. If light hits the ON region an excitatory signal is fed to the ganglion cell, whereas the light hitting the OFF region causes an inhibitory signal to the ganglion cell. The regions have historically misleadingly been called "excitatory" and "inhibitory" regions [19], however, these regions do not provide simple synaptic excitatory or inhibitory stimulus to post-synaptic cells (PSC). For example a dark stimulus in the OFF region delivers a substantial excitatory stimulus to the PSC. Thus the maximal response is delivered if the spatial distribution of the light stimulus is exactly the shape of the ON region and the intensity of light is sufficient. There are similar receptive field structures to enhance colour vision [20].

However, it must be acknowledged that the RFs in the form described above are only a concept for modelling and the real structure is much more detailed. It has been obvious since the days of discovering the concept of the RF that the isolated RF structure is an oxymoron due to the anatomical structure of the neuronal network in the retina. In early days these effects from outside of the classical RF (cRF) were called "the periphery effects". In more modern publications this effect of horizontal connections in the retina and beyond are included in the concept of the total RF (tRF). [22]

One example of applying the concept of RF excessively straight-forward is the Hermann grid illusion (see figure 2.3).The illusion was once interpreted with RFs of retinal ganglion cells. The explanation claimed that the illusive dots in the intersections were due to milder OFF region stimuli, and hence the intersections are perceived darker. However, the persistent illusion, even if the grid size and shape are varied, proved the explanation wrong. The explanation for the illusory effect requires higher feature integration models. One theory is provided by Schiller and Carvey [23].

2.1.4 Lateral Geniculate Nucleus

The ganglion cell axons conduct the neural responses from the retina deeper in the central nervous system. The axons form the optic nerve, which after a formation named the optic chiasm (or the chiasma), is called the optic tract. In the chiasma the axons of the optic nerve are redistributed to gather the axons concerning the same side of the visual field

(a)

(b)

Figure 2.2: On the left side: receptive fields shapes. On the right side: activations of the cells with corresponding receptive field. The stimuli is depicted below the abscissa. (a) Receptive field of the retinal ganglion cell or LGN neuron. (b) Receptive field of a simple cell in the primary visual cortex. Figure adapted from Carandini et al. [21].

from both eyes to different sides of the central nervous system: the right hemisphere of the visual field to the left and the left hemisphere of the visual field to the right. From the chiasma most of the axons convey the impulses to a subcortical nucleus named the lateral

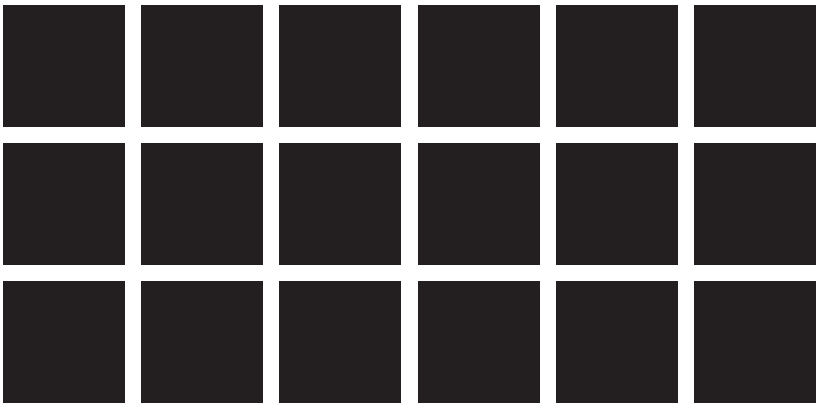


Figure 2.3: Hermann grid illusion.

geniculate nucleus (LGN) of the thalamus.

The LGN further processes the neural presentation of the visual field. It has a similar retinotopic structure and similar receptive fields as the earlier parts [24]. The colour processing in the LGN differs from the one done by the retina [25]. The optic nerve consists of three pathways conveying information to different layers of the LGN. The pathways are called the M-pathway, the P-pathway and the koniocellular pathway. The M-pathway projects to the magnocellular part of the LGN. These cells provide fast response and high contrast sensitivity but do not convey small details. The P-pathway projects to the parvocellular part of the LGN and provides information about spatial details but has slower responses [26]. The colour coding is also isolated to the detailed P-pathway [20]. The function of the koniocellular pathway is poorly known.

This simplified description of the visual system may induce an image of simple cell relay that modifies the signal on its way towards the final 'grandmother-cell'. It is clear that the visual system is much more complex. The visual system is a sophisticated network of cells. Lateral connections in each stage and feedback connections between the cortical levels and even to the LGN (showed in the feline visual system) form the fundamental basis of the system besides the basic feedforward connections between adjacent layers. In the LGN the feedback loops from the cortical areas affect what signals are transferred when to the cortical areas via the feedforward connections [27].

A small part of the ganglion cell axons do not project to the LGN but to the pulvinar, a

nucleus of the thalamus, via the superior colliculus in the midbrain. These axons form the extra geniculo-cortical pathway. It is very fast and thus important for guiding attention to salient stimuli [28].

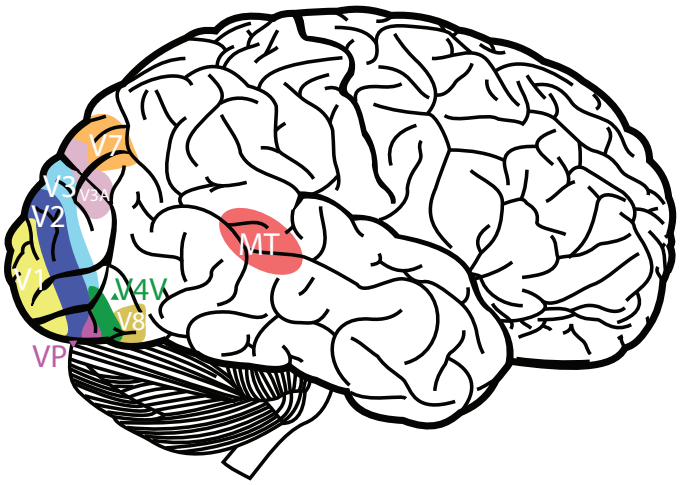
2.1.5 Cortical Regions of Vision

The synapses from the LGN finally lead the modulated signal to the primary visual cortex, i.e. V1. It is located around the calcarine sulcus in the occipital lobe. Beyond the V1 the signal is spread widely around the cortical areas, where visual processing is occupying large parts, possibly even one fourth, of the neocortex [29]. From V1 the signal is conducted to V2 and forward to V4, MT, to mention few. Some areas are depicted in 2.4. Despite extensive research the exact nature and functioning of most of the areas remain unclear.

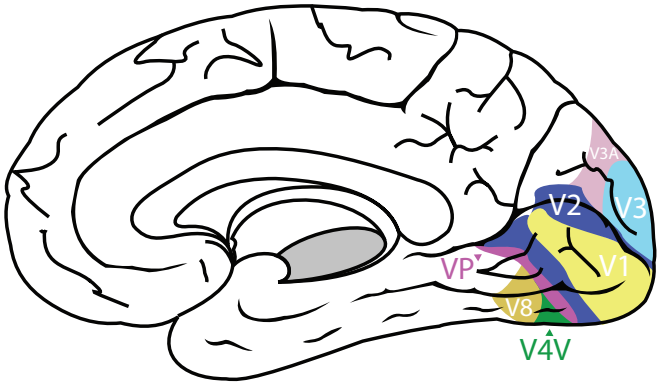
After V2 the visual system is divided into two large pathways called the dorsal and ventral streams. The function of these streams, sometimes also referred to as the "where" and "what" streams, was presented by Mishkin et al. [31] and was demonstrated in human experiments by Haxby et al. [32]. The theory divides the processing according to behavioural significance: the ventral stream deals with object recognition and leads close to areas involved to memory functioning, as the dorsal stream processes motion and spatial location aspects of perception, e.g. human motor activities [33] and thus leads towards the sensory and motor cortices.

The coding methods of the early cortical regions are relatively well understood and studies have even been made to determine the image shown to subjects solely by assessment of the cortical activation patterns or the activation in LGN caused by the image [34, 35]. One essential feature found in study of early cortical coding is retinotopy, i.e. the preservation of spatial relations of the environment [36].

The revelation of the V1 code was started by the influential work of Hubel and Wiesel [19]. Their work led to a consensus that the V1 processes visual information in terms of local contrasts, e.g. the edges in a perceived image are strongly represented. The V1 is anatomically arranged in neuronal columns, called hypercolumns. The hypercolumn consist of six distinct layers. In these layers one single location in the visual field is analyzed in all orientations and many spatial scales of contrast variation. These columns ultimately combine the common information gathered by both eyes. Anatomically, the primary visual cortex is bounded by a striping, visible to a bare eye that consists of the myelinated axons from the lateral geniculate body terminating in layer 4 of the neuronal columns. It is, thus, occasionally referred to as the 'striate cortex'. It has been estimated that the striate cortex contains about 140 million neurons. [37]



(a)



(b)

Figure 2.4: Examples of brain areas that process visual information. The upper part (a) of the figure presents the lateral right hemisphere and the lower part (b) the medial view of the right hemisphere. Most of the brain matter involved in early cortical processing of vision is in the calcarine sulcus. Image adapted from Tootell et al. [30].

The V1 cells have receptive fields as well. The receptive fields, however, differ considerably from the ones earlier in the processing chain. These RFs are elongated and hence selective to stimulus orientation. The existence of these RFs was first recorded by the pioneers Hubel and Wiesel in the feline cortex [19]. They found two distinct cell types that respond to stimuli with a small differences: the first class, i.e. the simple cells, are selective to orientation, magnitude, direction, size, and phase, whereas the other class, the complex cells have similar responses to all phases. The RFs and the activation are depicted in figure 2.2b for the simple cell. The activation of a complex cell is illustrated in figure 2.5

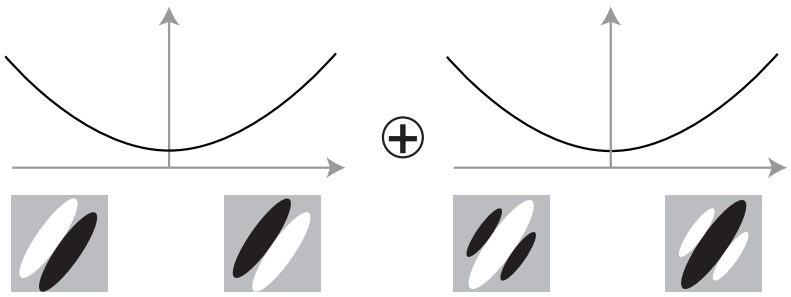


Figure 2.5: Actviation pattern for complex cell. Complex cell can be thought of as a sum of simple cells with quadrature phase-shift. Figure adapted from Carandini et al. [21].

Here, as well, the concept of RF is a very simplified tool for describing the functioning of the system. The RF explains only the effect of the feedforward synaptic links from the lower cortices, LGN, and retina. The lateral connections from the cortex per se, or feedback connections from higher parts of the system are not modelled by cRF. However the signals outside the cRFs are common and have a major influence on the way we perceive objects and scenes [38].

One interesting phenomenon stemming from the lateral and feedback connections of the visual system is illusory contour detection. Illusory contours are very powerful illusions appearing in areas that should or very likely could contain contours. Examples are given in figure 2.6. One can notice the apparent contours in figure 2.6. This raises the question of the meaning of being illusory. There has been a debate about the lowest level in the

visual system in which the contours exist. In the past the illusory activity was thought to be located at V2 and above. More recently evidence of "illusory activity" in as early as the striatal cortex, i.e. the first cortical area of visual processes, has been found by Lee and Nguyen [1]. The experiment was done with rhesus monkeys but there is similar evidence from human fMRI experiments [39].

Figure 2.6: Examples of the illusory contours. On the left is the famous Kaniza's triangle. On the right, viewer can find an s-shaped line between the parts of the shape.

Even though the illusory contours are not 'an illusion' even in the striatal cortex, they affect apparently in as high as the infero-temporal cortex (IT) [40]. This is often regarded as one of the final destinations of the dorsal pathway. Lesions in these areas cease the perception of illusory contours, suggesting that illusory contours are fundamentally related to object recognition.

The neural origin of visual phenomena of higher level is generally poorly understood. However, fMRI studies have proposed that e.g. the crowding effect takes place largely in a bottom-up manner in the visual system somewhere in the processing chain from V1 to V3. In these studies the brain functioning shows differences between crowded and non-crowded situations in V4 [41].

The descriptions in this section may give an excessively precise and simplistic image of the visual system. However, there is a vast amount of variability between individuals in

the quantitative and possibly even in the qualitative facts listed above. This complicates the modelling of this already very complex system.

2.2 Modelling Human Vision

Modelling of human vision can be divided into two categories: neuronal and behavioural. The approach using a neuronal basis is further divided into two modelling approaches with distinct aspects. These are descriptive and normative modelling. Descriptive modelling explains properties of the visual system using mathematical methods, whereas normative modelling utilizes mathematics to reason why the system works in a certain manner. Because of the aim of this work, the focus here is on descriptive modelling.

The behavioural modelling category is historically very significant, but the focus of the current research is on neuronal networks and neural signalling. However, neuronal functioning in high-level processes in human vision is not generally understood. Models for these phenomena are, thus, more or less process descriptions, i.e. behavioural models. These models appear in psychological and psychophysical contexts. The current models with neuronal basis are located to lower levels of vision than the models with weaker neuronal nexus describe, as for example the object recognition phenomena.

2.2.1 Descriptive models

Most models employ the concept of the linear RF. In fact the model for the RF is the beginning of computational modelling of vision. Current models have many modifications of the model of the RF presented in section 2.1. Mathematically the RF is described as a two-dimensional filter. This signal is modulated with the RF filter to model the response of the corresponding physiological system. Linear filters are generally utilized in signal processing using convolution.

$$g(x,y) = f(x,y) * h(x,y) = \sum_{\psi} \sum_{\chi} f(\chi,\psi) \cdot h(x-\chi, y-\psi) \tag{2.1}$$

Equation 2.1 shows the denotation and mathematical description of the discrete convolution in two-dimensional space, the χ and ψ are the samples in the signal, f is the signal, and h is the filter. The inverted filter is multiplied with the signal in every sample location of the signal.

The outputs of the filters are generally modified to resemble the firing rate of a cell. This means half-way rectifying (negative firing is generally not possible) and saturating the response in high stimulus intensities. This can be done mathematically or logically.

$$f(\alpha) = \min(d, \max(0, \alpha - c)) \tag{2.2}$$

$$f(\alpha) = d\frac{\alpha^2}{a+\alpha^2} \tag{2.3}$$

In equations 2.2 and 2.3 the d is the maximum response, c is the threshold, i.e. the minimum stimulus to elicit a response and α is the magnitude of the stimulus. In equation 2.3 a is another constant that defines the saturation behaviour of the function. Often, if the purpose is not to model the behaviour of a single simple neuron the half-way rectification can be substituted with the 'energy model' [42]. Physiologically the energy model describes the functioning of the V1 complex cells. Mathematically it is a complex function that is used to model the signal of a complex cell or a number of simple cells by taking the square of the modulus of the function.

2.2.2 Modelling of the Retina

Many of the models of the pre-cortical parts of the visual system discuss the neural properties of the retina and especially the output of the retinal cell complex. These models describe the non-linear behaviour in synaptic signalling [43, 44]. However, earlier parts of the neural visual system can be modelled with simplified linear filters as well [45]. The linear systems consider the retina as a holistic structure rather than a cell network, still counting in a large part of the properties of the retina.

The cell system of the retina can be described adequately for purposes of defining the information loss with difference of Gaussians (DoG), which are Gaussian filter systems. The parameters for the DoG filters must change with position when the retinal inhomogeneity is modelled. Additional components are used in some models to count in e.g. neural high-pass properties of the retinal system. This is a coarse simplification of the complex network of the retinal, receptor-interneuron-ganglion cell structure.

The DoG filters, sometimes called Mexican hat filters, are composed of a sum of two Gaussian envelopes with differing deviation and polarity.

$$\mathrm{DoG} = G_{\sigma_1} - G_{\sigma_2} = \frac{1}{\sqrt{2\pi}}\left(\frac{a}{\sigma_1}e^{-(x^2+y^2)/2\sigma_1^2} - \frac{b}{\sigma_2}e^{-(x^2+y^2)/2\sigma_2^2}\right) \tag{2.4}$$

In the equation the G_{σ_1} and G_{σ_2} are simple two dimensional Gaussian envelopes with deviations of σ_1 and σ_2 respectively. Now, setting the deviations to alter over space, and convolving this function with the signal, the main property of retina, i.e. location dependent band-pass attenuation, is modelled. functional features of the RFs are modelled.

$$\sigma_1 = \alpha_{\sigma_1} \cdot f(x,y) \tag{2.5}$$
$$\sigma_2 = \alpha_{\sigma_2} \cdot f(x,y) \tag{2.6}$$
$$f(x,y) = \sqrt{(x-x_0)^2 + (y-y_0)^2} \tag{2.7}$$

Where x_0 and y_0 are the coordinates of the centre of the fovea.

The DoG filter mimics the RF structure of a retinal ganglion cell. The RFs of retinal ganglion cells are larger and the density of the receptors is lower in the peripheral parts of the visual field. Thus the DoG filters are tuned to gather information from larger areas in parts that present the peripheral visual field.

To facilitate the computations it is possible to make coarse approximations, called two-dimensional Gaussian kernels. The Gaussian kernels are discrete matrices that represent Gaussian envelopes. Alternative, more sophisticated method, to perform the actions described above is to transform the signal and the filter function to the frequency space using the Fourier transformation. The Fourier transform decomposes the function in a linear sum of harmonic functions (sinusoids) [46]. Any function can be divided into such components. Moreover, there are computationally cheap methods to perform the Fourier transformation, Fast Fourier transform (FFT) for example.

$$F(x,y) = \sum\sum e^{-i(\omega_x x - \omega_y y)} f(x,y) \tag{2.8}$$

Equation 2.8 gives the discrete Fourier transform ($F(x,y)$) of an arbitrary discrete two-dimensional signal, i.e. of an image. The spatial frequencies in the x and y-directions are denoted by ω_x and ω_y, respectively.

In two-dimensional space the result of the Fourier transform is a topographical frequency map. The map represents all spatial frequencies that the signal contains, sorted, in polar coordinates. The direction from the centre point of the map represents the direction of the spatial oscillation and the distance from the centre point represents the frequency of the oscillation.

Now, taking piecewise inverse transforms of narrow rings of the map results an approximation for different spatial frequencies in the image. To model the retina the approximations

are weighted depending on the frequency and the distance from the point representing the fovea.

Some of the parameters for these filter systems are estimated indirectly. The sizes of the retinal RFs are suggested to correlate with the cortical representation, that each RF occupies equal size in the V1. This correlation is presented in the concept of the cortical magnification factor [47]. The cortical magnification resembles the relation between sizes in the retina, i.e. in the environment, and in the V1 cortex. Because of lesser convergence, the central areas of the retina have a bigger representation in the cortex. Depending on the source the cortical magnification factor has dozens of values. There has been a debate regarding the magnitude of the factor, but no consensus has been reached [48–50]. The definition of the magnitude of the cortical magnification is, hence, not unambiguous. However, $\sigma(x,y)$ in the DoG is inversely proportional to this factor.

The other parameters of the system: the spatially dependent amplitude of the system (a and b in equation 2.4) and the neural high-pass modulation are easier to define. Determining the values a and b as a function of the location in the visual field is done using a psychophysical test. In the test the spatial dependence of the contrast threshold [51] was determined with psychophysical experiment. The human vision model may also include a neural high-pass component in addition to the spatially altering low-pass component. The proposed high-pass component is suggested to have a value that is related to the spatial frequency measured in cycles per degree [45]. This relation is linear with a slope value of 1.

The system concerning these properties is not the most comprehensive model of the retina, but it describes rather well the information loss due to the properties of the retina. The form of the proceeded information (the retinal code) is often irrelevant albeit the retinal processing is suggested to start the attention modulation, according to Thorpe et al. [52]. In their model the temporal pattern of the spikes, i.e. the action potentials of the neurons, contain information.

However, due to the poor current knowledge of the processes in phenomena such as crowding, detailed neural models are impossible to utilize. The models developed for the quantification of crowding use either the signal modulated by the retinal DoG model or the plain image signal as their input.

Taking into account the LGN is rare. This is because its functions are quite poorly understood at the neural level. In general, the research on the LGN is focused on revealing the functional role of the nucleus. There is, however, little modelling on its neural functioning. For example there is a model in which the LGN divides information by statistical features of an image [53].

2.2.3 Models of Cortical Actions

The function family that is used to model the responses of V1 cells is called Gabor wavelets or Gabor functions. The idea of Gabor functions was developed in mid-twentieth century [54] and it was introduced to vision modelling in the same year by Marcelja and Daugman [55, 56]. The Gabor wavelet models all the critical properties of the V1 cell. It is tuned, i.e. sensitive, to orientation, frequency phase and size. Thus the complex shape of the RF of the V1 cells is reduced to a small number of parameters.

$$h(x,y) = w(x,y)\cos(\omega_x x + \omega_y y) \tag{2.9}$$

Equation 2.9 represents a product of a weight (w) and a harmonic function (cos), i.e. a two-dimensional Gaussian function and a sinusoidal grating. The dimensions are noted by x and y and the spatial frequency by ω. The envelope spatially localizes the harmonic function. Without the weight the harmonic function would apply to the whole signal and would thus be an ordinary Fourier component (explained in detail in section 2.1.2) of the signal. Thereby the Gabor transformation is generally referred to as localized spectral analysis. In complex form the Gabor filter can be written using the Euler formula:

$$h_{\sigma_x,\sigma_y,\omega,\phi}(x,y) = e^{-\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)} e^{i\omega x} \tag{2.10}$$

Now the response a system is a function of the size of the spatial localization (σ_x and σ_y), and the spatial frequency of the oscillation (ω). The orientation (α) can be modelled by introducing new coordinate systems (x_{new},y_{new}):

$$x = x_{\text{new}}\cos\alpha + y_{\text{new}}\sin\alpha \tag{2.11}$$

$$y = y_{\text{new}}\cos\alpha - x_{\text{new}}\sin\alpha \tag{2.12}$$

These equations may be substituted into equation 2.10.

There are many modifications of the original Gabor wavelet. The modifications are generally referred to as wavelet transforms. A model named Berkeley Wavelet Transform (BWT) [57] is specifically designed to represent the properties of V1 functioning as a local contrast detector. BWT preserves an approximation of all the features of the original Gabor functions but are computationally inexpensive.

The BWT model consist of eight mother wavelets, four odd and four even, with the orientations 0°, 45°, 90°, and 135°. The mother wavelets can be produced mathematically

with a unit step function, and the equations are found in the paper by Willmore et al. [57]. The mother wavelets are illustrated in figure 2.7. The mother wavelets are scaled and translated to produce daughter wavelets. The wavelets have spatial frequencies and orientation bandwidths that are comparable to biological values.

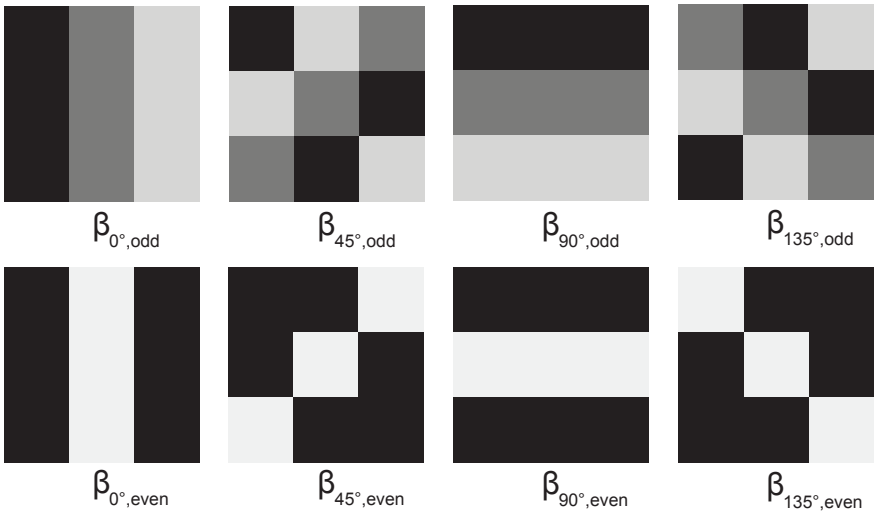


Figure 2.7: Mother wavelets of the BWT basis. The upper row shows the odd and the lower the even wavelets. The figure is adapted from, Willmore et al. [57]

$$\beta_{\theta,\phi}^{m,n,s} = \frac{1}{s^2}\beta_{\theta,\phi}(3^s(x-m), 3^s(y-n)) \tag{2.13}$$

In equation 2.13 $\beta_{\theta,\phi}^{m,n,s}$ is the daughter wavelet and $\beta_{\theta,\phi}$ the mother wavelet with orientation ϕ and parity θ that determines whether the wavelet is odd or even. The indices m and n determine the position and s denotes the scale. As can be seen the wavelets are scaled by the power of 3. This restricts the signal, i.e. the image, size to multiples of the wavelet size (3^s) and it causes the size of RFs to expand heavily in large s values.

The base of the BWT is self-inverting and reconstruction of the image is thus accomplished using

$$I(x,y) = \sum_{\theta,\phi} \sum_{m,n,s} w_{m,n,s}^{\theta,\phi} \beta_{\theta,\phi}(3^s(x-m), 3^s(y-n)) \tag{2.14}$$

Where $w_{m,n,s}^{\theta,\phi}$ are the coefficients of each wavelet in the particular signal. The wavelets have zero mean and thus the DC term (β_0) must be added to the inverse transform.

2.2.4 Normative Models

To explain the features of the wavelets and other filters described above normative modelling is needed. Normative modelling explains a structure as being optimized to perform a certain function. Thus it presumes the system is evolved to be ecologically valid. It describes the functioning of a system by coding the input signal in 'quasi-optimal' terms. In vision modelling this code is often referred to as the sensory code.

Normative modelling gives an explanation as to the way the visual system analyses the signals in terms of localized spatial frequencies. The statistical properties of the common stimuli have favoured certain coding methods. The common stimuli for a visual system are, of course, natural images, i.e. the scenes we encounter in everyday life. These properties are detected using mathematical methodology, e.g. employing frequency analysis on an image.

The frequency analysis, for example, is a very simple method but it reveals interesting facts [58]. Firstly it shows that the natural images have a lot more vertical and horizontal frequencies compared to the intermediate ones. Secondly, by analysing the frequency representation of an image, it is possible to place images in a space describing the location of the image in terms of it being outdoor-indoor and wide scene - close-up..

However, the statistical properties analyzed are usually of higher degree. These properties of natural stimuli have revealed other existing structures in natural images. The natural images are 'sparse distributed' and coded in in the sensory system according the same 'sparsity'. The 'sparse distribution' refers to high kurtosis value of the signals [59]. The high kurtosis in natural images stems from the universality of plain surfaces and rather sharp contours in borders of these surfaces. Research on these properties of natural images has revealed the purpose of the receptive field structure in work of Olshausen and Field [60].

2.2.5 Clutter Models

Beyond the descriptive and normative models are the models for psychophysically or psychologically exhibited phenomena in vision. This research yields important knowledge about the processes in visual perception and explains of the formation of perception. Moreover, the models contribute to the current development of artificial visual systems in more concrete ways than the models of the early visual system.

The basis of the perceptual models is on image processing. The perceptual models explain which properties in the pixel or dot groups in the image form objects and how the objects are processed further. Visual clutter is an important concept for the assessment of image properties by its perceptual workload.

Rosenholtz et al. [8] define visual clutter as a state in which excess items, or their representation or organization, lead to a degradation of performance in some (visual) task. Apparently there is no unambiguous definition about features, attributes, and factors that determine the amount of clutter, and about the mutual weighting of them. The models of visual clutter, however, attempt to define techniques that could assess the clutter in an image.

There are several models and techniques to assess clutter in an arbitrary image. The techniques include for example spatial information density assessment [61] and assessing how different image properties relate to subjective complexity estimate [62]. The models have evolved greatly from the historical measures of perceptual workload, such as set size or line count [63]. However, the methods for assessing the tendency of an image to cause perceptive workload are still quite naïve and ineffective.

One method in particular is of interest of this study. It is the clutter assessment toolkit by Rosenholtz et al. [8]. The methods of Rosenholtz et al. tries to derive and implement measures of visual clutter. The measures are Feature Congestion, Sub-band Entropy and Edge Density. All these techniques generate a map of the image to indicate the cluttered regions. The map is two-dimensional as the image.

Feature Congestion assesses spatial variations of colour, intensity and the edge orientation in an image. Sub-band Entropy estimates the information content of the image. These measures and techniques have very little to do with human physiology or neurology. There are, however, remarkable results on the correlation of Feature Congestion with subjective complexity, i.e. clutter rankings [64].

Despite the fact that the models lack a neural nexus they generate a feasible approximation of the result of the human perception. Certainly, modelling the construction of perception

is not the purpose of the clutter models; these models try to predict the result of the whole perceptual system. In the clutter models all possible deteriorating factors converge to one final measure of visual clutter.

The models embed various physiologically demonstrated processes such as crowding, masking, and deterioration of object recognition. The corrupting effects in visual search and other actions related to task are included as well.

The phenomena presented above have been more or less excessively studied separately. Knowledge on these subjects varies greatly, for example object recognition is a widely theorized and extensive subject, whereas the masking effect is well understood. Crowding effect that is the focus of this study, is discussed in detail below.

2.2.6 Saliency Models

A property in an image that attracts attention is saliency. Saliency modelling examines the attention transfer (attention route) over visual perceptions. Saliency models have a lot in common with the clutter models. Saliency can be thought of as an opposite of clutter. The saliency models use low-order features of the image to create a two-dimensional map to address the points in the images that pop out the most. In addition, a part of the saliency models include other parameters to be able to predict the probable route of attention on the image [65].

Albeit saliency and clutter maps share many features they have one fundamental distinction. Clutter maps assess the content globally while ranks in saliency maps are dependent on the environment. The saliency is always assessed locally, in relation to the flanking locations. The saliency of a region cannot, thus, be sorted globally without the context.

Apparently models that take low-level image features as their only input are incapable of modelling the top-down functioning of attention. However, few of the models define themselves as 'biologically inspired'. Biologically inspired refers to the extraction methods of the features from the image. Biologically inspired models treat the image similar to the early visual system and thus their extracted features are thought to be more accountable than features that are obtained with different filters. The other models in general are, however, quite efficient, as well.

Nevertheless, the features of saliency models are simple and low-level. For example the famous model by Itti et al. [66] takes colour, intensity, and orientation centre-surround differences as an input. Another model by Barth et al. [67] that defines the pre-attentive texture discrimination model is based on statistical studies [68] of eye movements on a

natural image. It uses higher-order statistical figures that define e.g. corners and junctions in the image.

The goal of assessing how attention is modulated in an arbitrary image is certainly a task of very complex mathematical modelling that comprises e.g. the top-down influences, i.e. the intentional aspects, the visual search protocols, effects of learning, etc. Despite the effect of top-down influences a study by Elazary and Itti [69] shows that the attention and regions-of-interest on natural images can be defined using saliency models that are based on low-level features. In the study the subjects were asked to mark the most interesting location in an arbitrary image. The study included a huge number of responses and the statistics of responses showed significant congruence with the saliency model of Itti et al. [66]. Although the study was only a statistical comparison, there are hallmarks of bottom-up saliency very early (V1) in the visual system [70].

One very apparent problem for attention guiding is eye movements. The attention can be guided to different parts of the visual field without eye movements (covert attention), but normally the eyes are directed towards the region of interest (overt attention). The saliency maps represent covert attention, i.e. the static situation in which the most salient regions attract attention, and thus gaze turn there. In other words, to model overt attention the eye location must be presented as a sequence of a covert saliency maps.

The pioneers Itti and Koch have modelled visual search as a sequence of saliency maps with an additional parameter called the inhibition of return [65]. Visual search is a studied subject as well. Studies have shown that human performance in visual search is very near to an ideal observer [71, 72].

Visual search is related to clutter as well as to saliency maps. The concept of clutter models must include the visual search strategies in the future to extend the usage of the concept of clutter to situations including eye movements. In current models, the clutter is related to a search task vice versa, the amount of clutter is assessed using the search performance as measure [64, 73].

2.2.7 Crowding Effect and Modelling

The models presented above are efficient in modelling the outcome of the visual system. The system and its mechanisms are, thus, not elucidated. For modelling the intermediate phenomena inside the system different methods and protocols are needed. The research on the actions inside the visual system is essential for improving the ‘biologically inspired’ measures and catching the details in the data processing chain.

The most crucial 'intermediate' phenomenon in the visual system is probably crowding. Its locus in the neural system is defined, but its causes are still unclear. Studies suggest that it defines the limits for face recognition [74], reading rate [3], and other recognition tasks [75]. Thus it is a ubiquitous effect for visual processing that requires object recognition.

Crowding is generally defined as the deleterious influence of flanking objects for identifying a target object. Bouma discovered in the 1970's that the spatial extent of the crowding effect is approximately linearly dependent on the eccentricity of the target [4]. The effect itself was discovered earlier.

Because crowding impairs the identification in clutter and it is the limit of e.g. reading rate, one could argue that it is the root cause of clutter. However, some essential factors of the holistic concept of clutter need other explanations. For example the pop out effect of a flickering target is not restricted by crowding.

In general, crowding is distinct from ordinary masking by its relation to detection. The crowding effect does not influence detection tasks. This means that after adding the crowding effect to perception every object will still remain visible but identification is deteriorated. Thus, crowding does not reduce the apparent contrast but rather the feature integration exceeds the detail resolution required for object identification.

There are a number of simple feature integration and other feature processing, suggested to be related to crowding, e.g. segmentation, contour integration, feature bonding, etc. May and Hess [76] has proposed that crowding is due to the overlapping of the "association fields". This "association field" would help the contour integration, similar to the illusory contours. All this applies well to the determined locus of the crowding phenomenon, V4: The anisotropy of its receptive fields is in line with the experimental anisotropy of the crowding effect. Physiological studies show that it is the area that combines the different features described above [77].

The area in which the flanking objects affect the recognition of the target object is called critical spacing. Bouma quantified it as a round area around the target object with a size of $0.5\phi^\circ$, ϕ° being the eccentricity of the target. Toet and Levi [78] have later specified that the area is elongated in the radial direction, being an ellipse described by the equation

$$s = \frac{s_0 + b\varphi}{\sqrt{1 + (\varepsilon^2 - 1)\frac{\varphi_V^2}{\varphi^2}}} \tag{2.15}$$

The modified version of equation 2.15 is provided by Pelli et al. [3]:

$$s = s_0 + \frac{b\varphi}{\sqrt{1 + (\varepsilon^2 - 1)\frac{\varphi_V^2}{\varphi^2}}} \tag{2.16}$$

This modification improves the fit to experimental data. In equations 2.15 and 2.16 s_0 refers to the zero component of critical spacing, i.e. critical spacing in 0° eccentricity. It is circa 0.06° of visual angle. The constant b is a factor for the length of the minor axis of the ellipse and is circa $0.25 - 0.3$. The constant ε is the ellipticity, Toet and Levi estimated it as equal to $2 - 3$. The radial eccentricity and the vertical component of eccentricity, are φ and φ_V, respectively.

The estimate by Toet and Levi [78] for the crowding area is about twice as long in the radial direction than in the tangential direction, according to the model. With this model, the distance between the target and the flankers is suggested to be the centre-to-centre distance between elements [2].

An interesting phenomenon, which is not included in the model, is the anisotropy of the crowding effect. The crowding effect is shown to depend on the direction in which the target is located [17]. The effect is moderately strong, the advantage for the lower visual field compared to the upper field being 17% with tangential flankers and up to 50% with radial flankers. These anisotropies suggest that the crowding is strongly correlated to the underlying physiology, i.e. the cortical representation of the visual field.

Crowding is inward-outward anisotropic as well. Flankers inward from the target disrupt identification more than flankers outward from the target. Motter and Simoni [79] suggest a simple solution; in cortical representation 'outward flankers' are closer to the target than flankers inwards from the target. Another problem that arises from the cortical representation is that at the level at which crowding occurs the remaining retinotopy is minor and the features dominate the representation. Thus the features most probably affect to the critical spacing [2, 17, 75].

Other important aspects that are left out of the model are the higher-level effects on crowding. Higher-level effects are for example the effect of the pattern formed by the target and the flankers. The complexity of the pattern is shown to affect the strength of the crowding effect. A number of property differences affect to crowding as well, e.g. shape, size and orientation [80]. The combination effect, i.e. the larger scale features the target and flankers form, is, as well, shown to influence the crowding effect [5]

Another higher-level aspect is the influence of attention. The influence of attention is controversial and under an intensive debate [7]. There is no consensus as to whether

attention is related to the magnitude of the crowding effect.

Considering the diversity of details in the crowding effect, the model described above is only a tentative approximation for the problem. Crowding is, however, an essential phenomenon in vision. It has real life relevance because of its limiting nature in object recognition. It is the restrictive "window" through which one perceives the environment. It limits information intake and is the limiting factor for visual search as well [81].

CHAPTER

THREE

METHODS

This chapter describes the computational and experimental methods used in this thesis. The computational methods were derived from the literature on information content, complexity and clutter in images. The existing models were further processed to incorporate the special features of the crowding effect.

The computations were performed using software developed for this thesis. A psychophysical experiment was conducted to evaluate the validity of the computational methods. The experiment was also to assess the crowding effect in natural images.

3.1 The Computations

The software was developed to model the information loss in the early visual system and to establish a model for the estimation of the crowding effect in an arbitrary image. The image is used as an input and an estimate of crowding is returned. The estimate contains both total crowding within the assessed area and a map containing crowding values for each location. Furthermore, the software allows the illustration of the filtering properties in the early visual system in the image. The information loss in the retina and local contrast detecting properties of early cortical areas, can be depicted using the illustration feature.

The program was written in the Python [82] programming language using the SciPy library [83]. The SciPy library contains tools necessary for scientific computations. In this section

the mathematical models constructed in this thesis are described in detail. For general descriptions of the software architecture refer to appendix A.

3.1.1 Pre-processing

The image file fed to the program was read in as an array. This array contains the greyscale values for each pixel in the image. The images were converted to greyscale by summation of the colour channels. The equation is as follows:

$$I(x,y) = 0.23I_{\mathrm{R}}(x,y) + 0.71I_{\mathrm{G}}(x,y) + 0.06I_{\mathrm{B}}(x,y) \tag{3.1}$$

The numerical constants in equation 3.1 are the measured and calculated ratios and $I_c(x,y), c \in \mathrm{R,G,B}$, is the display sub-channel for a one of the colours red, green or blue. The result was calculated by measuring the luminance of each pure colour channel with full intensity on the display used in the experiment. The measurements were made with a Minolta Chroma Meter CS-100.

3.1.2 Retinal Model

The retina function of the program mimics the filtering by retinal receptive fields (RF), including 'neural high-pass filtering' [45]. The filtering is performed in the spatial frequency domain. The concept is described in section 2.2.2. An example of the results of the retinal filtering is given in figure 3.1. Figure 3.1 illustrates only low-pass features of the retina function. The high-pass component is not illustrated due to the distrubting effect (it adds greyness to the image) but it is added to the computations of the crowding effect.

3.1.3 Sub-band Entropy

The image is first transformed to the frequency domain using an FFT algorithm. The frequency presentation is divided to narrow bands with high-degree Butterworth filters (20). Ideal filters were not used to prevent the ringing effect. The first part is extracted with a low-pass filter and the higher frequency bands using band-pass filters. The limits of the bands were selected to be fractions of octaves. Hence, the higher frequencies were analyzed in wider bands. The number of bands was set to 20. The bands approximate the receptive field distribution in the retina.

The extracted band was transformed to the spatial domain using an inverse FFT algorithm. The band-pass filtered image was attenuated with a modulator to model the retinal RF structure. The equations for the modulator is,

$$A(x,y,f_{\text{band}}) = e^{-Md(x,y)\widetilde{f_{\text{band}}}\alpha} \quad (3.2)$$

$$d(x,y) = \sqrt{(x-x_{\text{fovea}})^2+(y-y_{\text{fovea}})^2} \quad (3.3)$$

Equation 3.2 consists of the factor (α) which is achieved by fitting a linear model to the measurements by Campbell et al. [51] on contrast sensitivity around the visual space, and the approximative cortical magnification factor (M). The constants are multiplied with the median frequency of the band ($\widetilde{f_{\text{band}}}$) and the distance from the fovea ($d(x,y)$). The contrast sensitivity attenuation ($A(x,y)$) is calculated by weighting the product with an exponential function. All the x and y values refer to indices of the pixels of the image, unless mentioned otherwise. The attenuation modulator is subsequently multiplied by the neural high-pass component and the modulated signal is thus

$$S(x,y,f_{\text{band}}) = H(f_{\text{band}})\cdot I(x,y,f_{\text{band}})\cdot A(x,y,f_{\text{band}}) \quad (3.4)$$

$$= \widetilde{f_{\text{band}}} I(x,y,f_{\text{band}}) e^{-Md(x,y)\widetilde{f_{\text{band}}}\alpha} \quad (3.5)$$

In equation 3.4 $H(f_{\text{band}})$ is the high-pass component of the retinal transfer function and $I(x,y,f_{\text{band}})$ is the band-pass filtered input image.

All the bands are treated in a similar manner and the results are summed. The sum is scaled to have the original contrast in the foveal location.

3.1.4 V1 Model

The V1 model was implemented using the Berkeley wavelet transform (BWT). The BWT basis is introduced in section 2.2.3. The V1 model consist of an orthonormal set of wavelets representing the V1 RFs. The result represents the contrast detection functionality of the primary visual cortex. It was implemented due to requirements of the feature congestion measure presented below.

A pyramid representation of the input image is first built. The anti-aliasing for the lower resolution levels was made by averaging. The levels of the pyramid were convolved

with the wavelet basis to represent different scales of the RFs. The product was squared according to the 'energy model' (see section 2.2.1 and the paper by Hyvärinen et al. [42]). The BWT basis requires that the steps in the image pyramid reduce the size by the power of 3.

$$\mathbf{I}_{\text{filtered}} = \sum_{s}\sum_{w} (\text{BWT}(w) * \mathbf{I}_{\text{pyramid}}(s))^2 B_{\text{spline}}(s) \tag{3.6}$$

Equation 3.6 represents the procedure of applying the BWT. Each level (s) of the pyramid image ($\mathbf{I}$) is convolved with all wavelets (w). Spline interpolation was used to rescale the images in the pyramid. After rescaling all the convolutions were summed to obtain all the BWT coefficients.

3.1.5 Critical Spacing

The area in which the cluttering properties of an image were assessed is defined in the critical spacing model. In the model the crowding effect is caused only by the image area spatially nearby the target. The model is explained in section 2.2.7.

The area influencing crowding is defined as an ellipse around the target. The location of the target is defined manually and the critical spacing represents the "spot-light" area of covert attention around the target. The ellipse is elongated in the radial direction and has a minor axis to major axis ratio of 2 to 1. The length of the major axis is linearly dependent on the distance between the target and the fovea. The area can, thus, be defined by the following inequalities:

$$x < x_{\text{target}} + d_{\text{target}}\cos(\tau)\cos(\varphi_{\text{target}}) - 0.5 d_{\text{target}}\sin(\tau)\sin(\varphi_{\text{target}}) \tag{3.7}$$

$$y < y_{\text{target}} + d_{\text{target}}\sin(\tau)\cos(\varphi_{\text{target}}) - 0.5 d_{\text{target}}\cos(\tau)\sin(\varphi_{\text{target}}) \tag{3.8}$$

$$-\pi < \tau < \pi \tag{3.9}$$

In equations 3.7 and 3.8 x_{target} and y_{target} are the target coordinates and d_{target} is the target distance from the fovea. The parameter φ_{target} is the radial direction of the target from the fovea. The length of the major axis is the distance between the target and the fovea. The area, thus, extends halfways to the eccentricity of the target ($0.5\phi_{\text{target}}$). This is the spatial extent of crowding in the models presented below. The critical spacing procedure and retinal filtering are illustrated in figure 3.2

3.1.6 Sub-band Entropy

The first clutter assessment model implemented was the sub-band entropy measure used by Rosenholtz et al. [8]. The sub-band entropy computation utilizes the wavelet transform of the V1 model (see section 3.1.4) and the Shannon entropy. Equation for each wavelet can, thus, be written,

$$E_{w,s} = \sum_i -p_i \log p_i \tag{3.10}$$

In the critical spacing area each scale (s in eq. 3.10) and each wavelet (w in eq. 3.10) coefficients are binned to construct an approximative histogram of the contrast coefficient distribution (p in eq. 3.10). The entropies ($E_{w,s}$ in eq. 3.10) are summed over the wavelets and scales to obtain the clutter measure used in the prediction of crowding.

3.1.7 Feature Congestion Measure

A more sophisticated clutter tool by Rosenholtz et al. [8] utilized in this thesis was the feature congestion model. The feature congestion assesses the variance and covariance of certain image properties regionally. The properties used were the contour orientations and the contrast. These measures were applied to the critical spacing area only, similar to the sub-band entropy model.

The properties were computed using algorithms adapted from Rosenholtz et al. [8] The contrast was computed using the image pyramid (adapted from the V1 model in section 3.1.4) and a DoG filter. The obtained contrast was assessed regionally by filtering the image using an average filter. Approximate kernels for Gaussians were utilized to present the following filters.

$$G_{\text{filter}}(x,y) = \frac{1}{\sigma_{\text{I}}} e^{-(x^2+y^2)/2\sigma_{\text{I}}^2} - \frac{1}{\sigma_{\text{O}}} e^{-(x^2+y^2)/2\sigma_{\text{O}}^2} \tag{3.11}$$

$$G_{\text{average}}(x,y) = \frac{1}{\sigma_{\text{B}}} e^{-(x^2+y^2)/2\sigma_{\text{B}}^2} \tag{3.12}$$

The deviations for equations 3.11 and 3.12 were adapted from the estimates of Rosenholtz et al. [8]; $\sigma_I = 1.42$, $\sigma_O = 2.28$, and $\sigma_B = 6.0$. Finally the contrast variation of the regionally averaged results was computed as

$$\text{var}_{\text{contrast}} = E[\mathbf{I}^2] - E[\mathbf{I}]^2 \tag{3.13}$$

The contrast variation ($\text{var}_{\text{contrast}}$ in eq. 3.13) was computed for each scale in the image pyramid and the results were combined by taking the maximum value in each location over the scales.

The contour orientation was obtained in a similar manner. The wavelet model was utilized for contour calculations at each scale of the image pyramid. Again, the wavelet coefficients (i.e. the convolution results) were averaged using a Gaussian kernel ($\sigma = 3.5$) for each wavelet and scale. The orientation clutter was computed by taking the difference of the orthogonal wavelet coefficients, i.e. the vertical wavelet was subtracted from the horizontal wavelet and the 45° diagonal from the 135° diagonal wavelet.

The differences were then regionally averaged similarly to the contrast case ($\sigma = 4.6$). The covariance matrix of the two averaged differences was formed and the determinant taken, giving the clutter estimate:

$$\begin{aligned}
\text{var}_{\text{cardinal}} &= E[w^2_{s,0^\circ-90^\circ}] - E[w_{s,0^\circ-90^\circ}]^2 && (3.14)\\
\text{var}_{\text{diagonal}} &= E[w^2_{s,45^\circ-135^\circ}] - E[w_{s,45^\circ-135^\circ}]^2 && (3.15)\\
\text{cov} &= E[w_{s,0^\circ-90^\circ} \cdot w_{s,45^\circ-135^\circ}] - E[w_{s,0^\circ-90^\circ}] \cdot E[w_{s,45^\circ-135^\circ}] && (3.16)\\
\mathbf{V}^{cntr}_s &= \begin{vmatrix} \text{var}_{\text{cardinal}} & \text{cov} \\ \text{cov} & \text{var}_{\text{diagonal}} \end{vmatrix} && (3.17)
\end{aligned}$$

In equations 3.14, 3.15, and 3.16 w is the wavelet coefficient set averaged with G_{average} of scale s and denoted orientation. The fourth root of the determinant values are combined over the scales in a similar manner to the contrast case, i.e. by taking the maximum value in each location over all scales. The combined information contains the orientation clutter map.

The orientation and contrast clutter maps are combined into a feature congestion clutter map as follows:

$$\mathbf{V}_{FC} = \frac{\mathbf{V}_{\text{contrast}}}{0.0660} + \frac{\mathbf{V}_{\text{orientation}}}{0.0269} \tag{3.18}$$

The coefficients in equation 3.18 do not sum to one because of the missing colour measure. The final measure was reduced to a scalar by calculating the mean of the final map. The clutter maps are very similar to figure 3.2b.

3.1.8 Contrast Energy

In addition to the clutter estimates, crowding was assessed using a simple contrast energy measure in the critical spacing area. The contrast energy is a reasonably good complexity measure in restricted situations. It was implemented here because of the relatively small area of assessment, i.e. the critical spacing. The (Weber) contrast energy is computed by the equation

$$C_{\text{energy}} = \sum_y \sum_x \left(\frac{I(x,y) - \bar{I}}{\bar{I}} \right)^2 \tag{3.19}$$

In equation 3.19 $I(x,y)$ is the value of the pixel and $\bar{I}$ is the mean value of the pixels in the area.

3.1.9 Complexity Measure

The last tool for crowding estimation is a complexity measure for simple stimulus images adapted from Näsänen et al. [84]. The complexity measure evaluates the spatial deviation of the contrast energy within the critical spacing area and weights the result with the median frequency of the area.

The median frequency of the area was obtained by transforming the image into the frequency domain and then constructing a histogram of the image frequencies. Before transformation to frequency domain the mean was subtracted to prevent the edges of the critical area from causing distortions to the frequency representation.

The spatial spread for the contrast energy was computed by first evaluating the centre of the contrast energy by weighting the coordinates with the contrast values over the area:

$$p_x = \frac{\sum_x x \left(\frac{I(x,y)-\bar{I}}{\bar{I}} \right)^2}{\sum_x \left(\frac{I(x,y)-\bar{I}}{\bar{I}} \right)^2} \tag{3.20}$$

$$p_y = \frac{\sum_y y \left(\frac{I(x,y)-\bar{I}}{\bar{I}} \right)^2}{\sum_y \left(\frac{I(x,y)-\bar{I}}{\bar{I}} \right)^2} \tag{3.21}$$

In equation 3.20 $\bar{I}$, is the mean value over the image and $I(x,y)$ the value of the pixel with coordinates (x,y). The contrast of the image is then weighted with the distance from

the centre of gravity. The contrast energy of the weighted contrast map is subsequently computed and it is finally weighted with the mean frequency of the area. Thus the equation is

$$c_{\text{complexity}} = \sum_y \sum_x d_p(x,y) \left(\frac{I(x,y) - \bar{I}}{\bar{I}} \right)^2 \cdot \widetilde{f} \tag{3.22}$$

In the equation $d_p(x,y)$ is the distance from the centre of contrast energy and $\widetilde{f}$ is the median frequency of the critical area.

3.2 The Experiment

The psychophysical experiment measured contrast sensitivity for the letter object presented on natural image backgrounds in peripheral vision. The aim of the experiment was to induce the crowding effect and to validate the computational results. A novel technique, in this context, was used to present a letter stimulus with a varying contrast on natural image background.

3.2.1 Subjects

Four subjects (three males, age 25-58) participated in the experiment. All the subjects were employees of the Finnish Institute of Occupational Health at the time of the experiment. All subjects had normal or corrected to normal vision. The contrast sensitivity was tested and all the subjects exceeded 1.80 in the Pelli-Robson contrast sensitivity test performed in 6500 °K white-light so that the luminance at the test chart was $90\,\mathrm{cd/m^2}$.

3.2.2 Apparatus

The stimuli was generated using a Hewlett-Packard dc7100 CMT desktop computer with an ATI Radeon HD 2600 Pro graphics card and was presented on an Eizo FlexScan S2100 LCD-monitor. The declared pixel amount was 1600×1200 and the vertical and horizontal scanning frequencies 59 Hz and 61 Hz respectively. The declared pixel size was 0.270×0.270 mm. To evaluate the gamma correction needed, the luminance of the display was measured from an axis perpendicular to the image plane and 5° down-10°

right (same as the target location) at twelve different RGB values. The measurements were made with a Minolta Chroma Meter CS-100. A linear model was fitted to the measured data in logarithmic scale to obtain the gamma correction. The gamma correction value was 1.97. The computed and measured luminance values are presented in figure 3.3.

The software used for stimulus presentation was Psycho 1.0, a psychophysical stimulus presentation platform by Risto Näsänen. It uses a staircase algorithm to define psychophysical thresholds.

3.2.3 Stimuli

The software utilized in the experiment presents a 1024×1024 pixel (13.8° of visual angle to the horizontal and vertical directions (distance was set to 57 cm)) greyscale natural image on an interstimulus mid-grey (RGB value of 127 i.e. $128.5\,\mathrm{cd/m^2}$) background. A fixation cross was constantly visible at the centre of the display. The target letter was presented on the background image in the top-left visual field when fixating the fixation cross. The centre of the target was at an eccentricity 9.3° of visual angle (7.9° and 4.9° in horizontal and vertical directions, respectively). The targets were capital letters of the Courier New typeface.

There were seven background image conditions. Six conditions contained a natural outdoor image in various urban environments. The last image contained a plain mid-grey image (RGB value of 127). The images were presented in greyscale. Five letter sizes were used. The height of the letters varied from 1.19° to 3.56° of visual angle.

The stimuli were presented for 500 ms. The stimulus consisted of a background image and target letter chosen at random. The contrast of the target letter changed according to the specifications of the staircase algorithm. The specifications were set as follows: Three right answers were required to decrease the contrast of the letter with a factor of 1.41 while one wrong answer increased the contrast with factor of 1.41. A threshold was defined as the mean of six turning points (reversals).

A special protocol was needed in order to present the varying contrast of a natural image background. The contrast value on the letter had to be independent of the background grey level, and thus, the luminance. The protocol was developed by Risto Näsänen, the author of the stimulus software. In the protocol, the contrast was varied in accordance with the transparency of the target. The transparency levels for the letter were obtained using the transformation

$$T_* = c_{\text{Weber}} T_0 - c_{\text{Weber}} + 1 \tag{3.23}$$

$$c_{\text{Weber}} = \frac{L_{\max} - L_{\min}}{L_{\max}} \tag{3.24}$$

In equation 3.23, T_0 is the letter contrast if the letter is on a plain white background. $L_{\min}$ is the letter greyscale value and $L_{\max}$ the greyscale value of the background. The transformation T_* preserves the background signal and the contrast level of the target letter is set to $1 - c_{\text{Weber}}$. Instead of summing the greyscale values of the letter and the background, a product is taken to achieve all possible contrasts for the transformed letter:

$$S(x,y) = B(x,y)T_*(x,y) \tag{3.25}$$

In equation 3.25, $B(x,y)$ is the background. Presented on the signal $S(x,y)$, the semi-transparent letter now has the same contrast as the corresponding greyscale letter on a plain (equal to unity) background. Examples are given in figure 3.4

3.2.4 Procedure

The experiment took place in a dark room with the only source of illumination being the display. The subject was seated in front of the display, eyes in line with the fixation cross in the centre of the display. The distance of the forehead from the display was set to 57 cm and was held constant with a headrest. Using a distance of 57 cm, one centimetre on the display approximately corresponds to one degree of visual angle when the angle is small. An illustration of the experimental set-up is given in figure 3.5

The task was to identify the letter displayed. The gaze was constantly held on the fixation cross. The target was thus perceived in the peripheral vision. No feedback was given regarding the correctness of the response. There was no protocol for observing the gaze but a repeat button could be pressed if doubts about the eye movements emerged, after which the last trial was repeated using a new target.

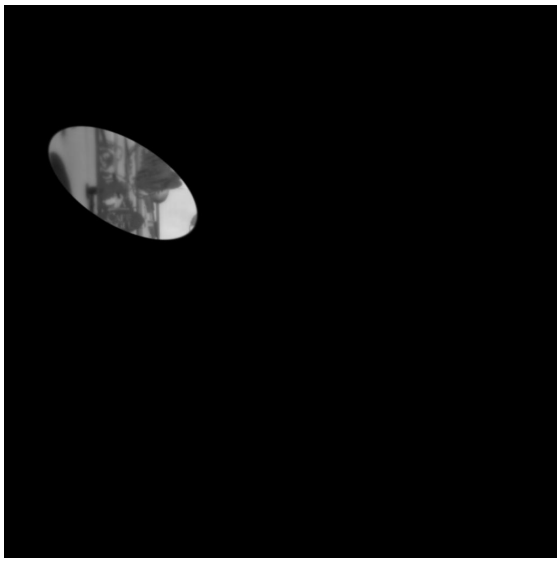
The experiment was repeated three times on different days for each subject. One experiment lasted for approximately 40 min. The subjects were encouraged to take breaks at appropriate occasions, i.e. when a threshold was determined. In one experiment each combination for all of the letter sizes and the background images was tested. The total number of the conditions was hence 35.

(a)

(b)

Figure 3.1: Illustration of the filtering procedure that mimics the low-pass features of the retina. The images are scaled and, thus, the filtering appears steep. (a) An example of an unfiltered background image (Kamppi) (b) A filtered background image.

(a)



(b)

Figure 3.2: Illustration of the filtering and critical spacing procedures. (a) An example of an unfiltered background image (Tuomiokirkko) (b) A critical spacing area of the background image. The image is filtered using the retinal model.

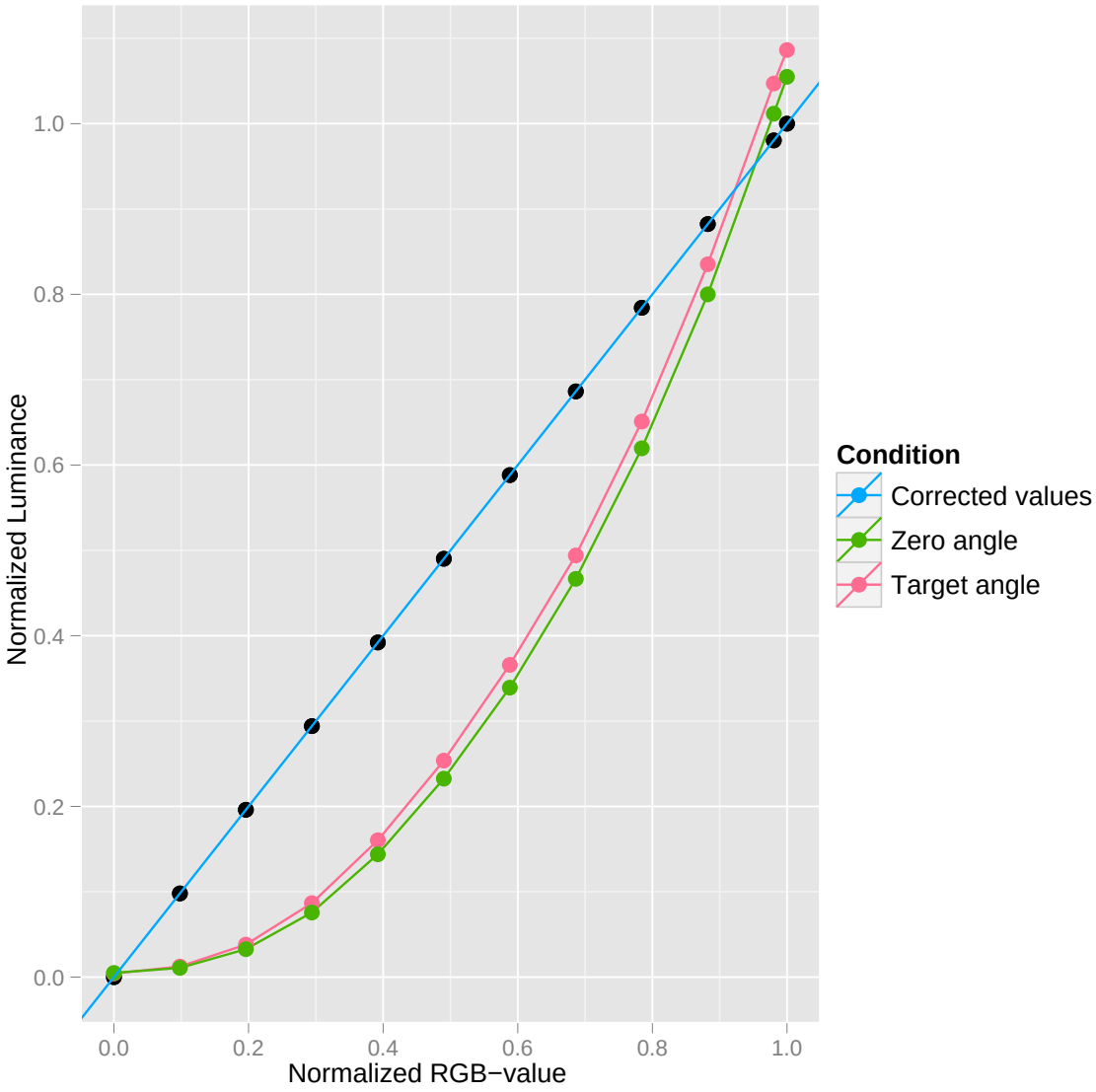


Figure 3.3: Measured luminance for the display without gamma correction and the calculated corrections made. Zero angle luminance values were measured in front of the display and at the target angle values used in the experiment.

(a)

(b)

Figure 3.4: Illustration of the experimental situation: Background image (Esplanadi) and stimulus with varying contrast. (The stimulus letters were not encircled in the experiment.) (a) A letter with full contrast and (b) a letter with a contrast of one third.

Figure 3.5: Illustration of the experimental set-up. The background image is in question is the Westminster.

CHAPTER

FOUR

RESULTS

This chapter presents the computational and experimental results. The main question was whether there is a simple way to quantitatively predict the crowding effect using computational methods. It was found that the crowding effect in natural images is clear. The computational prediction of the magnitude of the crowding effect in natural images is promising, as well.

The first part of the chapter presents the results of the psychophysical experiment and the latter part reports the relation between the computational models and the experimental results.

4.1 Experimental results

The experimental results show significant congruence across the participants. The results are roughly independent of the target stimulus size and hence, only a function of the background image. Figures 4.1a and 4.1b show the similarity of the results between the size categories. The three categories not illustrated gave similar results.

There is a minor cut-off effect in the smallest size category (letter height 1.19° of visual angle). However, all the results show a sigmoidal shape at contrast thresholds arranged in ascending order. The variance between the subjects is negligible. The effect of the background images is, thus, similar for each subject.

4.2 Results of the Models

Relations between the experimental results and the computational results were compared in different filtering settings. The different settings were compared using correlation tests and qualitative differences are presented in charts. The charts present the relations between the experimental results and computations categorized according to the background images.

Although the computations implemented to assess the intensity of the crowding effect (see chapter 3.1) were simplistic and based on image properties, the results achieved were promising. Table 4.1 shows correlation test results from Spearman's rank correlation test (Spearman's ρ). The null hypothesis was that the correlation between the experimental results and the predictive computations is positive.

The correlations show a clear relation between the psychophysical results and three out of four of the computed measures. Spearman's ρ also shows that with more complex measures, the best results are achieved with the unfiltered images. The contrast energy computations work best when using the full early visual system filter model. However, despite the high correlation, some measures show controversial effects in the qualitative assessment. The complexity measure fails to show a relation with the psychophysical experiment and the computed values.

Table 4.2 shows the linear dependence between the experimental results and the computations (assessed with Pearson's correlation). The strongest correlation was found between the simplest models (contrast energy and sub-band entropy) and the experimental results. The correlations were computed over all experimental data. The standard deviations

Table 4.1: Correlation calculated between the experimental results and the used computational methods using the Spearman's ρ. The correlations are categorized according to the applied visual system filters (RF filtering represents the low pass functioning of the retina). Only the results for a letter size of 2.96° of visual angle are presented. Other stimulus size categories gave similar results. (* means $p < 0.001$ and ** stands for not significant.)

Computational tool	**No filtering**		**RF filtering**		**RF and neural high-pass filtering**	
Contrast Energy	0.72	*	0.82	*	0.78	*
Sub-band Entropy	0.86	*	0.71	*	0.69	*
Feature Congestion	0.75	*	0.66	*	0.71	*
Complexity Measure	-0.74	**	0.04	**	-0.72	**

(sd), thus, depict the deviations between the size categories. The deviations appear to be moderately small.

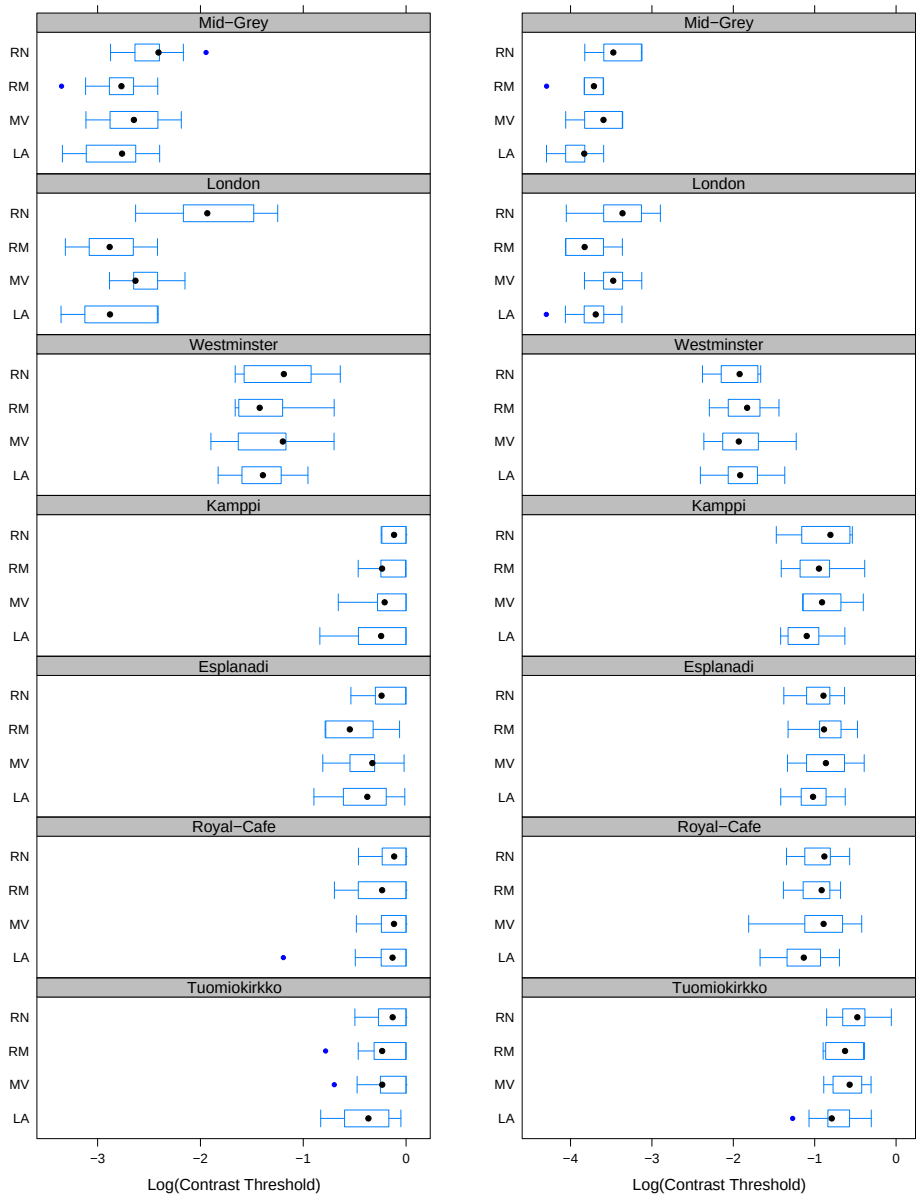
Despite the very similar correlation coefficients in the three measures (contrast energy, sub-band entropy, and feature congestion) there are qualitative differences between them. Figures 4.2 and 4.3 illustrate the differences between the contrast energy measure and the feature congestion measure in a situation with no pre-filtering in condition with letter height 2.96° of visual angle. Other size categories showed similar results.

The filtering settings greatly affect the results as well. Figures 4.4 and 4.5 show the experimental results as a function of contrast energy and feature congestion with full early visual system model. If these results are compared to the ones in figures 4.2 and 4.3 one can notice the effect of pre-filtering. The pre-filtering enhances the upward trend in the contrast energy model and amplify the margin between the backgound images in the feature congestion model.

Despite the highest correlation coefficient (see table 4.1), the sub-band entropy measure bears little relevance to the crowding effect intensity if the image is pre-processed with the visual system filters. This is exemplified by the difference of figures 4.6 and 4.7. Figure 4.6 shows no upward trend .

Table 4.2: Pearson product-moment correlation coefficients between the experimental results and used computational method. The correlations are categorized according to the used visual system filters (RF filtering represents the low pass functioning of the retina). The correlations are computed over all size categories. The standard deviation is denoted with sd.

Computational tool	**No filtering**		**RF filtering**		**RF and neural high-pass filtering**	
Contrast Energy	0.92	(sd 0.01)	0.87	(sd 0.01)	0.90	(sd 0.01)
Sub-band Entropy	0.92	(sd 0.02)	0.92	(sd 0.02)	0.92	(sd 0.02)
Feature Congestion	0.91	(sd 0.01)	0.84	(sd 0.01)	0.87	(sd 0.03)
Complexity Measure	-0.69	(sd 0.01)	0.22	(sd 0.05)	-0.73	(sd 0.01)



(a) **(b)**

Figure 4.1: Results of the psychophysical experiment. Contrast thresholds for letter sizes of (a) 1.19° and (b) 2.37° of visual angle. Black bullets (●) mark the median of the turning points in the threshold algorithm. Blue bullets (●) denote the outliers. The outliers are defined as reversals farther than 1.5 times the range between the second and fourth quartiles. The ends of the boxes denote the second and fourth quartile and the whiskers denote the minimum and maximum of the data.

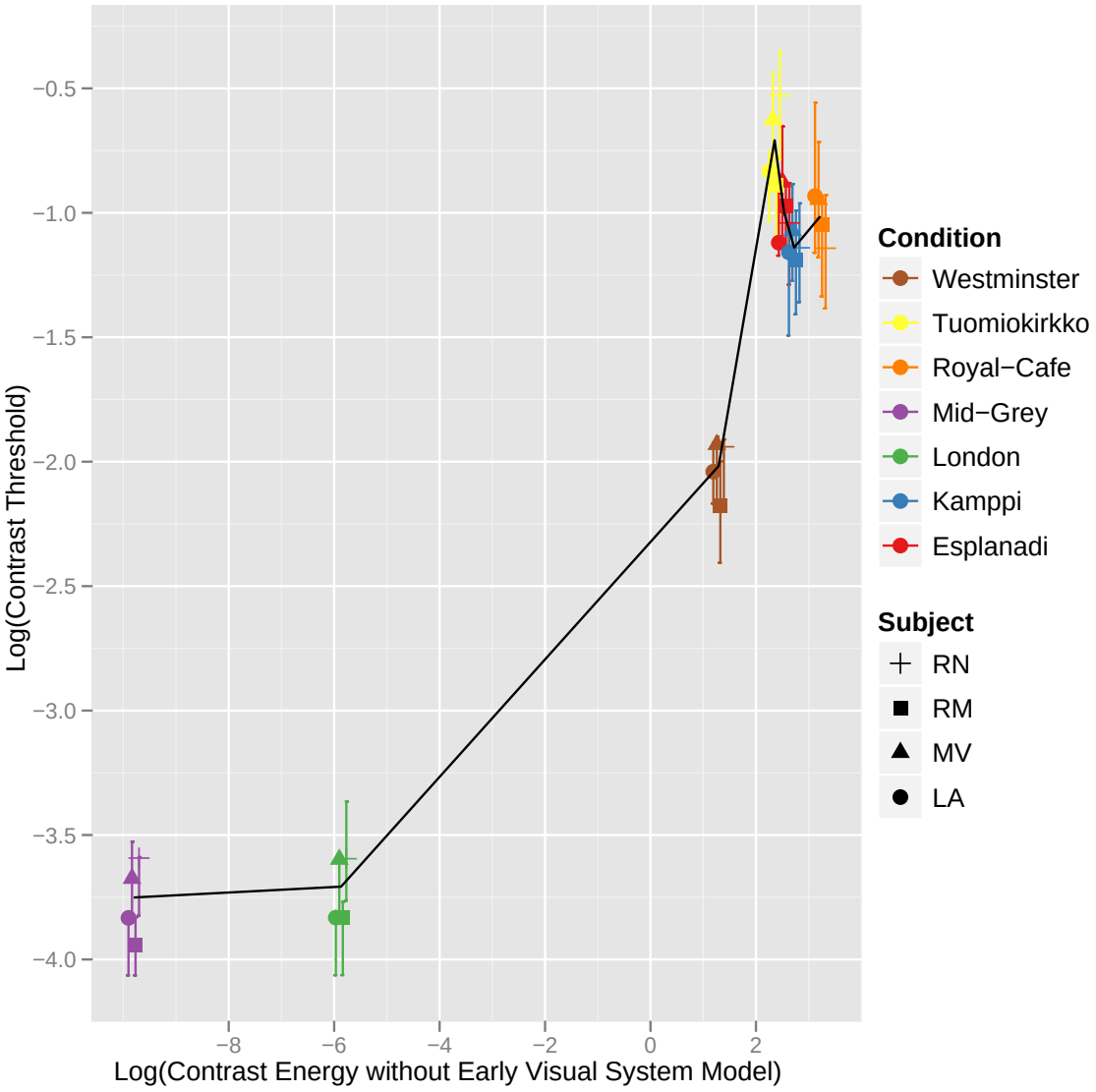


Figure 4.2: The logarithm of the experimental contrast thresholds as a function of the logarithm of the local contrast energy in the critical spacing area. The image was not filtered before computation of the local contrast energy. Only results of a letter size of 2.96° of visual angle are presented. The results were similar for the other size categories. Subjects are jittered along the x-axis to prevent overlap.

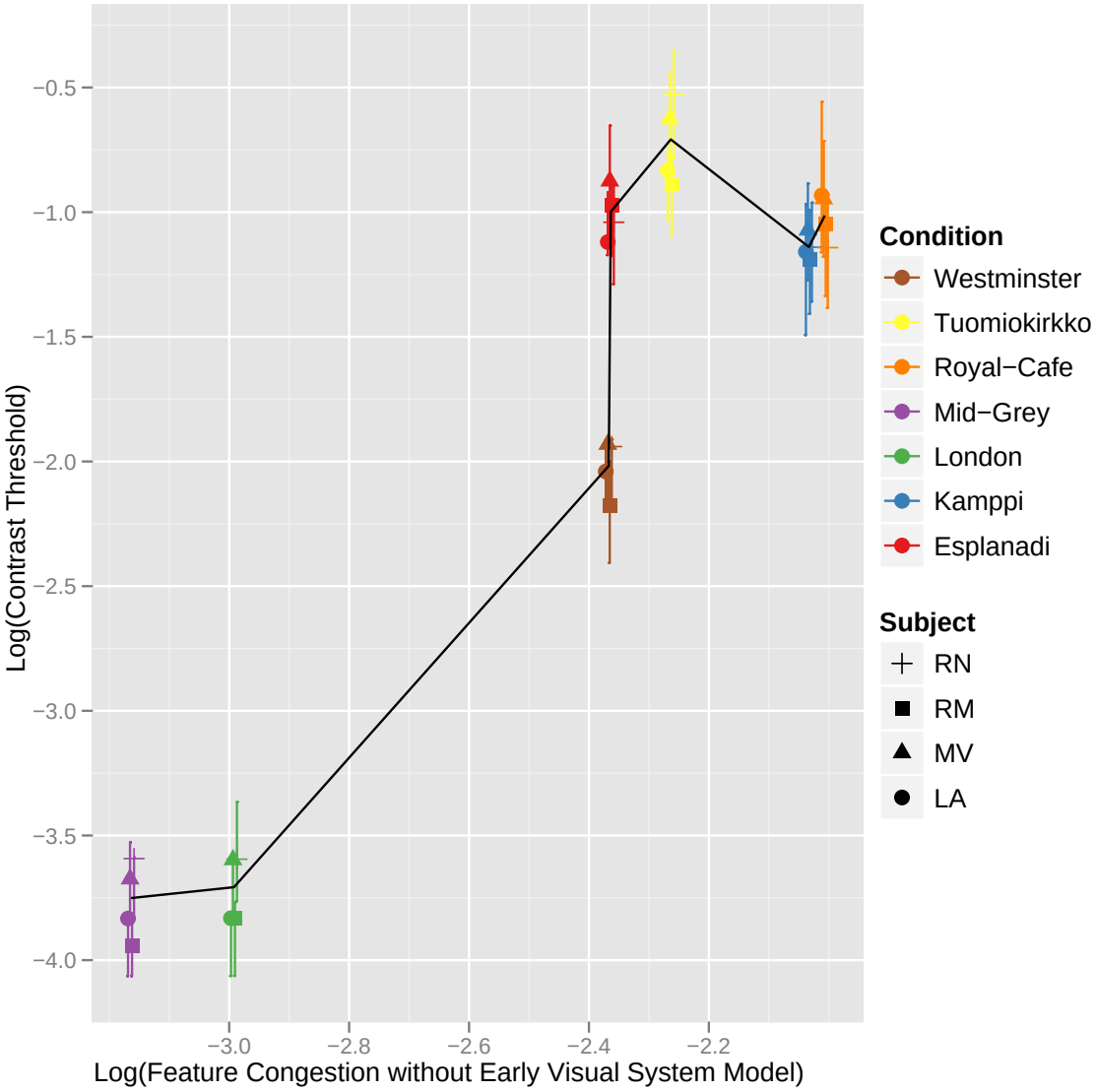


Figure 4.3: The logarithm of the experimental contrast thresholds as a function of the logarithm of the local feature congestion measure in the critical spacing area. The image was not filtered before computation of the feature congestion value. Only results of a letter size of 2.96° of visual angle are presented. The results were similar for the other size categories. Subjects are jittered along the x-axis to prevent overlap.

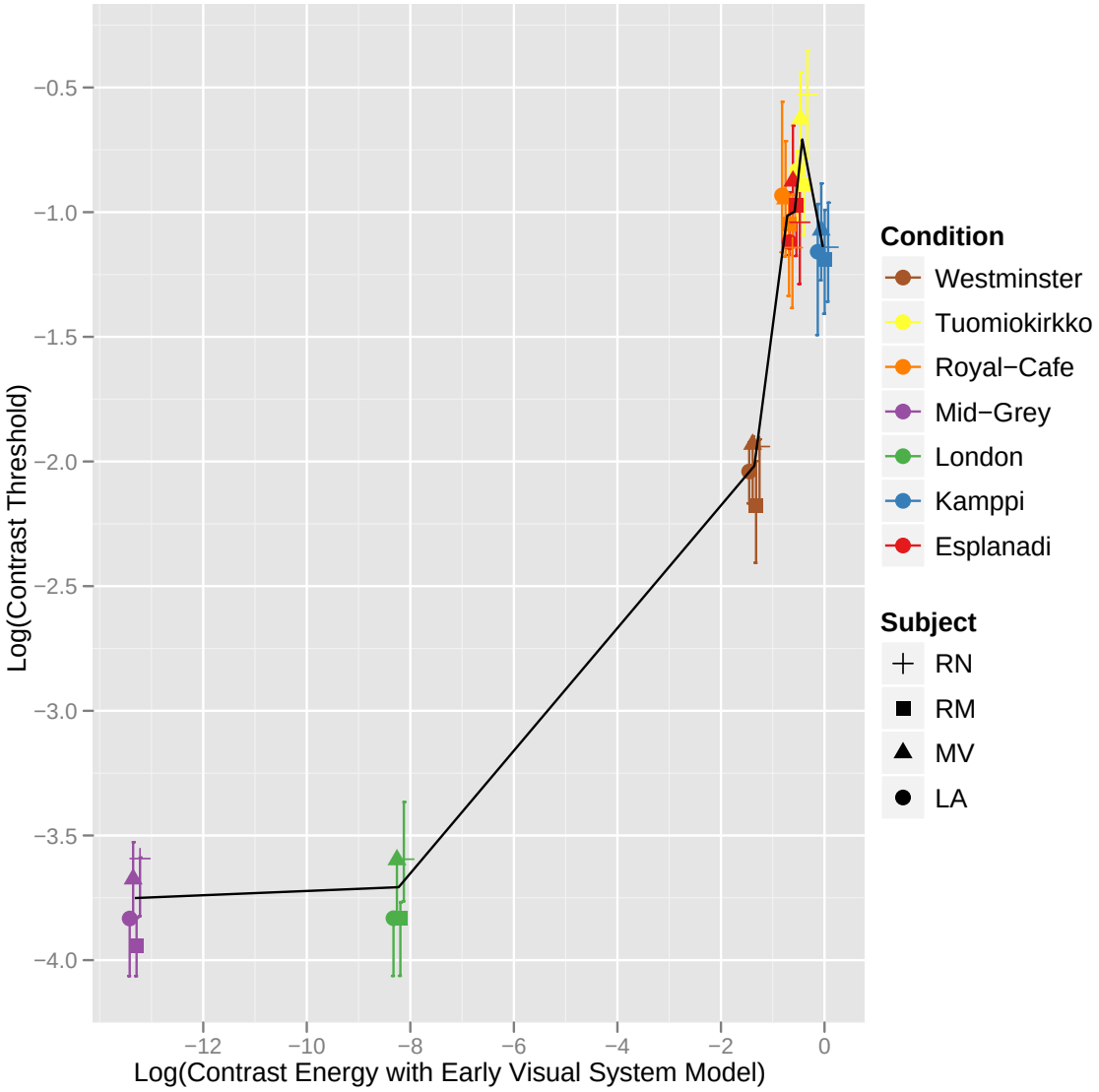


Figure 4.4: The logarithm of the experimental contrast thresholds as a function of the logarithm of the local contrast energy in the critical spacing area. The image was filtered with the early visual system model before computation of the local contrast energy. Only results of a letter size of 2.96° of visual angle are presented. The results were similar for the other size categories. Subjects are jittered along the x-axis to prevent overlap.

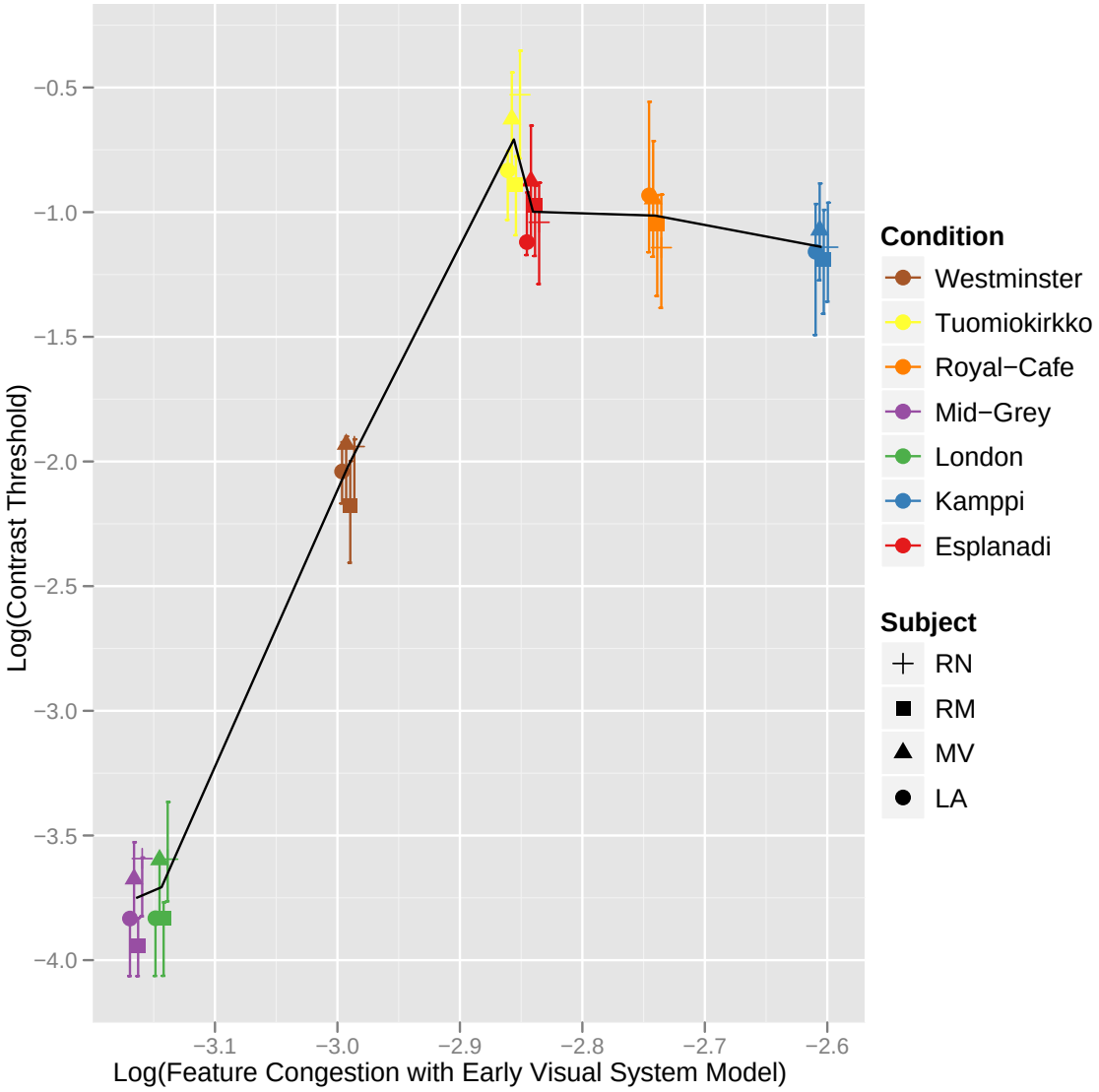


Figure 4.5: The logarithm of the experimental contrast thresholds as a function of the logarithm of the local feature congestion measure in the critical spacing area. The image was filtered with the full early visual system model before computation of the feature congestion value. Only results of a letter size of 2.96° of visual angle are presented. The results were similar for the other size categories. Subjects are jittered along the x-axis to prevent overlap.

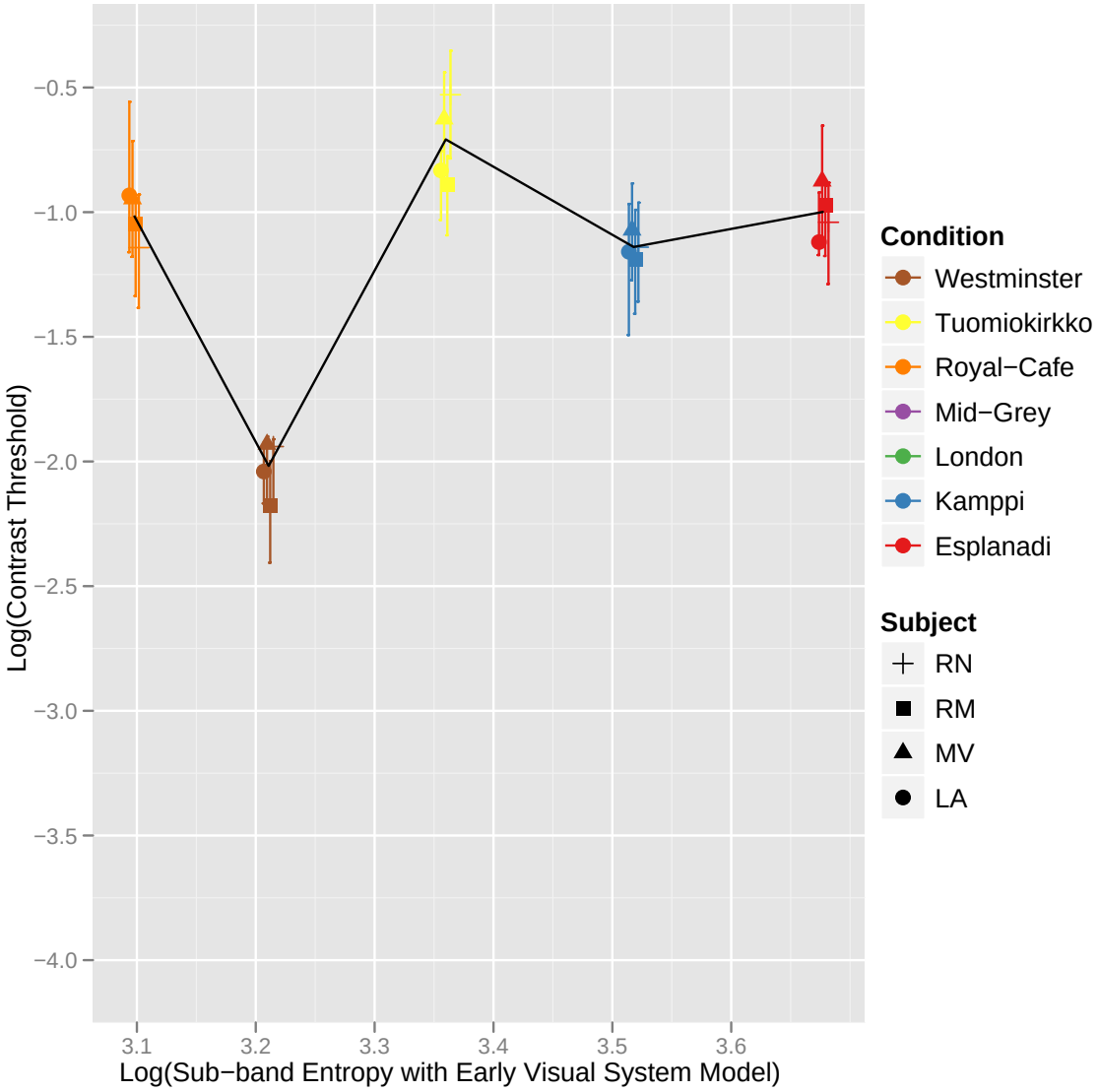


Figure 4.6: The logarithm of the experimental contrast thresholds as a function of the logarithm of the local sub-band entropy sum in the critical spacing area. The image was not filtered before computation of the sub-band entropy sum. Only results of a letter size of 2.96° of visual angle are presented. The results were similar for the other size categories. Subjects are jittered along the x-axis to prevent overlap.

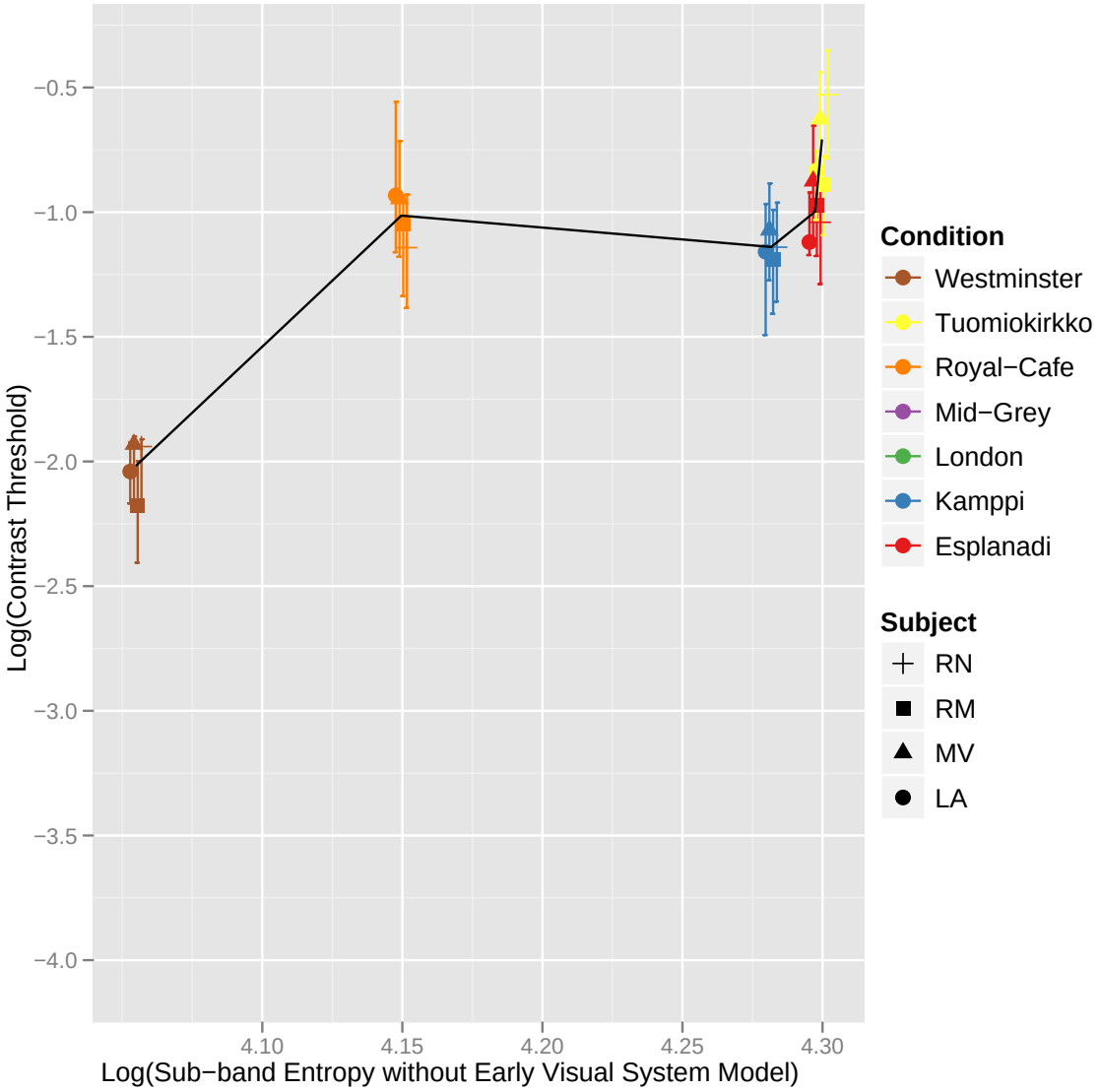


Figure 4.7: The logarithm of the experimental contrast thresholds as a function of the logarithm of the local sub-band entropy sum in the critical spacing area. The image was not filtered before computation of the sub-band entropy sum. Only results of a letter size of 2.96° of visual angle are presented. The results were similar for the other size categories. Subjects are jittered along the x-axis to prevent overlap.

CHAPTER

FIVE

DISCUSSION AND CONCLUSION

The study showed that the effect of crowding is apparent in natural images. The influence of basic image properties on the crowding effect was assessed. The results show a strong correlation between simple image statistics such as the local contrast energy and the experimental contrast thresholds in cluttered environments.

On the basis of the presented experimental and theoretical findings the methods developed in this thesis can be used to predict the crowding effect in an arbitrary image. The methods generate a crowding map based on the luminance information of the image. The map presents the locations of high crowding. The software developed is thus useful as an assessment tool e.g. for designers of desktop environments or augmented reality.

The primitive image statistics are coarse measures of the crowding effect. However, developing more sophisticated methods for the estimation of crowding magnitude would require more detailed information about the mechanisms of the phenomenon itself.

Below the main findings of the study are discussed. The experimental findings are examined, and comparisons with findings from previous studies are made. Future possibilities for the computational modelling of the crowding effect are also presented.

5.1 Result Remarks

The contrast thresholds of the target letter stimuli varied significantly depending on the background. The thresholds correlated with the local amount of specific features in the image. The features assessed were simple statistical properties of the image.

The varying contrast threshold itself showed congruent results among the participants. The change in threshold became saturated for each subject with an increase in the amount of local features in the image. The sigmoidal behaviour found in the peripheral contrast thresholds in complex natural backgrounds agree with previous research. An earlier study by Pelli et al. [2] showed similar saturation for letter targets on plain backgrounds and distracting letters with varying contrast. It can, thus, be argued that the crowding effect is apparent in natural backgrounds.

The study also implies that it is important to consider the effects of the early visual system. The results show significant differences in crowding prediction, depending on the visual system filtering. It is most probable that the significance of the early visual system is high in the crowding effect. The present study outlined the boundary conditions for the functioning of the visual system. Further research is needed to specifically elucidate the role of the early visual system in relation to visual stimuli and cortical activation.

5.2 Relevance of the Results

To the best knowledge of the author, there are no studies on the crowding effect in complex natural images. The study by Rosenholtz et al. [8] is the only study concerning natural images that also mentions the crowding effect. However, the clutter approach is significantly different and does not cover the crowding effect details.

The approach of the present study is quite different compared to the earlier research. Most of the research on the crowding effect attempts to discriminate the distinct factors for the crowding effect. This experiment is kept excessively simple to determine the main cause and the overall deterioration in visual perception in the peripheral vision. In this study the stimulus is chosen from everyday life. The method does not elucidate the mechanisms of the crowding effect but rather investigates the effect in a natural environment and searches for the relations between image statistics and the physiological restrictions of the perception. The methods do not consider the possibility that there might be differing amounts of ordinary masking or other phenomena involved. However, the findings showing the saturation in the most complex situations and the negligible effect of the stimulus size

indicate, that crowding is the main cause of the deterioration of the identification.

A specification of the individual factors causing the crowding effect was not possible in this study, due to the complexity of natural images. Assessing the effect of every image property separately would require enormous amounts of data and computations. The averaging over the local image patch around the stimulus was introduced to overcome this limitation. The average was computed to achieve general properties of images that could be used to quantify the tendency of image patches to cause crowding.

An in-depth study is required to resolve the interaction mechanisms of the features in natural images. It is beyond the scope of this study to define which features in the images jumble.

The averaged computational measures around the target patch correlated well with the experimental contrast thresholds. Computational measures assessed a group of features in the image patch. There were no significant differences in the results of three out of four of the implemented measures, although the complexity of the measures varied greatly. The most complex measure was feature congestion. It utilizes methods of early visual processing to assess image statistics. The sub-band entropy measures the information content and its distribution in separate spatial frequency bands. However, the best and the simplest measure was the sum of the contrast energy in the image patch. It gives a very simple description of the image content and worked very well in this study.

All the measures correlated with the contrast thresholds but all were quite coarse, as well. The contrast energy showed the most robust results. The contrast energy correlation coefficient was high in every condition (Spearman's ρ as high as 0.82), whereas the sub-band entropy showed quite good results merely in images with no early visual system filtering and the feature congestion showed good results only in the situation that involved the early visual system filtering.

5.3 Future Work

This study initiated the investigation of the crowding effect in complex natural images. A lot of research is still needed before the distinct features in natural images can be associated with the crowding effect. For the simplicity and the seminal nature of the research, the factors were kept minimal in the study. Only luminance was considered and the crowding was assessed in a single location only, for the sake of parsimony.

The chroma axis could be added to improved version of the crowding assessment protocol.

The inspirational models by Itti et al. [66] and Rosenholtz et al. [8] include colour assessment. A more detailed model of the early visual system and the crowding phenomenon could be produced as well. A research paradigm consisting of multiple target positions should be introduced with a more accurate model of the visual system.

The stationarity in the experimental results suggest that more accurate computational models could be developed for the crowding effect prediction tools. However, a clarification of the mechanisms causing the crowding effects is needed before the models describing the crowding effect can be improved.

A specific neuronal description of the phenomena causing the crowding effect would greatly contribute to the modelling of the crowding effect. This, however, requires a massive basic research on feature integration in the human visual system.

Nevertheless, development of the models for crowding is possible without excessive neuronal knowledge. The presented models ignore many details that the study on the crowding effect and the visual system in general have discovered. For example, the anisotropy in the crowding effect described in chapter 2, or the dependence on the pattern of the background could be constructed without additional research on neural mechanisms.

The essential target for further development of the crowding effect is the critical spacing model (see chapter 3 for details). The current model lacks the methods for assessing the feature integration in the surroundings of the target. According to the results of previous studies, it is most probable that feature integration affects the shape of the critical spacing area strongly. Subsequent studies in the field of assessing the crowding effect should, thus, consider these features in particular, possibly by transforming the whole process into a spatially limited feature space.

BIBLIOGRAPHY

[1] T. S. Lee and M. Nguyen. Dynamics of subjective contour formation in the early visual cortex. *Proc Natl Acad Sci U S A*, 98(4):1907–1911, Feb 2001. doi: 10.1073/pnas.031579998. URL http://dx.doi.org/10.1073/pnas.031579998.

[2] Denis G Pelli, Melanie Palomares, and Najib J Majaj. Crowding is unlike ordinary masking: distinguishing feature integration from detection. *J Vis*, 4(12):1136–1169, Dec 2004. doi: 10:1167/4.12.12. URL http://dx.doi.org/10:1167/4.12.12.

[3] Denis G Pelli, Katharine A Tillman, Jeremy Freeman, Michael Su, Tracey D Berger, and Najib J Majaj. Crowding and eccentricity determine reading rate. *J Vis*, 7(2):20.1–2036, 2007. doi: 10.1167/7.2.20. URL http://dx.doi.org/10.1167/7.2.20.

[4] H. Bouma. Interaction effects in parafoveal letter recognition. *Nature*, 226(5241): 177–178, Apr 1970.

[5] Endel Põder. Crowding with detection and coarse discrimination of simple visual features. *J Vis*, 8(4):24.1–24.6, 2008. doi: 10.1167/8.4.24. URL http://dx.doi.org/10.1167/8.4.24.

[6] Helena Ojanpää. *Visual search and eye movements: Studies of perceptual span*. PhD thesis, University of Helsinki, Faculty of Behavioural Sciences, 2006.

[7] Jeremy Freeman and Denis G Pelli. An escape from crowding. *J Vis*, 7(2):22.1–2214, 2007. doi: 10.1167/7.2.22. URL http://dx.doi.org/10.1167/7.2.22.

[8] Ruth Rosenholtz, Yuanzhen Li, and Lisa Nakano. Measuring visual clutter. *J Vis*, 7(2):17.1–1722, 2007. doi: 10.1167/7.2.17. URL http://dx.doi.org/10.1167/7.2.17.

[9] R. Navarro, P. Artal, and D. R. Williams. Modulation transfer of the human eye as a function of retinal eccentricity. *J Opt Soc Am A*, 10(2):201–212, Feb 1993.

[10] Th. Wertheim. Über die indirekte sehschärfe. *Zeitschrift für Psychologie und Physiologie die Sinnesorgane*, 7:172–187, 1891.

[11] D. R. Williams, P. Artal, R. Navarro, M. J. McMahon, and D. H. Brainard. Off-axis optical quality and retinal sampling in the human eye. *Vision Res*, 36(8):1103–1114, Apr 1996.

[12] URL `http://en.wikipedia.org/wiki/Eye`.

[13] R. H. Masland. Neuronal diversity in the retina. *Curr Opin Neurobiol*, 11(4): 431–436, Aug 2001.

[14] A. L. Hodgkin and A. F. Huxley. A quantitative description of ion currents and its applications to conduction and excitation in nerve membranes. *J. Physiol. (Lond.)*, 117:500–544., 1952.

[15] C. A. Curcio, K. R. Sloan, R. E. Kalina, and A. E. Hendrickson. Human photoreceptor topography. *J Comp Neurol*, 292(4):497–523, Feb 1990. doi: 10.1002/cne.902920402. URL `http://dx.doi.org/10.1002/cne.902920402`.

[16] C. A. Curcio and K. A. Allen. Topography of ganglion cells in human retina. *J Comp Neurol*, 300(1):5–25, Oct 1990. doi: 10.1002/cne.903000103. URL `http://dx.doi.org/10.1002/cne.903000103`.

[17] J. Intriligator and P. Cavanagh. The spatial resolution of visual attention. *Cognit Psychol*, 43(3):171–216, Nov 2001. doi: 10.1006/cogp.2001.0755. URL `http://dx.doi.org/10.1006/cogp.2001.0755`.

[18] C. Enroth-Cugell and J. G. Robson. The contrast sensitivity of retinal ganglion cells of the cat. *J Physiol*, 187(3):517–552, Dec 1966.

[19] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol*, 160:106–154, Jan 1962.

[20] Barry B Lee. Paths to colour in the retina. *Clin Exp Optom*, 87(4-5):239–248, Jul 2004.

[21] Matteo Carandini, Jonathan B Demb, Valerio Mante, David J Tolhurst, Yang Dan, Bruno A Olshausen, Jack L Gallant, and Nicole C Rust. Do we know what the early visual system does? *J Neurosci*, 25(46):10577–10597, Nov 2005. doi: 10.1523/JNEUROSCI.3726-05.2005. URL http://dx.doi.org/10.1523/JNEUROSCI.3726-05.2005.

[22] J. Allman, F. Miezin, and E. McGuinness. Stimulus specific responses from beyond the classical receptive field: neurophysiological mechanisms for local-global comparisons in visual neurons. *Annu Rev Neurosci*, 8: 407–430, 1985. doi: 10.1146/annurev.ne.08.030185.002203. URL http://dx.doi.org/10.1146/annurev.ne.08.030185.002203.

[23] Peter H Schiller and Christina E Carvey. The hermann grid illusion revisited. *Perception*, 34(11):1375–1397, 2005.

[24] Keith A Schneider, Marlene C Richter, and Sabine Kastner. Retinotopic organization and functional subdivisions of the human lateral geniculate nucleus: a high-resolution functional magnetic resonance imaging study. *J Neurosci*, 24 (41):8975–8985, Oct 2004. doi: 10.1523/JNEUROSCI.2413-04.2004. URL http://dx.doi.org/10.1523/JNEUROSCI.2413-04.2004.

[25] Stephen E. Palmer. *Vision Science: Photons to Phenomenology*. The MIT Press, 1999.

[26] L. C. Silveira and V. H. Perry. The topography of magnocellular projecting ganglion cells (m-ganglion cells) in the primate retina. *Neuroscience*, 40(1):217–237, 1991.

[27] Javier Cudeiro and Adam M Sillito. Looking back: corticothalamic feedback and early visual processing. *Trends Neurosci*, 29(6): 298–306, Jun 2006. doi: 10.1016/j.tins.2006.05.002. URL http://dx.doi.org/10.1016/j.tins.2006.05.002.

[28] D. L. Robinson and S. E. Petersen. The pulvinar and visual salience. *Trends Neurosci*, 15(4):127–132, Apr 1992.

[29] David C. Van Essen. *The Visual Neurosciences*, chapter Organization of Visual Areas in Macaque and Human Cerebral Cortex, pages 507–521. MIT Press, 2004.

[30] R. B. H. Tootell, N. K. Hadjikhani, J. D. Mendola, S. Matrrett, and A. M. Dale. From retinotopy to recognition: fmri in human visual cortex. *Trends in Cognitive Sciences*, 2:174–183, 1998.

[31] Mortimer Mishkin, Leslie G. Ungerleider, and Kathleen A. Macko. Object vision and spatial vision: two cortical pathways. *Trends in Neurosciences*, 6:414–417, 1983.

[32] J. V. Haxby, C. L. Grady, B. Horwitz, L. G. Ungerleider, M. Mishkin, R. E. Carson, P. Herscovitch, M. B. Schapiro, and S. I. Rapoport. Dissociation of object and spatial visual processing pathways in human extrastriate cortex. *Proc Natl Acad Sci U S A*, 88(5):1621–1625, Mar 1991.

[33] M. A. Goodale and A. D. Milner. Separate visual pathways for perception and action. *Trends Neurosci*, 15(1):20–25, Jan 1992.

[34] Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, Mar 2008. doi: 10.1038/nature06713. URL `http://dx.doi.org/10.1038/nature06713`.

[35] G. B. Stanley, F. F. Li, and Y. Dan. Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. *J Neurosci*, 19(18):8036–8042, Sep 1999.

[36] P. T. Fox, F. M. Miezin, J. M. Allman, D. C. Van Essen, and M. E. Raichle. Retinotopic organization of human visual cortex mapped with positron-emission tomography. *J Neurosci*, 7(3):913–922, Mar 1987.

[37] G. Leuba and R. Kraftsik. Changes in volume, surface estimate, three-dimensional shape and total number of neurons of the human primary visual cortex from midgestation until old age. *Anat Embryol (Berl)*, 190(4):351–366, Oct 1994.

[38] Alessandra Angelucci, Jonathan B Levitt, Emma J S Walton, Jean-Michel Hupe, Jean Bullier, and Jennifer S Lund. Circuits for local and global signal integration in primary visual cortex. *J Neurosci*, 22(19):8633–8646, Oct 2002.

[39] J. D. Mendola, A. M. Dale, B. Fischl, A. K. Liu, and R. B. Tootell. The representation of illusory and real contours in human cortical visual areas revealed by functional magnetic resonance imaging. *J Neurosci*, 19(19):8560–8572, Oct 1999.

[40] K. R. Huxlin, R. C. Saunders, D. Marchionini, H. A. Pham, and W. H. Merigan. Perceptual deficits after lesions of inferotemporal cortex in macaques. *Cereb Cortex*, 10(7):671–683, Jul 2000.

[41] A. Cyrus Arman, Susana T. L. Chung, and Bosco S. Tjan. Neural correlates of letter crowding in the periphery. *Journal of Vision*, 6:804, 2006.

[42] A. Hyvärinen, J. Hurri, and P. O. Hoyer. *Natural Image Statistics - A modern approach to visual neuroscience and image processing*. Not, 2007.

[43] M. A. Freed. Rate of quantal excitation to a retinal ganglion cell evoked by sensory input. *J Neurophysiol*, 83(5):2956–2966, May 2000.

[44] Garrett T Kenyon, Bryan J Travis, James Theiler, John S George, Gregory J Stephens, and David W Marshak. Stimulus-specific oscillations in a retinal model. *IEEE Trans Neural Netw*, 15(5):1083–1091, Sep 2004. doi: 10.1109/TNN.2004.832722. URL `http://dx.doi.org/10.1109/TNN.2004.832722`.

[45] J. Rovamo, J. Mustonen, and R. Näsänen. Neural modulation transfer function of the human visual system at various eccentricities. *Vision Res*, 35(6):767–774, Mar 1995.

[46] R. N. Bracewell. *The Fourier Transform and Its Applications*. McGraw-Hill Education, 1978.

[47] A. Cowey and E. T. Rolls. Human cortical magnification factor and its relation to visual acuity. *Exp Brain Res*, 21(5):447–454, 1974.

[48] J. Rovamo and V. Virsu. An estimation and application of the human cortical magnification factor. *Exp Brain Res*, 37(3):495–510, 1979.

[49] N. Drasdo. The neural representation of visual space. *Nature*, 266(5602):554–556, Apr 1977.

[50] S. M. Anstis. Letter: A chart demonstrating variations in acuity with retinal position. *Vision Res*, 14(7):589–592, Jul 1974.

[51] F. W. Campbell, J. J. Kulikowski, and J. Levinson. The effect of orientation on the visual resolution of gratings. *J Physiol*, 187(2):427–436, Nov 1966.

[52] S. Thorpe, A. Delorme, and R. Van Rullen. Spike-based strategies for rapid processing. *Neural Netw*, 14(6-7):715–725, 2001.

[53] Thibaud Debaecker and Ryad Benosman. Bio-inspired model of visual information encoding for localization: from the retina to the lateral geniculate nucleus. *J Integr Neurosci*, 6(3):477–509, Sep 2007.

[54] D. Gabor. Theory of communication. *J. IEE (London)*, 93:429–457, 1946.

[55] S. Marcelja. Mathematical description of the responses of simple cortical cells. *J Opt Soc Am*, 70(11):1297–1300, Nov 1980.

[56] J. G. Daugman. Two-dimensional spectral analysis of cortical receptive field profiles. *Vision Res*, 20(10):847–856, 1980.

[57] Ben Willmore, Ryan J Prenger, Michael C-K Wu, and Jack L Gallant. The berkeley wavelet transform: a biologically inspired orthogonal wavelet transform. *Neural Comput*, 20(6):1537–1564, Jun 2008. doi: 10.1162/neco.2007.05-07-513. URL `http://dx.doi.org/10.1162/neco.2007.05-07-513`.

[58] Antonio Torralba and Aude Oliva. Statistics of natural image categories. *Network*, 14(3):391–412, Aug 2003.

[59] David J. Field. What is the goal of sensory coding? *Neural Comput.*, 6(4):559–601, 1994. ISSN 0899-7667. doi: http://dx.doi.org/10.1162/neco.1994.6.4.559. URL `http://dx.doi.org/10.1162/neco.1994.6.4.559`.

[60] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, Jun 1996. doi: 10.1038/381607a0. URL `http://dx.doi.org/10.1038/381607a0`.

[61] Allison Woodruff, James Landay, and Michael Stonebraker. Constant information density in zoomable interfaces. In *AVI '98: Proceedings of the working conference on Advanced visual interfaces*, pages 57–65, New York, NY, USA, 1998. ACM.

[62] A. Oliva, M. L. Mack, M. Shrestha, and A. Peeper. Identifying the perceptual dimensions of visual complexity of scenes. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society.*, 2004.

[63] Julian E. Hochberg. *Perception*. Prentice-Hall, 1964.

[64] Ruth Rosenholtz, Yuanzhen Li, Zhenlan Jin, and Jonathan Mansfield. Feature congestion: A measure of visual clutter. *J Vis*, 6(6):827–827, 6 2006. ISSN 1534-7362. URL `http://journalofvision.org/6/6/827/`.

[65] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res*, 40(10-12):1489–1506, 2000.

[66] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Journal PAMI*, 20(11):1254–1259, Nov. 1998. doi: 10.1109/34.730558.

[67] E. Barth, C. Zetzsche, and I. Rentschler. Intrinsic two-dimensional features as textons. *J Opt Soc Am A Opt Image Sci Vis*, 15(7):1723–1732, Jul 1998.

[68] C Zetzsche, H K Schill, G Deubel, and E Krieger. Investigation of a sensorimotor system for saccadic scene analysis: an integrated approach. In *From Animals to Animats 5: Proceedings of the fifth international conference of simulation of adaptive behavior*, 1998.

[69] Lior Elazary and Laurent Itti. Interesting objects are visually salient. *J Vis*, 8(3): 3.1–315, 2008. doi: 10.1167/8.3.3. URL `http://dx.doi.org/10.1167/8.3.3`.

[70] Li Zhaoping. Attention capture by eye of origin singletons even without awareness–a hallmark of a bottom-up saliency map in the primary visual cortex. *J Vis*, 8(5): 1.1–118, 2008. doi: 10.1167/8.5.1. URL `http://dx.doi.org/10.1167/8.5.1`.

[71] Jiri Najemnik and Wilson S Geisler. Optimal eye movement strategies in visual search. *Nature*, 434(7031):387–391, Mar 2005. doi: 10.1038/nature03390. URL `http://dx.doi.org/10.1038/nature03390`.

[72] Jiri Najemnik and Wilson S Geisler. Eye movement statistics in humans are consistent with an optimal search strategy. *J Vis*, 8(3):4.1–414, 2008. doi: 10.1167/8.3.4. URL `http://dx.doi.org/10.1167/8.3.4`.

[73] Preeti Verghese and Suzanne P McKee. Visual search in clutter. *Vision Res*, 44(12):1217–1225, Jun 2004. doi: 10.1016/j.visres.2003.12.006. URL `http://dx.doi.org/10.1016/j.visres.2003.12.006`.

[74] Risto Näsänen and Helena Ojanpää. How many faces can be processed during a single eye fixation? *Perception*, 33(1):67–77, 2004.

[75] Dennis M Levi. Crowding–an essential bottleneck for object recognition: a mini-review. *Vision Res*, 48(5):635–654, Feb 2008. doi: 10.1016/j.visres.2007.12.009. URL `http://dx.doi.org/10.1016/j.visres.2007.12.009`.

[76] Keith A May and Robert F Hess. Ladder contours are undetectable in the periphery: a crowding effect? *J Vis*, 7(13):9.1–915, 2007. doi: 10.1167/7.13.9. URL `http://dx.doi.org/10.1167/7.13.9`.

[77] P. De Weerd, R. Desimone, and L. G. Ungerleider. Cue-dependent deficits in grating orientation discrimination after v4 lesions in macaques. *Vis Neurosci*, 13(3):529–538, 1996.

[78] A. Toet and D. M. Levi. The two-dimensional shape of spatial interaction zones in the parafovea. *Vision Res*, 32(7):1349–1357, Jul 1992.

[79] Brad C Motter and Diglio A Simoni. The roles of cortical image separation and size in active visual search performance. *J Vis*, 7(2):6.1–615, 2007. doi: 10.1167/7.2.6. URL `http://dx.doi.org/10.1167/7.2.6`.

[80] F. L. Kooi, A. Toet, S. P. Tripathy, and D. M. Levi. The effect of similarity and duration on spatial interaction in peripheral vision. *Spat Vis*, 8(2):255–279, 1994.

[81] Björn N S Vlaskamp and Ignace Th C Hooge. Crowding degrades saccadic search performance. *Vision Res*, 46(3):417–425, Feb 2006. doi: 10.1016/j.visres.2005.04.006. URL `http://dx.doi.org/10.1016/j.visres.2005.04.006`.

[82] URL `http://python.org/`.

[83] URL `http://scipy.org/`.

[84] R. Näsänen, H. Kukkonen, and J. Rovamo. Relationship between spatial integration and spatial spread of contrast energy in detection. *Vision Res*, 34(7):949–954, Apr 1994.

APPENDIX

A

THE SOFTWARE

The Python software was developed according to the principles of model-view-controller (MVC) design pattern. The program architecture separates the functionalities and consists of three classes.

1. Graphical user interface (GUI)
2. Mathematical computation
3. Image presentation and marking

In the current version, the GUI class includes help and guidance functions and the software is hence not usable without the GUI. The software has, however, been designed to allow the implementation of non-GUI (command line) applications. The mathematical core of the software tool is therefore fully separated from the GUI, facilitating command line usage.

The GUI class consist of the guide and input/output functions. A lot of attention is paid to the readability of the code. The functions are grouped so that the data and building functions are separated. The protocol is illustrated in listing A.1. The input/output functioning of the GUI class is separated similarly.

Listing A.1: Example of the code in the GUI class

```
# civs_gui.py
program_name='CiVS'
```

```
program_version='0.8'
########################################
# Main terminal window class for image
# analysis program CiVS
# Program analyzes clutter and mimics
# a human visual system
#
# Author: Lauri Ahonen
#         Finnish Institute of
#         Occupational Health
########################################
import wx
import os
import civs_motor
import civs_illu


## Terminal window class ================
class mainWindow(wx.Frame):

    # Initial Actions ---------------------
    def __init__(self, parent, id):
        self.ACTIVE_CHILD = None
        wx.Frame.__init__(self, parent, id, program_name, size=(640, 480))
        self.manipulator = civs_motor.Manipulate(parent=self)
        icon = wx.Icon('ico.ico', wx.BITMAP_TYPE_ICO)
        self.SetIcon(icon)
        self.panel = self.makeTerminalPanel()
        self.Bind(wx.EVT_CLOSE, self.onCloseWindow)
        self.makeMenuBar()
        self.makeToolBar()
        self.makeStatusBar()
        self.initWriteTerminal()
    # -------------------------------------

    # Sets the current frame (image the user wants to operate)
    def setActiveChild(self, active):
        self.ACTIVE_CHILD = active
    # -------------------------------------

    # Menu bar is created first ----------
    def menuData(self):
        return(("&File",
                ("&Open", "Open an existing image", self.onOpen),
                ("E&xit", "Terminate the program", self.onCloseWindow)),
               ("&Filter",
                ("Optic and Retinal",
                 "Optical features and retinal physiology mimicking filtering.", self.
                     onOpt),
                ("All parts of LGN",
                 "Combined Lateral Geniculate Nucleus presentation", self.onLGN),
                ("Striate cortex","Illustration of the image coded by V1", self.onV1),
                ("", "", ""),
                ("Complexity evaluation (Ruth Rosenholtz)",
                 "Quantified complexity in terms of Rosenholtz", self.onRuthComp),
                ("Complexity evaluation (Risto Nasanen)",
                 "Quantified complexity in terms of Risto", self.onRistoComp)),
```

```
                ("&Help",
                 ("About", "Product Information", self.onInfoDialog)))


    def makeMenuBar(self):
        menuB = wx.MenuBar()
        for each in self.menuData():
            label = each[0]
            items = each[1:]
            menuB.Append(self.makeMenu(items), label)
            self.SetMenuBar(menuB)
```

The image presentation class handles the illustration of the filter effects and displaying the images in frames. The functional purpose of the image presentation class is to register the user-provided information about the fixation and target positions on the image. The information is saved by the image frame object. Example of the code for this functionality is presented in listing A.2.

Listing A.2: Example of the code in the illustration class

```
# Drawing the image and the fixation mark
    def OnPaint(self, event):
        self.dc = wx.PaintDC(self)
        self.dc.SetPen(wx.TRANSPARENT_PEN)
        self.dc.DrawBitmap(wx.BitmapFromImage(self.image), 0, 0)
        self.dc.SetBrush(wx.RED_BRUSH)
        if self.fixx or self.fixy is not None:
            self.dc.DrawCircle(self.fixx,self.fixy,3)
        self.dc.SetBrush(wx.GREEN_BRUSH)
        if self.tarx or self.tary is not None:
            self.dc.DrawCircle(self.tarx,self.tary,3)
    # ------------------------------------

    # Set mark on fixation and target points ---------
    def drawImg(self,pos,s):
        if s is 0:
            self.fixx = pos.x
            self.fixy = pos.y
        else:
            self.tarx = pos.x
            self.tary = pos.y

        self.dc.Clear()
        self.dc.DrawBitmap(wx.BitmapFromImage(self.image), 0, 0)
        if self.fixx or self.fixy is not None:
            self.dc.SetBrush(wx.RED_BRUSH)
            self.dc.DrawCircle(self.fixx,self.fixy,3)
        if self.tarx or self.tary is not None:
            self.dc.SetBrush(wx.GREEN_BRUSH)
            self.dc.DrawCircle(self.tarx,self.tary,3)
    # ------------------------------------
```

The computational properties of the program are provided by the computations class. This class utilizes packages for scientific computing (scipy.signal, scipy.ndimage and scipy.fftpack) and implements the algorithms presented in chapter 3. In the current version the experimental conditions must also be set here. There is readiness for the computations to run in stand-alone console-mode.

In addition to the methods presented in chapter 3 the class implements optical features that is not in use in the current version. It is implemented to model the optical features of the eye but omitted due to the reasons presented in chapter 2.

Example code is printed in listing A.3

Listing A.3: Example of the code in the mathematical class

```
# User changeable parameters in the beginning
    def experimentConditions(self):
        """ Physical parameters of test conditions """

        return {'pixel size':0.271, # mm
                'display distance':570, # mm
                '0.5 degrees in radians':0.00872664625997,
                'radian in degrees':57.2957795,
                'oblateness of simple cells':0.75, # height-width ratio
                'pyramide levels':4}

        """ _ _ _ _ _ _ _ _ _  _ _ _ _ _ _ _ _ _ """
    # ------------------------------------

    # Berkeley wavelet transform -wavelets
        """ e.g.
                  [ 0 -1  1]
                A*[ 1  0 -1]    /A is normalizing constant (make to sum to unity)
                  [-1  1  0]
        """
    def mamaWavelets(self):
        return (('0o',(1/numpy.sqrt(6))*numpy.array([[-1,0,1],[-1,0,1],[-1,0,1]])),
                ('0e',(1/numpy.sqrt(18))*numpy.array([[-1,2,-1],[-1,2,-1],[-1,2,-1]])),
                ('45o',(1/numpy.sqrt(6))*numpy.array([[-1,1,0],[1,0,-1],[0,-1,1]])),
                ('45e',(1/numpy.sqrt(18))*numpy.array([[-1,-1,2],[-1,2,-1],[2,-1,-1]])),
                ('90o',(1/numpy.sqrt(6))*numpy.array([[-1,-1,-1],[0,0,0],[1,1,1]])),
                ('90e',(1/numpy.sqrt(18))*numpy.array([[-1,-1,-1],[2,2,2],[-1,-1,-1]])),
                ('135o',(1/numpy.sqrt(6))*numpy.array([[0,-1,1],[1,0,-1],[-1,1,0]])),
                ('135e',(1/numpy.sqrt(18))*numpy.array([[2,-1,-1],[-1,2,-1],[-1,-1,2]])))
    # ------------------------------------

    # Translate wx.Image -object to numeric matrix structure
    def image2mtx(self, image):
        width = image.GetWidth()
        height = image.GetHeight()
        rgb = numpy.fromstring(image.GetData(),numpy.uint8)
        length = rgb.size
        R = rgb[range(0,length,3)]
        G = rgb[range(1,length,3)]
        B = rgb[range(2,length,3)]
```

```
        gray = 0.23*R + 0.71*G + 0.06*B
        normLuminance = gray/gray.max()
        mtx = normLuminance.reshape(height,width)

        return mtx
    # ------------------------------------

    # Convert matrix structure to wx.Image -object
    def mtx2image(self, filteredMtx,gamma=None):
        imgLayer = filteredMtx.flatten()
        if gamma != None: # Gamma correction if needed
            imgLayer = numpy.sqrt((imgLayer-imgLayer.min())*imgLayer.max())
        normLayer = numpy.uint8((imgLayer-imgLayer.min())*(220/imgLayer.max()))
        lll = numpy.repeat(normLayer,3)
        filteredImg = wx.EmptyImage(filteredMtx.shape[1],filteredMtx.shape[0])
        filteredImg.SetData(lll.tostring())

        return filteredImg
    # ------------------------------------

    # Retinal filtering for image ---------
    def retinal(self,image,name,fixpos,targpos=0):

        # - - - - - - - - - - - - - - - - -
        mtx = self.image2mtx(image)
        # - - - - - - - - - - - - - - - - -

        height = numpy.float32(mtx.shape[0])
        width = numpy.float32(mtx.shape[1])
        fixx = (fixpos.x-width/2)/width
        fixy = (fixpos.y-height/2)/height

        # Visual angle determination  - -
        conditions = self.experimentConditions()
        pixelSize = conditions['pixel size']
        distanceToDisplay = conditions['display distance']
        rads = conditions['0.5 degrees in radians']
        degs = conditions['radian in degrees']
        obl= conditions['oblateness of simple cells']
        aH = 2*numpy.arctan(0.5*height*pixelSize/distanceToDisplay)*degs
        aW = 2*numpy.arctan(0.5*width*pixelSize/distanceToDisplay)*degs
        # Maximum frequency in image: amount of pixel pairs in visual degree
        maxFreq = (numpy.tan(rads)*distanceToDisplay)/pixelSize
        # - - - - - - - - - - - - - - - - -

        # Transform to frequency space  -
        fourier = scipy.fftpack.fft2(mtx)
        zeroComp = fourier[0,0]
        fourier[0,0] = 0
        fourier = scipy.fftpack.fftshift(fourier)
        # - - - - - - - - - - - - - - - - -

        # Filter and a modulator grids  -
        grid = numpy.float64(numpy.mgrid[-height/2:height/2,-width/2:width/2])
        # Normalization
        nGrid = grid[0]/grid[0].shape[0], grid[1]/grid[1].shape[1]
        # Radius for Butterworth filter
```

```
        radius = numpy.hypot(normGrid[0], normGrid[1])
        # - - - - - - - - - - - - - - - - - -

        # Cortical magnification factor (fitted value)
        M = 1
        # Ellipse for spatial modulation (ellipse dimensions from Wertheim 1891)
        ellipse = numpy.hypot(M*aH*(abs(nGrid[0]-fixy)),M*obl*aW*(abs(nGrid[1]-fixx)))
        # - - - - - - - - - - - - - - - - - -

        # Optical features of eye (not in use)
        #A1 = 0.1743
        #A2 = 0.0392
        #B1 = 0.0362
        #B2 = 0.0172
        #C1 = 0.215
        #C2 = 0.00294
        #cosGrid = numpy.cos(numpy.pi*nGrid[0]), numpy.cos(numpy.pi*nGrid[1])
        #modulated =( (1 - C1 + C2*cosGrid[0]) * numpy.exp(-A1 * numpy.exp(A2 * cosGrid
            [0]))
        #               + (C1 - C2*cosGrid[1]) * numpy.exp(-B1 * numpy.exp(B2 * cosGrid
            [1])) )


        #  Bands (octave fractions) in which the image frequencies are treated
        octaves = numpy.logspace(-10,-1,num=20,base=2)
        # - - - - - - - - - - - - - - - - - -

        # Loop to filter each band separately
        prev = 0
        for band in octaves:
            # Calculating the band
            if prev == 0:
                trtBand = numpy.float64((1 / (1.0 + (radius / band)**(40))))
            else:
                trtBand = numpy.float64((1 / (1.0 + (radius / band)**(40)))
                                        -(1 / (1.0 + (radius / prev)**(40))))
            # The defined band in the image to spatial space
            currentBand = scipy.fftpack.ifft2(scipy.fftpack.ifftshift(trtBand*fourier))
            # Spatial modulation for the band of image
            midFreq = numpy.mean((band,prev))*maxFreq
            # Constant to adjust the power of filter (coarse assessment from
            alpha = 0.1
            # Demo filtering without neural high-pass
            if targpos != 0: #|| if NEURAL HP-FILTERING GIVES BAD RESULTS
                modulator = numpy.exp(-numpy.abs(ellipse)*midFreq*alpha)
            # "Real case" simulation with equation:
            else:
                nHP = midFreq # Neural high pass (Rovamo et al.)
                modulator = nHP*numpy.exp(-numpy.abs(ellipse)*midFreq*alpha)
                # Demo
            if prev == 0:
                allBands = modulator*currentBand.real # Small imaginary parts
            else:
                allBands += modulator*currentBand.real
            prev = band

        # Add Fourier transform zero component (a0)
```

```
        filtMtx =  allBands + numpy.abs(zeroComp)/mtx.size
        # - - - - - - - - - - - - - - - - - - -
        filtered = self.mtx2image(filtMtx)
        # - - - - - - - - - - - - - - - - - - -
        self.mama.dispOnTerminal('retinal',name)
        # - - - - - - - - - - - - - - - - - - -
        return filtered, filtMtx
    # ------------------------------------
```